# The effective resolution of correlation filters applied to natural scenes

Michel Vidal-Naquet  Manabu Tanifuji
Riken, Brain Science Institute
Hirosawa 2-1, Wako Shi
Saitama 351-0198, Japan
michel@brain.riken.jp

## Abstract

*In this paper, we measure the responses of image patches, used as filters, on different image ensembles and examine how the responses are affected by reducing the resolution of the image ensembles. By comparing the set of responses obtained at high and reduced resolutions, we find that for the ensembles of natural and object images (cars), there is a limit resolution of about 15x15 and 10x10 pixels, respectively, beyond which the filter responses are significantly affected by resolution reduction. We support the result by a simple theoretical analysis based on image ensemble statistics.*

*There are two consequences to this result. First, it provides a natural working resolution, determined solely from the image ensemble statistics, to which higher resolution templates can be reduced without losing a significant amount of information. This can be used, in particular, to reduce the search space for useful visual features in many applications. Secondly, in contrast to many studies, it suggests that features that are more complex than Gabor patches can be effectively used as first layer filters and combined in order to represent more complex shapes and appearances.*

## 1. Introduction

In computer image analysis, useful visual information is generally extracted from an input image by first applying a set of filters, or features, in order to simplify the image representation and facilitate further high level tasks. Success of the feature extraction step is critical for applications such as object recognition or determining correspondences in pairs of images, and depends largely on the quality of the feature selection. While many approaches use low-level descriptors based on local operations that can be computed efficiently [16, 3, 6], other popular methods consist of using a simpler representation based on feature appearance, i.e. the two-dimensional gray-level matrix. Appearance based features that were tested throughout the literature vary greatly in size and complexity, from small and generic [12, 15, 23] to intermediate complexity image fragments, often selected with an exhaustive search [8, 21, 1, 18], and full-object templates [20]. In these approaches, computation of visual similarity is performed by convolution or Normalized Cross-Correlation (NCC). Reasons for their popularity include mathematical interpretation (projection onto a feature space), invariance to contrast changes (with NCC), and also because the features can be generic as well as class-specific. The fact that correlation type operations perform rigid, pixel-wise comparisons between image patches, however, forms a critical limitation of many of these methods. Natural variation in the appearance of visual objects and scenes suggests that, on the contrary, some degree of flexibility should be incorporated into the similarity measure. Several approaches that account for local appearance variability have therefore been proposed, such as decomposing rigid features into sub-features [4, 9], using global gray-level or local gradient histograms [17, 7], and the blurring of fine details by combining a small number of feature principal components [5].

In this paper, rather than understanding how to deal with local variability in images, we attempt to uncover the visual properties of image ensembles that can be effectively captured by rigid matching of patches, using NCC as a visual similarity measure. In particular, we aim to estimate a limit of visual detail, i.e. resolution, that is determinant for the set of filter responses, measured over some image ensembles, such as the ensembles of natural images, car images, or random noise images. For this, we compute a set of responses of square *filter patches* measured on a large set of randomly selected *test patches*, represented at some high resolution, and observe the evolution of the responses as the resolution of the filter and test patches is progressively reduced. Because low frequency components dominate the spectrum in natural and car images [13, 19], the response set of patches on natural and car image ensembles is not significantly af-

fected by the low-pass filtering up to some limit resolution beyond which the set of responses is significantly affected. The evolution of the response set is measured by comparing the responses at high resolution to the responses at lower resolutions with Normalized-Cross Correlation. In addition to simulations, we perform a theoretical analysis that shows how the limit depends on the spectrum of the image ensembles and supports the quantitative results. We find that for natural images and cars, the limit resolution is of about 15x15 and 10x10 pixels, respectively.

The rest of the paper is organized as follows. In Section 2, we describe the image ensembles used in our experiments. In Section 3, we explain how we compute the patch response sets and our method to determine the effective resolution of the patches. In Section 4, we provide a theoretical analysis that explains the evolution of the filter-patch response sets as the resolution is reduced. In Section 5, we show the simulation results demonstrating the resolution limit for natural images and cars. We also provide comparisons with ensembles of random images. Finally, in Section 6, we provide a discussion of the results and a conclusion.

