Object Detection at Multiple Scales Improves Accuracy

Stanley Bileschi
Massachusetts Institute of Technology
Cambridge MA
bileschi@mit.edu

Abstract

For detecting objects in natural visual scenes, several powerful image features have been proposed which can collectively be described as spatial histograms of oriented energy. The HoG [3], HMAX C1 [12], SIFT [10], and Shape Context feature [2] all represent an input image using with a discrete set of bins which accumulate evidence for oriented structures over a spatial region and a range of orientations. In this work, we generalize these techniques to allow for a foveated input image, rather than a rectilinear raster in order to improve object detection accuracy. The system leverages a spectrum of image measurements, from sharp, fine-scale image sampling within a small spatial region to coarse-scale sampling of a wide field of view. In the experiments we show that features generated from the foveated input format produce detectors of greater accuracy, as measured for four object types from commonly available data-sets.

1. Introduction

In the field of object detection, often the features used to represent an input are more important to accurate performance than the statistical techniques used to learn patterns of those features. Whereas the earliest detectors used simple cues such as gray-scale values, wavelet coefficients, or histograms of RGB values, modern techniques can attribute much of their success to features such as Lowe's SIFT [10], Dalal's Histogram of Oriented Gradients (HoG) [3], the visual Bag-Of-Words [16], and hierarchical networks of selectivity and invariance, such as Poggio's HMAX network, LeCun's convolutional network, among notable others [9, 12, 8, 7, 6].

Most image feature algorithms ingest input brightness samples at regular spatial intervals. This work adapts the HoG and HMAX feature to input a spectrum of brightness values, sampled densely at a fine scale at the center of the target, and coarsely further away. That is to say, the aim of this work is to improve on existing methods via adaptation to a foveated input.

The motivation behind such an approach is that recent vision experiments exploring larger spatial regions (e.g., [17, 18, 11]) suggest that robust object detection may often be a matter of context as much as appearance. We should sample images at multiple scales so as to have access to context, shape, and texture.

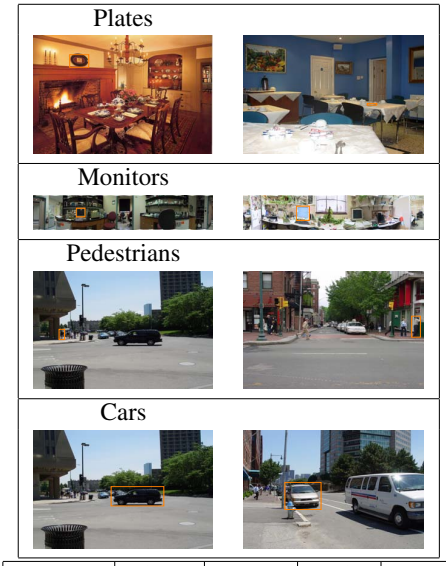
It is critical to clearly distinguish between two separate notions of visual scale. In one sense, scale refers to the size of the visual region under scrutiny, with respect to the object in question. At a large scale, the target is only a small part of the region. The other type of scale has to do with the size of the filters used to inspect the properties of the visual area, and has more to do with the frequency domain. Many visual feature algorithms begin with processing an image with many scales of Gabor filters, for instance, and each filter responds most strongly to brightness modulations at it's own scale. It is the first notion of scale which is the focus of this work. Hopefully it will be obvious which scale is meant by context.

There are two major experimental efforts within this work. First we will explore how the scale of the input image, relative to the size of the target, affects the performance of an object detector, independent of available resolution. This experiment will uncover the relative utility of differing scales. Secondly, multiple scales will be included into the same classifier in a simple early-fusion framework. This experiment will show that features from different scales provide complementary information. Adaptations to both HMAX and HoG are described and tested.

2 The HoG and HMAX Features

The Histogram of Oriented Gradients (HoG) algorithm and the Hierarchal Maximization Architecture (HMAX) algorithm are well-known, successful methods for converting an input image into a mathematical

978-1-4244-2175-6/08/$25.00 ©2008 IEEE



	Plates	Mons.	Peds.	Cars
nObjs.	234	395	1449	5799
nScenes	46	195	852	3090

Figure 1. Two full source data scenes for each of the four object types. Targets are annotated with orange bounding boxes. Note that the target is generally much smaller than the scene as a whole.

description, or feature vector [3, 12]. HoG was designed as an adaptation to SIFT [10] to maximize the detection accuracy on a pedestrian detection task, while HMAX was designed primarily to mimic the behavior of the early stages of the mammalian ventral visual stream. For the purposes of this work, we will only be using the first layers of HMAX, S1 and C1.

