

1 Statistical roots of the proposed research

1.1 Statistical modeling in vision

The availability of large amount of data, the richness and diversity of the structures underlying the data, and the difficulty of the tasks of understanding the data are the driving forces to advance the state of the art of statistical modeling, learning and computing, because they challenge conventional statistical thinking and force us to extrapolate from conventional statistical theories and methods. Vision is such a fertile field that holds such a promise.

The problem of vision is still a deep mystery. There are bewildering varieties of visual patterns in the natural scenes, including object patterns such as human figures, horses, birds, and texture patterns such as foliage, grasses, rivers, as well as the articulate and complex motions of these patterns. How does the biological visual system represent, learn, and recognize these patterns from image data collected by the retina? This problem seems effortlessly easy for biological vision, but proves to be extraordinarily difficult for computer vision.

In principle, the visual patterns can be represent by statistical models that characterize the regularities and variabilities of the image patches that correspond to these patterns, and developing statistical models for visual patterns is of central importance to understanding the problem of vision. In practice, image data for visual patterns are plentiful and just one-click away on google, and carefully annotated images are readily available from a number of large and free data bases [27] [65] [26] [43] [31] [7] [30] [78]. However, there are still very few realistic statistical models available for representing and modeling the rich and diverse image data [14] [4] [85] [73].

Finding statistical models for visual patterns is a great challenge but also a great opportunity for statisticians, who are equipped with a wealth of tools for modeling and computing. Statisticians have the potential to make fundamental and transformative contributions to solving this deep scientific problem. Advances in this area will surely enrich and deepen our knowledge of statistical modeling, learning and computing in general.

For the past 14 years, the two PIs have been working on searching for statistical models and computational algorithms for representing and recognizing visual patterns [72] [85] [81] [39], continuing the pattern theoretical approach pioneered by Grenander [38] and advocated by Mumford [50]. The PIs' work on Markov random field model [75] for texture patterns and the more recent active basis model [73] for object patterns have been recognized by the nominations for the David Marr prize in the International Conference on Computer Vision in 1999 and 2007 respectively. The proposed research is a continuation of the PIs' research program.

The rest of this section reviews our recent work on active basis model, which has mainly been supported by the NSF grant DMS-0707055. So it also serves as a partial review of results from *prior NSF support*.

1.2 Statistical root in linear regression

Modern linear regression. Modern linear regression problems are often characterized by the fact that the number of predictors or regressors is much greater than the dimensionality of the response vector. The active basis model is founded on such a linear regression structure. To fix notation, let $(\mathbf{I}(x), x = (x_1, x_2) \in D)$ be an image defined on a rectangular lattice or domain D , where $x = (x_1, x_2)$ indexes the pixels of \mathbf{I} . The linear regression model is of the following form:

$$\mathbf{I}(x) = \sum_{i=1}^n c_i B_i(x) + U(x), \quad (1)$$

where $(B_i(x), i = 1, \dots, n)$ are a small number of basis elements selected from a dictionary of such elements, $(c_i, i = 1, \dots, n)$ are the coefficients, and $U(x)$ is the unexplained residual image. In the

notation and terminology of linear regression, we may arrange $\mathbf{I}(x)$, $B_i(x)$ and $U(x)$ as column vectors (if the image is 100×100 , then each of \mathbf{I} , B_i , U is a 10,000 dimensional vector), so that \mathbf{I} is the response vector, and $(B_i, i = 1, \dots, n)$ are the predictor vectors or regressors.

The regressors $(B_i, i = 1, \dots, n)$ are selected from a large dictionary of regressors. In our work, the regressors are localized, elongated, and oriented Gabor wavelets [16].

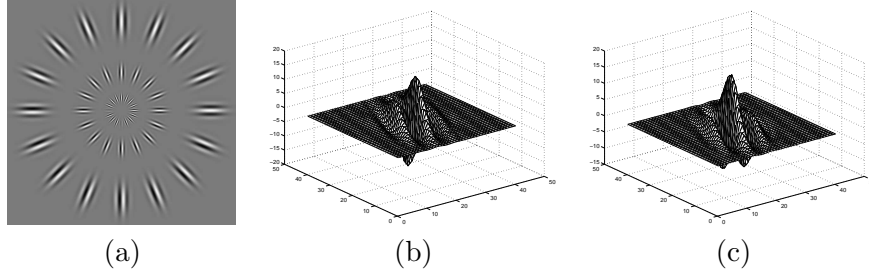


Figure 1: (a) A sample of Gabor wavelets at different locations, orientations, and scales. (b) A Gabor sine wavelet. (c) A Gabor cosine wavelet.

Dictionary of regressors. Figure 1(a) displays a sample of Gabor wavelet elements at different locations, orientations and scales. Figure 1(b) and (c) display two examples of Gabor wavelets. Specifically, the Gabor wavelets are translated, rotated, and dilated versions of the following function: $G(x_1, x_2) \propto \exp\{-[(x_1/\sigma_1)^2 + (x_2/\sigma_2)^2]/2\}e^{ix_1}$, which is a sine-cosine wave multiplied by a Gaussian function. This Gaussian function is elongated along the x_2 -axis, with $\sigma_2 > \sigma_1$, and the sine-cosine wave propagates along the shorter x_1 -axis. We truncate the function to make it locally supported on a finite rectangular range, so that it has a well defined length and width, and the function is 0 outside this rectangular range.