## 2. Image ensembles used and patches

In this section, we describe the different image ensembles that we tested, the extraction of the patches and explain exactly what we mean by resolution.

### 2.1. Images Ensembles

In our experiments, we used an ensemble of 600 natural images, taken from the commonly used database introduced in [22], and a set of 1275 car images, mostly seen from the side, appearing at scales varying by a factor of about 2, with some clutter in the background. The car images were downloaded from the internet. The natural images were of size 1024x1536, and the car images were of about 60x82 pixels. Examples of these images are shown in Figure 1.

In addition, we also tested patches from an ensemble of random images, with a predefined Fourier amplitude spectrum of the form $f^{-0.5}$, where $f$ is the L2 norm of the 2-D frequency, in order to confirm our theoretical model (Section 4). Examples of patches, randomly selected from the different image ensembles and at different resolutions are shown in Figure 2.

### 2.2. Definition of resolution

We define the resolution of a patch $P_s$ by its size $s$, corresponding to the Nyquist frequency after filtering with a blurring Kernel $\mathbf{G}_s$. The bold font represents the Kernel in Fourier space. The low-pass blurring Kernel used is the Hanning window, commonly used for image sub-sampling.

Figure 1. Examples of images from the natural and car image sets used. Natural images are log-scaled for presentation.

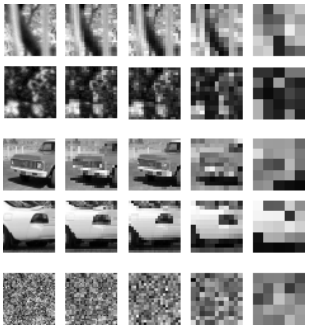

Figure 2. Examples of image patches extracted from natural scenes (top 2 lines), cars (lines 3 and 4) and random images with Fourier spectrum of the form $f^{-0.5}$ (lower line). The patches are shown at different resolutions along the horizontal axis, left to right: 41 (full resolution), 31, 21, 11 and 5 pixels per side.

The finest resolution was $s_0 = 41$, and we tested various resolution for $s$, down to $s = 5$.

For example, the top right patch shown in Figure 2, $P_5$, of size 5x5, was obtained by filtering and down-sampling $P_{41}$: $P_5 = \downarrow F^{-1}(\mathbf{P}_{41}.\mathbf{G}_5)$, where $F^{-1}$ is the inverse Fourier transform and $\downarrow$ is the down-sampling operator.

## 3. Patch responses and Resolution Similarity Function

For a given image ensemble, we want to understand how the responses of a filter patch, extracted from the ensemble and measured on a large number of patches from the same ensemble, are affected when the resolution of the patches is reduced. We first define the Patch Response Vector and then the Resolution Similarity Function (RSF), that measures the similarity between patch responses at different resolutions.

### 3.1. Patch Response Vector

The set of responses of a filter patch $P_s$, of resolution characterized by the size $s$, is represented by a vector $\mathbf{V}_s$ that we call the Patch Response Vector (PRV), whose entries $V_{i,s}$ are defined by equation (1):

$$V_{i,s} = P_s \otimes T_{i,s} = \downarrow (P_{s_0} * G_s) \otimes \downarrow (T_{i,s_0} * G_s) \quad (1)$$

where: $i \in [1..N]$, $\otimes$ represents the NCC operator, $*$ is the convolution operator, $T_{i,s}$ represents a test patch indexed by $i$, and $N$ is the number of test patches.

The test patches $T_{i,s}$ and the filter patch $P_s$ are from the same image ensemble. In our simulations, described in Section 5, we used $N = 2000$ test patches.

### 3.2. Resolution Similarity Function

To compare the responses of a patch for different resolutions, we define the Resolution Similarity Function (RSF), that is obtained by computing the normalized cross-correlations between the Patch Response Vector of the patch at the highest resolution $s_0 = 41$, and the Patch Response Vector of the patch obtained at some lower resolution $s$. The RSF is given by equation (2):

$$RSF(\mathbf{V}_{s_0}, \mathbf{V}_s) = \mathbf{V}_{s_0} \otimes \mathbf{V}_s \tag{2}$$

In Section 4 we provide a theoretical analysis of the RSF and in Section 5, we show quantitative simulation results for the different image ensembles.