These features, and other features like them, have been used to accurately detect a wide variety of objects in natural scenes [4, 15]. Along with SIFT and Shape Context [2], they produce feature vectors which are accumulations of evidences for spatial patterns of oriented structures within the input. The HoG and HMAX features begin by calculating a gradient magnitude and direction at each pixel. This is followed by a pooling and normalization stage which computes representative statistics while reducing the dimensionality.

Note that these features, and others, [2, 14], do perform a form of multi-scale processing by filtering at multiple resolutions. Again, this is different from the multiscale processing than investigated here, i.e., the size of the input relative to the size of the target.

3 Data

These experiments required a suitable database of labeled objects within their natural contexts. It was important that the data included a wide background field around the objects, in order to explore larger scales. Furthermore, many labeled examples of the targets were necessary in order to have enough data to train a classifier and still perform statistically significant tests. Two objects in the LabelMe database [13] and two in the StreetScenes DB. [1] were found to meet the constraints. Fig. 1 illustrates typical examples of these data and relates each object to the number of labeled examples and the total number of images.

For negative examples, it was necessary to chose from a distribution similar to that of the positive data, to prevent learning spurious statistics unrelated to the presence of the object. Locations and images were chosen from the same marginal distributions as the positive data. Any candidate negative whose minimum bounding square intersected the bounding square of a positive example with an intersect to union ratio greater than .25 was rejected. An independent set of negatives were chosen for each positive class. Unlabeled examples can sometimes be included due to imperfections in the ground truth. These represent a very small minority of the actual negative examples.

4 Single Scale Experiments

Object detectors are trained and tested using an existing image feature, but the input image is varied in scale, relative to the size of the ground truth hand-drawn label. This experiment is repeated for two choices of image feature (HoG and HMAX), using two different classifiers (gentleBoost [5], and a linear-kernel SVM), on the four object databases (Pedestrian, Car, Plate and Monitor), as described in Section 3.

Each experimental condition is executed as follows. First a set of positive and negative images were cropped from the database. The crop region was selected by first finding the minimum square bounding box around the object, and then scaling that box by some scale factor. The scale factors ranged from a minimum of $\frac{1}{2}$ to a maximum of 16 times the size of the original box. Figure 2 illustrates the set of bounding boxes extracted for an example pedestrian. The small scales are indeed smaller than the target object and may be completely within the object. The largest scales leave the target object as a very small part of the window, most of the window is background or clutter. When the crop region extended beyond the image, pixels were filled in by assuming the image was symmetric across the edge. Some small artifacts are visible in Fig. 2.

The positive and negative crops are all converted to grayscale and resized to 128×128 pixels using MATLAB's bilinear `imresize` function. This size was chosen to match the experiments of [15, 3] as closely as possible. The images were then converted into the target feature format, and a classifier was trained and tested using 5 random training and testing splits. In these experiments 75% of the data was used for training, and the remaining 25% for testing.

Figure 3 plots the average equal-error-rate (EER) of the resulting ROC curves as a function of scale. The blue circles indicate systems trained and testing using gentleBoost, and the red $\times$s SVMs, though there is no statistical difference. Scale index 4 is the scale factor where the extraction boundaries are equal to the minimum square bounding box enclosing the ground-truth polygon.

Two conclusions are to be drawn. Firstly, in all cases there is a preferred scale, larger than the minimum bounding box, which reliably produces the most accurate detections. As the crop region grows larger or shrinks smaller than this preferred scale, the performance suffers. Secondly, we see that that the performance of the detector is strong even at scales very different from the preferred scale. This suggests that there is discriminative information in these measurements.

5 Multi-Scale Experiments

Now multiple scales will be used simultaneously, and results compared to the single-scale results in Fig. 4. Sec. 5.1 will simply use the concatenation of 3 feature vectors from differing scales. Sec. 5.2 will correct for the increased amount of information input to the system, by using lower resolution inputs.

5.1 3 Full fields

Here we will determine whether a simple multi-scale approach will outperform a single, optimized scale. Features from 3 separate scales are fed into the classifier simultaneously.