We can translate, rotate and dilate $G(x_1, x_2)$ to obtain a general form of the Gabor wavelets: $B_{x_1, x_2, s, \alpha}$, centered at $x = (x_1, x_2)$ and tuned to orientation α and scale s . $B_{x, s, \alpha} = (B_{x, s, \alpha, 0}, B_{x, s, \alpha, 1})$, where $B_{x, s, \alpha, 0}$ is the even-symmetric Gabor cosine component, and $B_{x, s, \alpha, 1}$ is the odd-symmetric Gabor sine component. We always use Gabor wavelets as pairs of cosine and sine components. We normalize both the Gabor sine and cosine components to have zero mean and unit ℓ_2 norm.

For an image $\mathbf{I}(x)$, with $x \in D$, we can project it onto a Gabor wavelet $B_{x, s, \alpha, \eta}$, $\eta = 0, 1$. The projection of \mathbf{I} onto $B_{x, s, \alpha, \eta}$, or the Gabor filter response at (x, s, α) , is $\langle \mathbf{I}, B_{x, s, \alpha, \eta} \rangle = \sum_{x'} \mathbf{I}(x') B_{x, s, \alpha, \eta}(x')$. We write $\langle \mathbf{I}, B_{x, s, \alpha} \rangle = (\langle \mathbf{I}, B_{x, s, \alpha, 0} \rangle, \langle \mathbf{I}, B_{x, s, \alpha, 1} \rangle)$, and the local energy is computed by $|\langle \mathbf{I}, B_{x, s, \alpha} \rangle|^2 = \langle \mathbf{I}, B_{x, s, \alpha, 0} \rangle^2 + \langle \mathbf{I}, B_{x, s, \alpha, 1} \rangle^2$.

The dictionary of Gabor wavelets is $\Omega = \{B_{x, s, \alpha}, \forall (x, s, \alpha)\}$. We can discretize the orientation so that $\alpha \in \{o\pi/O, o = 0, \dots, O-1\}$, that is, O equally spaced orientations (the default value of O is 15 in our experiments).

Over-completeness and sparsity. The dictionary Ω is called “over-complete” because the number of wavelet elements in Ω is larger than the number of pixels in the image domain, since at each pixel x , there can be many wavelet elements $B_{x, s, \alpha}$ tuned to different orientations α and scales s . In statistics, this is often called the “small n and large p” problem [11].

For an image $(\mathbf{I}(x), x \in D)$ we seek to represent it by

$$\mathbf{I}(x) = \sum_{i=1}^n c_i B_{x_i, s, \alpha_i}(x) + U(x), \quad (2)$$

where, corresponding to Equation (1), $B_i(x) = B_{x_i, s, \alpha_i}(x) \in \Omega$. In the experiments described in this proposal, we mostly fixed the scale parameter s (e.g., the length of the Gabor wavelets is 17 pixels). Even though each B_{x_i, s, α_i} has the same size as the image \mathbf{I} , B_{x_i, s, α_i} is non-zero only on a small rectangular support, so we may consider each B_{x_i, s, α_i} to be a “stroke” for “sketching”

the image \mathbf{I} . Thus the linear regression model (2) translates an image with a large number (e.g., 100×100) of pixels into a sketch of a small number (e.g., 50) of strokes, and these strokes should capture geometric information in image \mathbf{I} .

Variable selection. The set of wavelet elements $\mathbf{B} = (B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ can be selected from Ω by the matching pursuit algorithm [48], which seeks to minimize $\|\mathbf{I} - \sum_{i=1}^n c_i B_{x_i, s, \alpha_i}\|^2$ by the following greedy scheme:

- [0] Initialize $i \leftarrow 0$, $U \leftarrow \mathbf{I}$.
- [1] Let $i \leftarrow i + 1$. Let $(x_i, \alpha_i) = \arg \max_{x, \alpha} |\langle U, B_{x, s, \alpha} \rangle|^2$.
- [2] Let $c_i = \langle U, B_{x_i, s, \alpha_i} \rangle$. Update $U \leftarrow U - c_i B_{x_i, s, \alpha_i}$.
- [3] Stop if $i = n$, else go back to 1.

In the experiments presented in this proposal, we hand-pick n . n can be selected by principled criteria.

The matching pursuit algorithm has been widely used in signal processing. It is actually the forward selection algorithm for variable selection in linear regression. One may adopt penalized least squares methods such as lasso [63] or basis pursuit [12], which can be solved by LARS [25]. See also [8] [57] [11]. One may also adopt a Bayesian approach by assuming a mixture of Gaussian prior on the coefficients [37] [13] [69]. In our work, we find that matching pursuit works well. The reason is that the Gabor wavelets are small relative to the size of the image \mathbf{I} , so each B_{x_i, s, α_i} only explains away a small piece of \mathbf{I} , thus matching pursuit is not as greedy as it appears.

Primary visual cortex. Why do we use a dictionary of localized, elongated and oriented wavelet elements such as Gabor wavelets as regressors? The answer is that they give sparse coding of natural images, which contain edges as the most prominent structures, and edges can be efficiently represented by such wavelets. A formal mathematical justification is given by the work of Donoho and Candes (1999) on curvelets [9]. A statistical justification is given by Olshausen and Field (1996) [51]. They collect a large sample of image patches of natural scenes, and then learn the dictionary of basis elements by minimizing a lasso-like criterion over both the coefficients and the basis elements. One may also integrate out the coefficients in a Bayesian treatment [44] [52]. The learned basis elements closely resemble the Gabor wavelets. The learning bears some similarity to independent component analysis [5] and factor analysis, but also has significant differences in terms of over-completeness and sparsity. Olshausen and Field (1996) propose that sparse coding in the form of the model (2) is used by the primary visual cortex, where the regressors $\{B_{x, s, \alpha}\}$ correspond to simple cells in the primary visual cortex.