## 4. Analysis of the Resolution Similarity Function

In this section, we provide a theoretical analysis of the RSF as a function of the resolution $s$.

We start by expressing the entries of the Patch Response Vectors as a function of the patch Fourier components and the Fourier components of the blurring Kernel $G_s$. We also assume the patches, in pixel space, are normalized to mean 0 and standard deviation 1, which is done a-priori in NCC matching:

$$V_{i,s} = P_s \otimes T_{i,s} = K \sum_j P_s^j . \overline{T_{i,s}^j} \tag{3}$$

$$= K \sum_j P_{s_0}^j . \overline{T_{i,s_0}^j} . |G_s^j|^2 \tag{4}$$

$$V_{i,s_0} = K \sum_j P_{s_0}^j . \overline{T_{i,s_0}^j} \tag{5}$$

Here, $K$ is a normalizing constant, $P_s^j$ and $T_{i,s}^j$ represent the complex Fourier component $j$ or the filter patch $P_s$ and the test path $T_{i,s}$ respectively, at resolution $s$. The sums are over all the complex Fourier components $j$. $\overline{x}$ is the conjugate operator. $|x|$ represents the norm. Eq. (4) results from the definition in eq. (1). We can ignore the down-sampling.

We can now express the RSF:

$$RSF(\mathbf{V}_{s_0}, \mathbf{V}_s) = \mathbf{V}_{s_0} \otimes \mathbf{V}_s \tag{6}$$

$$= \frac{\sum_i V_{i,s} . V_{i,s_0}}{\sqrt{\sum_i {V_{i,s}}^2 . \sum_i {V_{i,s_0}}^2}} \tag{7}$$

where the index $i$ goes over all the test patches.

In eq. (7), we assumed that the mean of $V_{i,s}$ and $V_{i,s_0}$ is 0. This is justified when the number of test patches is sufficiently large so that $V_{i,s}$ and $V_{i,s_0}$ cover uniformly the whole spectrum of responses, from -1 to 1.

We now consider the upper term of eq. (7) and replace $V_{i,s}$ and $V_{i,s_0}$ by their definition in eqs. (4) and (5):

$$\sum_i V_{i,s} . V_{i,s_0} = \tag{8}$$

$$\sum_i \sum_j P_{s_0}^j . \overline{T_{i,s_0}^j} . |G_s^j|^2 . \overline{\sum_j P_{s_0}^j . \overline{T_{i,s_0}^j}} \tag{9}$$

$$= \sum_{i,j} |P_{s_0}^j|^2 . |T_{i,s_0}^j|^2 . |G_s^j|^2 \tag{10}$$

In eq. (9), we used the fact that $V_{i,s_0} = \overline{V_{i,s_0}}$, since $V_{i,s_0}$ is a real number. In eq. (10), the fourth order terms representing the correlations between the components of different frequencies, that should appear, actually cancel out by assuming the different frequency components are not correlated over the image ensemble. This is not necessarily true, but we found that they are dominated by the terms present in eq. (10) (data not shown here). This assumption is also justified by the very good fit between the experimental curves and theoretical predictions shown in Section 5.

Finally, we assume that individual patches of the same ensemble have a Fourier amplitude spectrum close to the average amplitude spectrum over the ensemble, that we denote $\mathbf{Q}_{s_0}$ [13, 19]. This is an approximation for natural and car image patches, while it is exact for the $f^{-0.5}$ noise images. The final expression for $\sum_i V_{i,s} . V_{i,s_0}$ becomes:

$$\sum_i V_{i,s} . V_{i,s_0} = \sum_j {Q_{s_0}^j}^4 . |G_s^j|^2 \tag{11}$$

We can apply similar considerations for both sums in the denominator of eq. (7), and obtain the final expression of the RSF:

$$RSF(\mathbf{V}_{s_0}, \mathbf{V}_s) = \frac{\sum_j {Q_{s_0}^j}^4 . |G_s^j|^2}{\sqrt{\sum_j {Q_{s_0}^j}^4 . \sum_j {Q_{s_0}^j}^4 . |G_s^j|^4}} \tag{12}$$

Equation (12) entirely describes the behavior of the RSF as a function of the resolution (size), through the resolution dependent blurring Kernel $G_s$, and the average Fourier amplitude spectrum of the image ensemble, represented by $\mathbf{Q}_{s_0}$.