First HoG or HMAX features were calculated from three scales independently as in the previous experiment. Scale factors 2, 6, and 10 were selected, corresponding to scales smaller than the object, slightly larger than the object, and much larger than the object (scale-factors 0.63, 1.59 ,and 4). These scales were chosen since they represent a wide range, but none so small or or so large as to severely impair performance, as can be seen from Fig. 3. A boosting classifier is trained on each of the three sets of features *independently*, noting the features from which the stumps were derived. Then a single monolithic boosting classifier is trained on the union of those three selected sets of features. This more

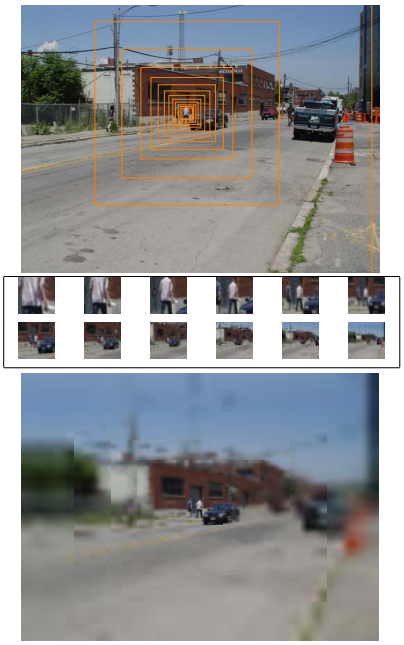


Figure 2. Top: Orange boxes indicate the sizes of the crops used in the scale sensitivity tests. The largest box extends beyond the image. Middle: Uniform resolution extractions from each box. Bottom: Pedestrian reconstructed from overlaid crops.

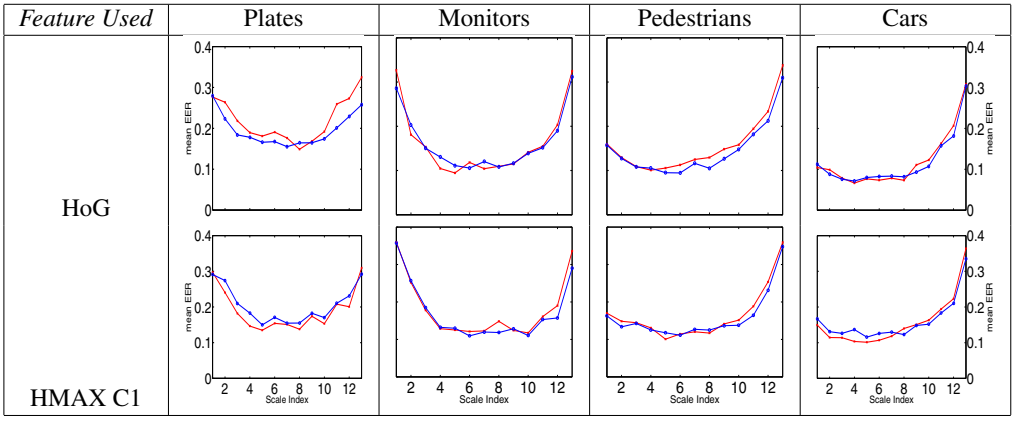


Figure 3. Detection performance as a function of input scale. Detailed in Sec. 4. Scales are indexed from factor .5 to 16 (too small to too large). Results are illustrated in the form of ROC Equal Error Rate (EER), averaged over 5 independent trials. Independent of the object, learning machinery (Blue for GentleBoost, Red for SVM), or representation; the scale of the input image affects the detection rate in a predictable way.

complicated approach was chosen because of computational limits on high-dimensional data. Figure 4 illustrates the results of this experiment, again in terms of equal error rate, plotted against the results of the previous experiment. The solid horizontal line indicates the results of this experiment, that is, the mean EER of the classifier trained with features from multiple scales (dotted lines indicate the standard deviation of the 10 trials). The red line shows the classifier trained with HMAX features and the blue line HoG features, though they are not statistically different.

For each object tested, the classification score from the multi-scale approach outperforms the best score from a classifier trained at any single scale. These results support the assertion that complementary information from different scales can be leveraged to improve system-wide performance, even when the underlying image feature and statistical learning method are unchanged.