Now the question is, what is the purpose of sparse coding and what is beyond the linear regression model (2)?

1.3 Active basis model

A simple exercise may offer an answer to the above question. Suppose we want to represent image patches of objects of the same category, and for simplicity let us assume for the present that these image patches are defined on the same lattice, and the objects in these images appear at the same location, scale, and pose. See, for instance, the three deer images in Figure (2.b). Then let us consider what happens if we pursue a sparse coding for these images simultaneously.

Multiple response vectors sharing common regressors. To fix notation, let $\{\mathbf{I}_m, m = 1, \dots, M\}$ be the set of training images defined on a common lattice D . Since the objects in \mathbf{I}_m are from the same category, we may want to represent them by $\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_i, s, \alpha_i} + U_m$, where the multiple response vectors $\{\mathbf{I}_m\}$ share the same set of regressors $\mathbf{B} = (B_{x_i, s, \alpha_i}, i = 1, \dots, n)$. This \mathbf{B} can be considered a common template for the training images. Because there can be shape deformations in the objects, we may allow the basis elements in \mathbf{B} to perturb their locations and orientations. We call such \mathbf{B} an active basis, which is a mathematical model for a deformable

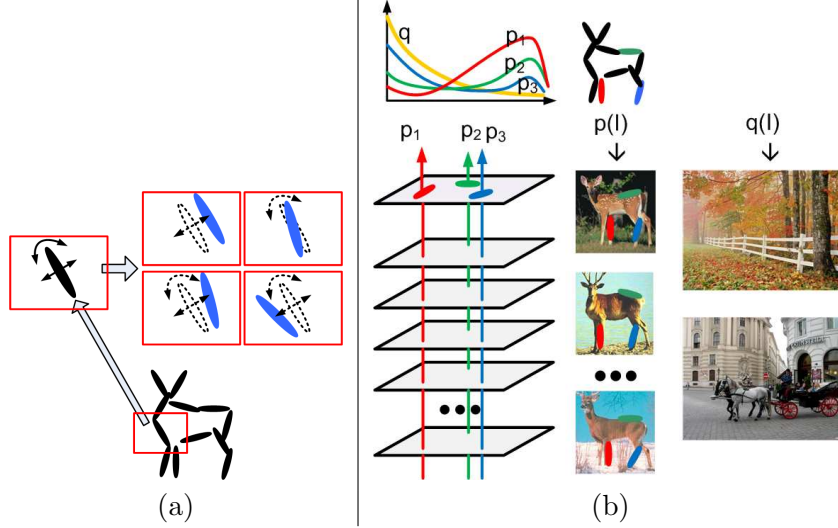


Figure 2: (a) An active basis template $\mathbf{B} = (B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ of a deer, where each Gabor wavelet element B_{x_i, s, α_i} is illustrated by an elongated ellipsoid. An element B_{x_i, s, α_i} (black ellipsoid) can slightly shift its location and orientation and change to $B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}$ (blue ellipsoids) for coding image \mathbf{I}_m . (b) The elements of the active basis \mathbf{B} are shared by all the training images $\{\mathbf{I}_m, m = 1, \dots, M\}$ of deer, subject to local perturbations $(\Delta x_{m,i}, \Delta \alpha_{m,i}, i = 1, \dots, n)$ that deform the active basis template \mathbf{B} . The elements are selected in the order of the Kullback-Leibler divergence between the foreground distribution p_i of the Gabor filter responses pooled from training images of deer, and the background distribution q pooled from the two natural images of rural and urban scenes.

template [3]. The linear regression model then becomes

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} + U_m, \quad m = 1, \dots, M. \quad (3)$$

For each image \mathbf{I}_m , the wavelet element B_{x_i, s, α_i} is perturbed to $B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}$, where $\Delta x_{m,i}$ is the perturbation in location, and $\Delta \alpha_{m,i}$ is the perturbation in orientation. $\mathbf{B}_m = (B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}, i = 1, \dots, n)$ is the deformed template for representing image \mathbf{I}_m .

Sparse coding for generalization. We call $(\Delta x_{m,i}, \Delta \alpha_{m,i}, i = 1, \dots, n)$ the activities or perturbations of the basis elements for image m . The sparse coding in terms of Gabor wavelets enables us to generalize to similar shapes by perturbing the parameters of the Gabor wavelets, i.e., locations, orientations, and coefficients. Let $A(\alpha) = \{(\Delta x = (d \cos \alpha, d \sin \alpha), \Delta \alpha) : d \in [-b_1, b_1], \Delta \alpha \in [-b_2, b_2]\}$ be the set of all possible activities for a basis element tuned to orientation α (default values: $b_1 = 6$ pixels, and $b_2 = \pi/15$). Figure (2.a) illustrates an active basis template of a deer, where each basis element is illustrated by an elongated ellipsoid.

Extended matching pursuit. We can pursue the active basis template \mathbf{B} by minimizing the following least squares criterion $\sum_{m=1}^M \|\mathbf{I}_m - \sum_{i=1}^n c_{m,i} B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}\|^2$.