## 5. Experimental results

In this section, we present the simulation results. We show the RSF curves as a function of resolution (size) obtained experimentally for the different image ensembles, together with the RSF predicted by equation (12).

The results are summarized in Figure 3. The curves in the figure were obtained as follows. For a given image ensemble, 200 filter patches were selected randomly. For each filter patch, 2000 test patches were randomly chosen from the same ensemble. The RSF curve was then obtained according to equations (1) and( 2). The curves shown in the figure are the average curves over the 200 RSF curves obtained for the individual filter patches. To plot the theoretical predictions, we computed $Q_{s_0}$ by averaging the norm of the 2000 test patches Fourier amplitudes. We then plugged $Q_{s_0}$ and the spectrum of the filtering Kernel $\mathbf{G}_s$ into equation (12). 1-D cuts of the amplitude spectra $Q_{s_0}$, for the different image ensembles, are shown in Figure 4, illustrating in particular the stronger components at high frequencies in the noise images.

In Figure 3, we observe the graphs from right to left, and see that the curves for natural and car image ensembles remain very close to 1, up to the lower resolutions of $s = 15$ and $s = 10$ respectively, when the curves cross the RSF value of 0.99. Beyond these resolutions, the curves appear to drop more significantly. This indicates that the response properties of the patches are not affected by loss of visual detail until these lower limits. As shown in our theoretical analysis, the variations of the RSF, and thereby the values of the limit resolutions, are directly related to the form of the image ensemble amplitude spectrum, that is roughly of the form $f^{-1}$ for both cars and natural scenes [13, 19]. In contrast, the RSF of the $f^{-0.5}$ noise images falls off sharply at the onset of resolution reduction, because the low frequency components are less dominant. The RSF value is 1 for $s = 41$, since this is the original resolution of the patches.

While the RSF curves of individual filter patches do not have to be equal to one another, the deviations of the RSF values are necessarily very small up to the limit resolutions because the curves hold tightly to the maximum value of 1. Therefore, we omitted indications of variance for clarity.

We illustrate the resolution limits by also showing examples of filter-patches at different resolutions, with their individual RSF values, in Figures 5, 6 and 7. The different figures correspond to the three different image ensembles tested. Plotting the values displayed in these figures in the $(RSF(s), s)$ plane would yield an individual RSF curve. For example, from left to right, top to bottom in Figure 5, the RSF value of the natural filter-patch remains 1 (up to a precision of 0.01) until a resolution of $s = 15$, then it begins to drop. For the car filter-patch, the limit observed in Figure 6 is $s = 11$. For the $f^{-0.5}$ noise image patch, the decrease in the RSF value is immediate.

## 6. Discussion and conclusion

In the present study, we measured the responses of filter-patches to a large set of test-patches, extracted from within a same image ensemble, using Normalized Cross-Correlation as the matching measure. By comparing the set of responses obtained for high and reduced image patch resolutions, we showed, in simulations and by a simple theoretical calculation, how the set of patch responses depends on the image ensemble spectral statistics. In particular, we found that the filter-patches respond in an almost identical manner at high

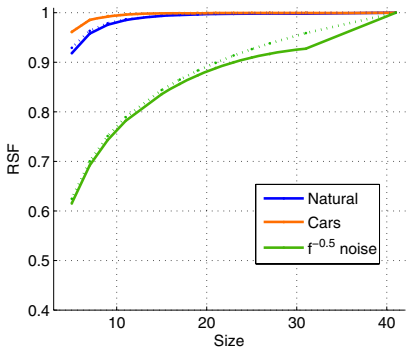


Figure 3. Average RSF curves for the different image ensembles tested. Blue: natural. Orange: cars. Green: random $f^{-0.5}$ noise. Solid lines represent the RSF curves obtained experimentally. Dotted lines indicate the theoretical curve predicted by eq. (12). The theoretical curves overlap with the experimental ones for the natural and car ensembles. The discrepancy between the theoretical curve and the experimental curve for the random $f^{-0.5}$ noise images for the higher resolutions is due to the fact we did not take into consideration the down-sampling. We observe that as the resolution is reduced, the RSF curves remain at about 1 for the natural and car images, until the resolution reaches about 15 pixels and 10 pixels respectively, beyond which the curves begin to drop. For the random images, in contrast, the RSF curve begins to drop at the onset of resolution reduction.