5.2 8 Fields of $\frac{1}{16}$-Resolution

It was shown above that a classifier with access to features from multiple scales can outperform the same machine with features from only the best single scale. A fair criticism is that there was more input to the multi-scale classifier. It had 3 128×128 fields whereas the single scale classifier had only one. In this experiment we apply the same methodology as above, but use 8 fields of 32×32 resolution to address this concern. We simply apply a suitable HoG-like algorithm to each 32×32 image independently, building a feature vector by concatenating the values from each scale. With the 8 scales used here, the Foveated Histogram of Gradients (FHoG) feature for this input produces 2592 total features.

For plates and pedestrians, the results of this experiment were not significantly different from the above multi-scale experiment, with mean EER $=$ .129 and .080, respectively. For monitors and cars the results were still better than the best single scale, at mean EER $=$ .105 and .088. This represents 12% fewer errors in the worst case (plates).

6 Summary and Next Steps

The contribution of this work is to clearly demonstrate the value of a multi-scale object-detection approach, when multi-scale information is available. It was first shown that targets can be detected across a broad range of scales, and that there is a preferred scale which is slightly larger the size of the object itself. The hypothesis that information from multiple scales is complementary was supported in Sec. 5.1 by training a classifier using features from several different scales. The hypothesis was bolstered further in Sec. 5.2 by using lower resolution images from each scale, and maintaining high levels of accuracy. Our next steps are to continue to critically explore the space of multi-scale image features, so as to design features which are both discriminative for a wide variety of object types, and computationally inexpensive.

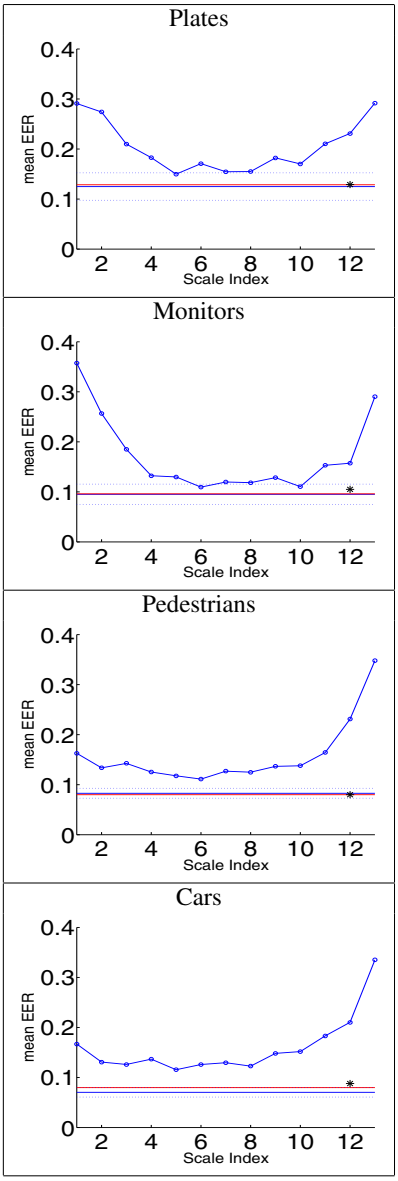


Figure 4. Single Scale vs. Multi-Scale. Detailed in Sec. 5.1, classifiers with multi-scale input are compared to the previous results from Fig. 3 (using HMAX and boosting). The solid lines indicate the mean EER of the multi-scale classifier. Red = HMAX, Blue = HoG. Dotted lines indicate std. dev. The black star indicates the results from the experiments of Sec. 5.2

References

[1] http://cbcl.mit.edu/software-datasets/streetscenes/.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2:886–893, 2005.

[4] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Stanford University Technical Report*, 1998.

[6] K. Fukushima. Neocognitron capable of incremental learning. *Neural Networks*, 17(1):37–46, January 2004.

[7] G. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Comp.*, 18(7):1527–1554, July 2006.

[8] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiology*, 160(1):106–154, 1962.

[9] Y. LeCun, Huang, and Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *CVPR*, 2004.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[11] R. Perko and A. Leonardis. Context driven focus of attention for object detection. In L. Paletta and E. Rome, editors, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint (WAPCV 2007)*, volume 4840, chapter 14, pages 216–233. Springer LNAI, December 2007.

[12] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[13] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 2007.

[14] H. Schneiderman and T. Kanade. A statistical model for 3d object detection applied to faces and cars. *CVPR*, 2000.

[15] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Object recognition with cortex-like mechanisms. *In: IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:411–426, 2007.

[16] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. *ICCV*, 1:370–377, 2005.

[17] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):153–167, 2003.

[18] L. Wolf and S. Bileschi. A critical view of context. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2006.