- [0] Initialize $i \leftarrow 0$. For $m = 1, \dots, M$, initialize $U_m \leftarrow \mathbf{I}_m$.
- [1] $i \leftarrow i + 1$. Select $(x_i, \alpha_i) = \arg \max_{x, \alpha} \sum_{m=1}^M \max_{(\Delta x, \Delta \alpha) \in A(\alpha)} |\langle U_m, B_{x + \Delta x, s, \alpha + \Delta \alpha} \rangle|^2$.
- [2] For $m = 1, \dots, M$, retrieve $(\Delta x_{m,i}, \Delta \alpha_{m,i}) = \arg \max_{(\Delta x, \Delta \alpha) \in A(\alpha_i)} |\langle U_m, B_{x_i + \Delta x, s, \alpha_i + \Delta \alpha} \rangle|^2$. Let $c_{m,i} \leftarrow \langle U_m, B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} \rangle$, and update $U_m \leftarrow U_m - c_{m,i} B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}$.
- [3] Stop if $i = n$, else go back to 1.

Primary visual cortex. In Step 1 and Step 2 of the above algorithm, we need to perform a local maximization over $(\Delta x, \Delta \alpha) \in A(\alpha)$. Interestingly, Riesenhuber and Poggio [56] observe that complex cells of the primary visual cortex appear to perform local maximum pooling of the responses from simple cells. From the perspective of the active basis model, this corresponds to

estimating the activities $(\Delta x_{m,i}, \Delta \alpha_{m,i})$. Therefore, if we are to believe Olshausen and Field's theory on wavelet sparse coding [51] and Riesenhuber and Poggio's theory on local maximum pooling [56], then the active basis model is a compelling logical consequence if we want to reverse engineer the generative model.

1.4 Statistical root in exponential family distributions

There is yet one more issue to be fixed. The above algorithm is driven by the least squares criterion, which implicitly assumes (or works best under the assumption) that U_m is Gaussian white noise, and that $c_{m,i}$ follows a diffused prior distribution. This is clearly a wrong assumption because there are strong edges in the background, and the distribution of $\{c_{m,i}, m = 1, \dots, M\}$ should be estimated from the training images in a spirit similar to empirical Bayes and linear mixed model. So we need more sophisticated statistical modeling.

Density substitution. To simplify notation, let $x_{m,i} = x_i + \Delta x_{m,i}$, and $\alpha_{m,i} = \alpha_i + \Delta \alpha_{m,i}$, and $B_{m,i} = B_{x_{m,i}, s, \alpha_{m,i}}$. We can write the deformed template $\mathbf{B}_m = (B_{m,i}, i = 1, \dots, n)$. For simplicity, we assume that the basis elements in the deformed template $(B_{m,i}, i = 1, \dots, n)$ are orthogonal to each other, or in other words, they do not overlap (in practice, we allow small overlap). This is often the case for the linear representation produced by the matching pursuit algorithm. Then $c_{m,i} = \langle \mathbf{I}_m, B_{m,i} \rangle$, and U_m lies in the subspace that is orthogonal to \mathbf{B}_m . Let $C_m = (c_{m,i}, i = 1, \dots, n)$, and with slight abuse of notation, we also use U_m to denote the coordinate of \mathbf{I} in the subspace orthogonal to \mathbf{B}_m . Then $p(\mathbf{I}_m | \mathbf{B}_m) = p(C_m)p(U_m | C_m)$, where the linear mapping between \mathbf{I}_m and (C_m, U_m) is orthogonal. $p(C_m)$ can be estimated from the training images. To model $p(U_m | C_m)$, we introduce a reference model $q(\mathbf{I})$, so that $q(\mathbf{I}_m) = q(C_m)q(U_m | C_m)$ under the same linear mapping. We assume that $p(U_m | C_m) = q(U_m | C_m)$. Then $p(\mathbf{I}_m | \mathbf{B}_m) = q(\mathbf{I}_m)p(C_m)/q(C_m)$. This is a density substitution scheme that has been used by Friedman (1987) [33] for projection pursuit density estimation. Such a scheme also works if C_m is non-orthogonal, or non-linear, or even a discrete reduction of \mathbf{I}_m [74].

The least squares criterion implicitly assumes a Gaussian white noise $q(\mathbf{I})$, under which $(c_{m,i}, i = 1, \dots, n)$ are independent for orthogonal \mathbf{B}_m . The problem of Gaussian white noise is that it cannot account for strong edges in the background $q(U_m | C_m)$. So we assume a $q(\mathbf{I})$ that reproduces the marginal distribution of $c = \langle \mathbf{I}, B_{x,s,\alpha} \rangle$ in natural images, while still maintaining the independence of $(c_{m,i}, i = 1, \dots, n)$, at least approximately. Such a $q(\mathbf{I})$ has been explicitly constructed by the co-PI and Mumford [82]. Let $q(c)$ be this marginal distribution, which can be pooled from natural images, such as the two images of rural and urban scenes in Figure (2), see also [60] [61]. Under $q(c)$, $r = |c|^2$ has a very long tail, reflecting the fact that there are strong edges in natural images or residual background U_m . We then further assume that given \mathbf{B}_m , $(c_{m,i}, i = 1, \dots, n)$ are also independent under $p(C_m)$. This gives us the following model:

$$p(\mathbf{I}_m | \mathbf{B}_m) = q(\mathbf{I}_m) \prod_{i=1}^n \frac{p_i(c_{m,i})}{q(c_{m,i})}. \quad (4)$$

Figure (2.b) illustrates this idea, where p_i and q in this figure are the distributions of $r = |c|^2$ under $p_i(c)$ and $q(c)$ respectively.