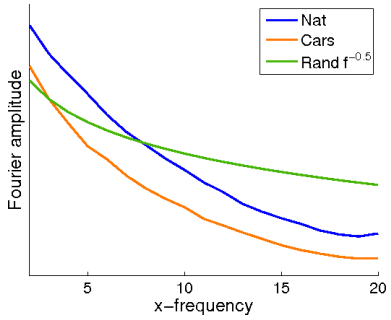


Figure 4. Cut of the log-amplitude spectra $\mathbf{Q_{s_0}}$ along the y-frequency 0, for the different image ensembles. Blue: natural. Orange: cars. Green: $f^{-0.5}$random noise. x-axis: horizontal frequencies. y-axis: log of the amplitude spectrum. The curve for the noise ensemble was translated vertically for presentation. The high frequencies are more dominant for the random images, implying the RSF function falls off more sharply as the resolution is reduced (Figure 3).

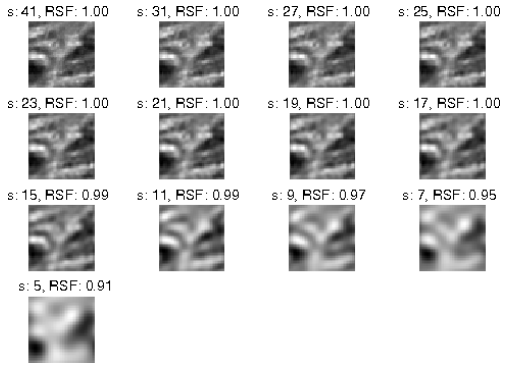


Figure 5. Examples of RSF values obtained for different sizes of a natural image patch filter. The RSF value is given for each size of the filter. For the display, the patches were not down-sampled. The RSF begins to drop after the resolution passes below a size of 15 pixels.

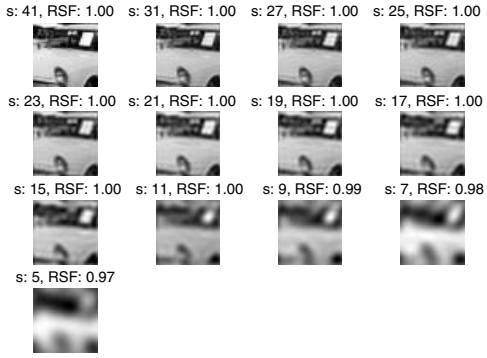


Figure 6. Car image ensemble. Examples of RSF values obtained for different sizes of a patch filter. The RSF drops when the resolution is below 11 pixels.

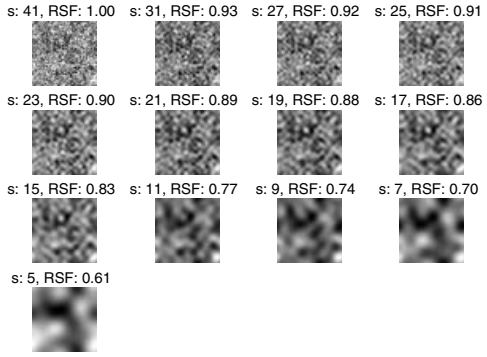


Figure 7. $f^{-0.5}$ noise images. Examples of RSF values obtained for different sizes of a $f^{-0.5}$ patch filter . We observe there is no resolution limit, the RSF value drops at the onset of resolution reduction.

resolutions (41x41 pixels) and at reduced resolutions, up to the limit of about 15x15 pixels for natural images and 10x10 pixels for the cars. In contrast, the responses of patches from a random noise image ensemble, where the lower frequency components of the images are less dominant, were significantly affected at the onset of resolution reduction.