Exponential tilting. We further parametrize $p_i(c)$ to be the following exponential family distribution:

$$p(c; \lambda) = \frac{1}{Z(\lambda)} \exp\{\lambda h(|c|^2)\} q(c), \quad (5)$$

where $\lambda > 0$ is the parameter. Let $r = |c|^2$, $Z(\lambda) = \int \exp\{\lambda h(r)\} q(c) dc = \mathbb{E}_q[\exp\{\lambda h(r)\}]$ is the normalizing constant, and $\mu(\lambda) = \mathbb{E}_\lambda[h(r)]$ is the mean parameter. $h(r)$ is a monotone increasing function. We assume $p_i(c) = p(c; \lambda_i)$.

Sufficient statistics. We use the following function for the sufficient statistics $h(r)$:

$$h(r) = \xi \left[\frac{2}{1 + e^{-2r/\xi}} - 1 \right]. \quad (6)$$

$h(r)$ behaves like $h(r) = r$ for small r , but $h(r) \rightarrow \xi$ (e.g., $\xi = 6$) as $r \rightarrow \infty$. The reason for such a transformation that saturates at a fixed value for large r is as follows. Let $q(r)$ be the distribution of $r = |c|^2 = |\langle \mathbf{I}, B \rangle|^2$ under $q(c)$ where $\mathbf{I} \sim q(\mathbf{I})$. We may implicitly model $q(r)$ as a mixture of $p_{\text{on}}(r)$ and $p_{\text{off}}(r)$, where p_{on} is the distribution of r when B is on an edge in \mathbf{I} , and p_{off} is the distribution of r when B is not on an edge in \mathbf{I} . $p_{\text{on}}(r)$ has a much heavier tail than $p_{\text{off}}(r)$. Let $q(r) = (1 - \rho_0)p_{\text{off}}(r) + \rho_0 p_{\text{on}}(r)$, where ρ_0 is the proportion of edges in the natural images. Similarly, let $p_i(r)$ be the distribution of $r = |c|^2$ under $p_i(c)$. We can model $p_i(r) = (1 - \rho_i)p_{\text{off}}(r) + \rho_i p_{\text{on}}(r)$, where $\rho_i > \rho_0$, that is, the proportion of edges sketched by the selected basis element B_i is higher than the proportion of edges in the natural images. Then, as $r \rightarrow \infty$, $p_i(r)/q(r) \rightarrow \rho_i/\rho_0$, which is a constant.

1.5 Learning and inference algorithms

Learning algorithm. The learning algorithm is very similar to the extended matching pursuit algorithm of Subsection (1.3), except that it is a greedy scheme for maximizing the likelihood of the non-Gaussian model (4) and (5), instead of the least squares criterion. Specifically, we want to pursue the common template \mathbf{B} and deform it to \mathbf{B}_m for each \mathbf{I}_m , so that there is a big contrast between the distribution of $\{c_{m,i}, m = 1, \dots, M\}$ and the marginal distribution $q(c)$. This contrast, or more specifically, the Kullback-Leibler divergence, is monotone in $\sum_{m=1}^M h(|c_{m,i}|^2)$, which serves as the pursuit index, similar to the pursuit index of the projection pursuit. This index essentially counts the number of edges sketched by B_i .

[0] Initialize $i \leftarrow 0$. For $m = 1, \dots, M$, initialize $R_m(x, \alpha) \leftarrow \langle \mathbf{I}_m, B_{x,s,\alpha} \rangle$ for all (x, α) .

[1] $i \leftarrow i + 1$. Select

$$(x_i, \alpha_i) = \arg \max_{x, \alpha} \sum_{m=1}^M \max_{(\Delta x, \Delta \alpha) \in A(\alpha)} h(|R_m(x + \Delta x, \alpha + \Delta \alpha)|^2).$$

[2] For $m = 1, \dots, M$, retrieve

$$(\Delta x_{m,i}, \Delta \alpha_{m,i}) = \arg \max_{(\Delta x, \Delta \alpha) \in A(\alpha_i)} |R_m(x_i + \Delta x, \alpha_i + \Delta \alpha)|^2.$$

Let $c_{m,i} \leftarrow R_m(x_i + \Delta x_{m,i}, \alpha_i + \Delta \alpha_{m,i})$, and update $R_m(x, \alpha) \leftarrow 0$ if $|\langle B_{x,s,\alpha}, B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} \rangle|^2 > \epsilon$. Then estimate $\hat{\lambda}_i = \mu^{-1}(\sum_{m=1}^M h(|c_{m,i}|^2)/M)$.

[3] Stop if $i = n$, else go back to 1.

See Figure (2) for illustration. We allow small overlap or correlation between $(B_{m,i}, i = 1, \dots, n)$ (e.g., $\epsilon = .1$). The maximum likelihood estimation of λ_i only involves translating the mean parameter back to the natural parameter by $\mu^{-1}()$.

Figure (3) illustrates the results of the learning algorithm. Figure (4) illustrates the learning process.

Inference algorithm. After learning the template $\mathbf{B} = (B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ and estimating $\Lambda = (\lambda_i, i = 1, \dots, n)$, we can use the learned deformable template to find the object in a testing image \mathbf{I} , by fitting the following model: $\mathbf{I} = \sum_{i=1}^n c_i B_{x+x_i+\Delta x_i, s, \alpha_i+\Delta \alpha_i} + U$, where the location of the object, x , is unknown. The maximum likelihood estimation of the location of the object is accomplished by the following inference algorithm:

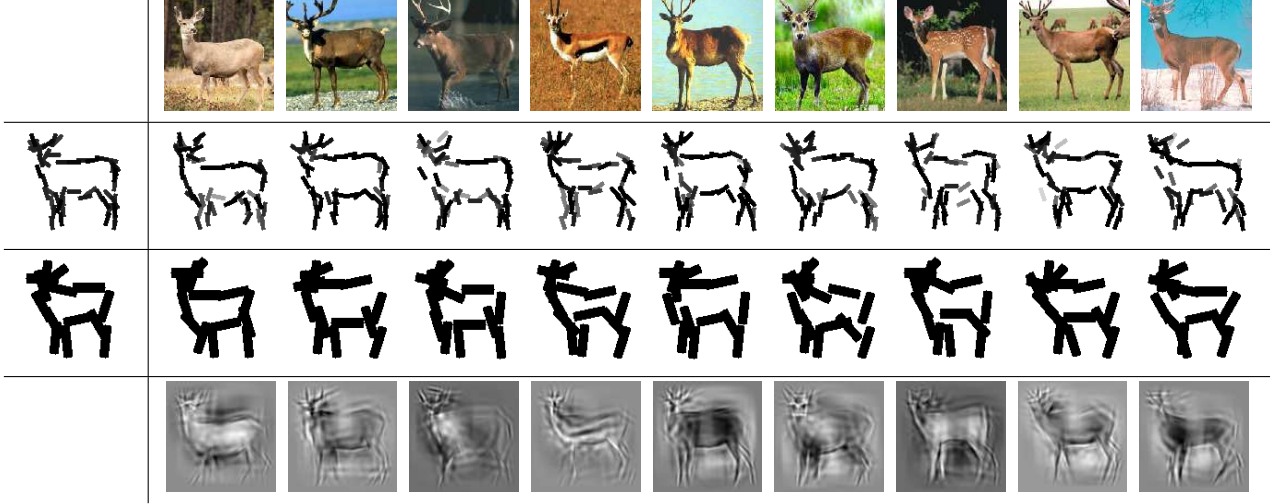


Figure 3: Learning the active basis model. Each Gabor wavelet element is illustrated by a bar with the same location, orientation, and length as the element. The first row displays $\{\mathbf{I}_m, m = 1, \dots, M = 9\}$. The second row: the first plot is the active basis template $\mathbf{B} = (B_i = B_{x_i, s, \alpha_i}, i = 1, \dots, n = 50)$. The rest of the plots are the deformed templates $\mathbf{B}_m = (B_{m,i} = B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}})$. The third row: the same as the second row, except that s is about twice as large, and $n = 14$. The last row displays the linear reconstruction $\mathbf{I}_m^{\text{syn}} = \sum_{i=1}^n c_{m,i} B_{m,i}$, where $n = 100$, and $(B_{m,i}, i = 1, \dots, n)$ contains Gabor wavelet elements at multiple scales.

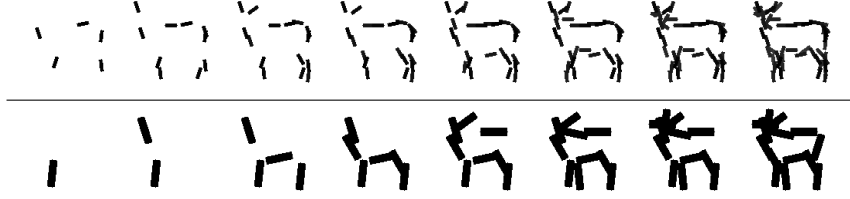


Figure 4: The learning algorithm sequentially selects the elements of the active basis $\mathbf{B} = (B_i, i = 1, \dots, n)$. The first row displays the learned template at the smaller scale with $n = 5, 10, 15, 20, 25, 30, 40, 50$. The second row: $n = 1, 2, 4, 6, 8, 10, 12, 14$ at the larger scale.

[1] For every pixel x , compute the log-likelihood ratio (foreground p versus background q) of x ,

$$l(x) = \sum_{i=1}^n \left[\lambda_i \max_{(\Delta x, \Delta \alpha) \in A(\alpha_i)} h(|\langle \mathbf{I}, B_{x+x_i+\Delta x, s, \alpha_i+\Delta \alpha} \rangle|^2) - \log Z(\lambda_i) \right].$$

[2] Find the MLE of x : $\hat{x} = \arg \max_x l(x)$. For $i = 1, \dots, n$, retrieve

$$(\Delta x_i, \Delta \alpha_i) = \arg \max_{(\Delta x, \Delta \alpha) \in A(\alpha_i)} |\langle \mathbf{I}, B_{\hat{x}+x_i+\Delta x, s, \alpha_i+\Delta \alpha} \rangle|^2.$$

[3] Return the location \hat{x} , and the translated and deformed template $(B_{\hat{x}+x_i+\Delta x_i, s, \alpha_i+\Delta \alpha_i}, i = 1, \dots, n)$.