First, the results suggest a basic lower limit resolution for the initial filter measurements in computer vision applications, because the responses of patches at higher resolutions, providing in principle more visual information, are actually fully explained by the responses of patches represented at the lower resolution limit. The resolution limits we found in our examples imply that features with a richer frequency content than Gabor type filters, derived by different encoding principles [10, 2], can be effectively used to extract information from the visual environment. Such features are shown in Figures 5 and 6. Reducing the resolution further leads to a deterioration of the responses and thereby a loss of information about the visual world. We can note that our result is in line with recent physiological studies, that suggest that primary visual cortex neurons may represent more complex stimuli than simple Gabors [14, 11].

Secondly, these results suggest that numerous computer vision applications using correlation based operators may benefit, in terms of computational complexity, from reducing the resolution of the filters down to the lower resolution limits that depend on the image statistics. More generally, the results suggest a method to incorporate prior knowledge about the image statistics, that can be used in learning applications. For example, in recognition methods that are based on selecting useful visual features with an exhaustive search through a large set of image patches, as in [21, 1, 18], the lower resolution limit can reduce the size of the feature search space.

We can mention several difficulties with our method that should be addressed. First, although we can clearly observe a cut-off in the car and natural image curves shown in Figure 3, it is not clear where to place the limit resolution exactly. In our experiments, we determined limit resolutions by an arbitrary RSF threshold of 0.99, but a more appropriate way would be to determine the threshold in an application dependent manner. Second, it is not clear whether the value of the initial feature size, 41x41 pixels here, will affect the RSF curves significantly. This is not expected, however, due to the scale-invariance properties of natural image ensembles [13]. Finally, the method for comparing the patch responses at different resolutions (eq. 2) may be critical for determining the resolution limits. For example, the comparison could be performed by computing the Mutual Information (MI) between the Patch Response Vectors (Section 3), rather than using NCC. MI can, in principle, capture more complex dependencies between random variables than can NCC. However, the use of MI would raise other issues, such as how to determine the binning. NCC is a straightforward method for comparison that captures the essential common

properties of two random vectors, and leads to a simple theoretical analysis.

In conclusion, we found a rough limit resolution for image patches when they are used as a first stage in the processing of natural scenes and object images. This limit depends solely on the spectral statistics of the image ensemble. Using higher resolution patches would not be more informative because their response properties do not differ significantly from the responses of patches at the lower limit resolution. We also found that the features of the limit resolution are more complex than simple Gabor type features, commonly used in computer vision applications, suggesting that more complex features can be effectively used in the first stages of visual processing. Finally, our results and analysis suggest that knowledge about image ensemble spectral statistics can be incorporated into applications requiring learning, such as object recognition, in order to simplify the computational complexity of the algorithm and facilitate the learning stage. This will be tested in future work.

## References


[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004. 1, 5

[2] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997. 5

[3] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proceedings of the International Conference on Computer Vision*, pages 634–640, 2003. 1

[4] B. Epshtein and S. Ullman. Feature hierarchies for object classification. In *Proceedings of the International Conference onComputer Vision*, pages 220–227, 2005. 1

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Conf on Computer Vision and Pattern Recognition*, 2003. 1

[6] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45, 2001. 1

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004. 1

[8] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981. 1

[9] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 11–18, 2006. 1

[10] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:560–561, 1996. 5

[11] B. Olshausen and D. Field. How close are we to understanding v1? *Neural Computation*, 17:1665–1699, 2005. 5

[12] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of International Conference on Computer Vision*, 1998. 1

[13] D. L. Ruderman and W. Bialek. Statistics of natural images:scaling in the woods. *Phys. Rev. Lett.*, 73:814–817, 1994. 1, 3, 4, 5

[14] N. C. Rust, O. Schwartz, J. Movshon, and E. Simoncelli. Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, 46:945–956, 2005. 5

[15] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recoognition. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, July 1998. 1

[16] S. Smith and J. Brady. SUSAN - a new approach to low level image processing. *Int. Journal of Computer Vision*, 23(1):45–78, May 1997. 1

[17] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. 1

[18] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. of the IEEE Conf on Computer Vision and Pattern Recognition*, pages 762–769, 2004. 1, 5

[19] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14:391–412, 2003. 1, 3, 4

[20] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991. 1

[21] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5:682–687, 2002. 1, 5

[22] J. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond. B*, (265):359–366, 1998. 2

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001. 1