Figure (5) shows two examples of inference. In each example, we search over multiple resolutions of the testing image because the scale of the object in the testing image is unknown. The resolution that achieves the maximum log-likelihood score is selected.

Cortex-like structure. The computation of $l(x)$ in Step 1 of the above inference algorithm can be accomplished by the following three steps:

[1] For all (x, α) , compute $\text{SUM1}(x, \alpha) = h(|\langle \mathbf{I}, B_{x, s, \alpha} \rangle|^2) = h(|\sum_{x'} \mathbf{I}(x') B_{x, s, \alpha}(x')|^2)$.



Figure 5: Inference. In each block, the left is the testing image \mathbf{I} , and the right is the translated and deformed template $(B_{\hat{x}+x_i+\Delta x_i, s, \alpha_i+\Delta \alpha_i}, i = 1, \dots, n)$.

[2] For all (x, α) , compute $\text{MAX1}(x, \alpha) = \max_{(\Delta x, \Delta \alpha) \in A(\alpha)} \text{SUM1}(x + \Delta x, \alpha + \Delta \alpha)$.

[3] For all x , compute $\text{SUM2}(x) = \sum_{i=1}^n [\lambda_i \text{MAX1}(x + x_i, \alpha_i) - \log Z(\lambda_i)]$. $l(x) = \text{SUM2}(x)$.

These three steps can be implemented by a cortex-like structure, which computes the SUM1 maps, MAX1 maps, and SUM2 maps consecutively. Step 1 corresponds to the simple cells of the primary visual cortex or V1 [51]. Step 2 corresponds to the complex cells of V1 [56]. Step 3 is a consequence of our active basis model. One may hypothesize that it corresponds to cells beyond V1.

1.6 Novelty in statistical modeling and computing

The active basis model is novel in statistical modeling and computing in the following aspects. (1) It allows the perturbations in the regressors to account for the variabilities in the response vectors. (2) It models regression coefficients by non-Gaussian distributions. (3) It uses non-Gaussian model for residuals. (4) The learning algorithm combines key features of both matching pursuit for explaining away individual images and projection pursuit for estimating the probability distribution of the whole training sample of images. (5) It provides a new perspective on sparse coding. In addition to denoising [19], compression [21], compressed sensing [10], here sparse coding is used for generalization by varying the parameters of the sparse coding elements.

Reproducibility. We have worked hard to develop reproducibility webpages for our research, and will continue to do so. In particular, the following webpage contains a wealth of data and source code: <http://www.stat.ucla.edu/~ywu/ActiveBasis.html>

2 Proposed Projects

The proposed research is based on the active basis model, but is by no means an incremental continuation of it. The proposed activities go far beyond the active basis model. They include an adventure in a completely new sparse land, an investigation of a most fundamental issue in learning, and a unification of two popular approaches to image representation and description.

2.1 A new sparse land of hyper regressors

What is beyond wavelets, edgelets, curvelets etc? Donoho and collaborators perhaps have done more than any others in developing sparse representations based on linear elements such as wedgelets [18], edgelets [20], curvelets [9] and beamlets [41]. These representations take advantage of the fact that natural images or images of geometric shapes mostly contain edges at different scales. The question is: What is beyond these elements? In an analogy to language, if these elements as “letters,” then what are the “words” so that these “words” lead to even sparser representations?

Artists’ intuition. The question may have already been answered by artists. Figure (6) shows three examples taken from two recent books on teaching children how to draw animals and other objects by sketching a very small number of elementary geometric shapes [42] [64]. In particular, the first row displays the steps of drawing a horse using an ellipsoid for the body and parallel bars

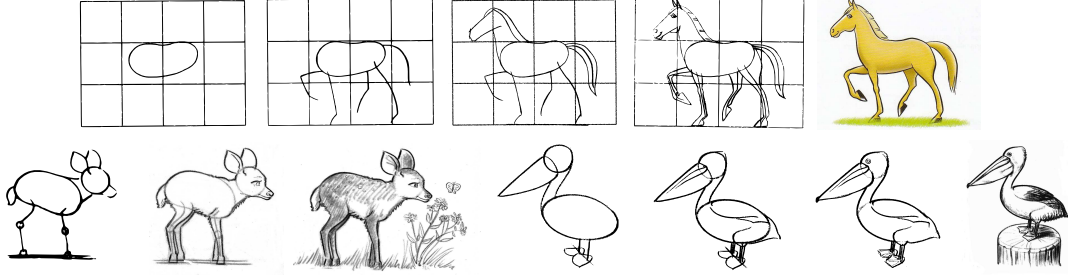


Figure 6: Drawing animals using a very small number of elementary geometric shapes [42][64].

for the legs and so on. The second row displays the drawing of deer and pelican, where for each animal, the first plot illustrates a sparse representation based on elementary shapes. The artists’ amazing intuition is essentially a highly sparse and symbolic representation of animal shapes. We call such a representation the “shape script,” a term coined by the co-PI in his previous work [23]. A shape script is a highly sparse and symbolic representation that can be extremely useful for learning and inference of object patterns because it captures essential dimensions of the object shapes.

Active basis model for elementary shapes. Interestingly, the shape script can still be described in the linear regression framework, except that the regressors are not Gabor wavelet elements, but elementary shapes, such as ellipsoids, angles, parallel bars, etc. Also interestingly, these elementary shapes can be described by active basis templates. We call such elementary active bases hyper regressors, which are compositions of the original regressors of Gabor wavelet elements.



Figure 7: The horse body and head can be detected and represented by the active basis models for ellipsoids. The first block displays the observed image. In each of the remaining two blocks, the first plot is the designed template. The second image is superposed with the translated template before deformation. The third image is superposed with the translated and deformed template.

Figure (7) illustrates the idea of elementary active bases. The first block displays the observed image of a horse. We design an image of a horizontal ellipsoid. Then we train the active basis model on this image using the learning algorithm. The first image in the second block shows the learned template. We then match the template to the observed image using the inference algorithm. The second image in this block shows the translated template at the detected location before deformation. The third image in this block shows the translated and deformed template that sketches the body of the horse. The third block shows that a vertical ellipsoid template can represent the head of the horse. It is important for the ellipsoids to be represented by the active basis models so that the templates can deform to fit the parts of the object.

Hyper regressors with hyper parameters. An active basis model can be written in the following form $\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_{m,i}, s, \alpha_{m,i}} + U_m = C_m \mathbf{B}_m + U_m$, where \mathbf{B}_m is a deformed template that is composed of Gabor wavelet elements. By a recursion of the above model, we propose to learn the following model that is a composition of K hyper regressors:

$$\mathbf{I}_m = \sum_{k=1}^K C_{m,k} \mathbf{B}_{x_{m,k}, s_{m,k}, \rho_{m,k}, \alpha_{m,k}}^{(t_k)} + U_m, \quad (7)$$

where t_k is the type of the k -th hyper regressor (e.g., ellipsoid, angle, parallel bars, etc.), which

is endowed with its hyper parameters, such as the overall location x , scale s , aspect ratio ρ and orientation α . Similar to the active basis, we may allow perturbations of these hyper parameters, and the perturbations can be quite large because the sizes of the elementary shapes are much larger than the Gabor wavelets. Such perturbations cause global deformations of the hyper regressors, so that the model is more capable of modeling large deformations and articulations of object shapes. In addition, on top of the hyper parameters, we also allow the perturbations of the location, scale, and orientation parameters of the Gabor elements that belong to each hyper regressor. This causes the local deformations of the hyper regressor. So model (7) is a recursive compositional model [36] [83].

Cortex-like structure. The inference of model (7) can be accomplished by a cortex-like structure in the form of recursive sum-max maps, which involves maximization over parameters of the Gabor elements for inferring local deformations as well as maximization over hyper-parameters for inferring global deformations [74].

We propose to design elementary geometric shapes as hyper regressors. We also propose to learn the hyper regressors from natural images. This is similar to learning object parts, but sparsity must be aggressively enforced. In addition, we propose to design and learn hyper regressors for representing generic images of natural scenes, instead of specific categories of objects.

Continuation of sparse coding principle. The Gabor wavelets (or edgelets, curvelets etc.) provide sparse coding of the image data $\mathbf{I} = \sum_{i=1}^n c_{m,i} B_{x_{m,i}, s, \alpha_{m,i}} + U_m$, with residual U_m . The locations and orientations $(x_{m,i}, \alpha_{m,i}, i = 1, \dots, n)$ can be considered the “shape data,” which can be further coded. The shape script model provides such a sparse coding of the shape data based on elementary shapes. When coding $(x_{m,i}, \alpha_{m,i}, i = 1, \dots, n)$, we need to account for the residuals in them, and the residuals in the shape data are exactly the activities in the active basis model. So active basis model is a natural step if we want to continue the sparse coding principle.

Novelty and significance. The shape script model brings sparse coding to a whole new level. It has the potential to transform the field of image representation and modeling.

2.2 A deeper and broader issue: generative vs discriminative learning

Root in statistics. Nowadays a commonly discussed and debated (sometimes the debate is with oneself) topic in machine learning and computer vision communities, the relationship between generative and discriminative approaches for learning was first studied in statistics: Efron (1975) compared the efficiencies of the linear discriminant analysis based on multivariate Gaussian models (generative approach) and logistic regression (discriminative approach), and obtained insightful results.

The study of this issue is often hindered by the lack of realistic generative models. The active basis model, being a simple generative model that is highly non-Gaussian and reasonably realistic, has the potential to contribute to this fundamental topic.

A commonly used argument to support the discriminative approach [15] [32] [2] is that if the goal is simply to separate positive examples from negative examples, then constructing a generative model is an overkill: difficult but unnecessary. However, it is believed that a realistic generative model has the advantage that it can learn efficiently from a small training sample (as D. Geman puts it: what matters is not when sample size goes to infinity, what matters is when sample size goes to zero). Moreover, generative models do not rely on negative examples. In addition, they can be conveniently learned in unsupervised settings by EM-like schemes.

Extrapolative generalization. The PIs would like to add one more important point to support the use of generative models, or more specifically, sparse representations: they enable extrapolative generalization. For instance, generalizing from a horse to a deer, while at the same time, knowing what tells them apart. This can be done only if we identify essential dimensions along which to extrapolate. We propose to study the issue of extrapolative generalization in the project on

shape scripts discussed in the previous section. In this section, we shall propose activities to study supervised and unsupervised learning.

Comparison with adaboost. We propose to compare the active basis model and the adaboost method [32] [67] for supervised learning, where the weak classifiers of adaboost are based on thresholding $\text{MAX1}_m(x, \alpha) = \max_{(\Delta x, \Delta \alpha) \in A(\alpha)} |\langle \mathbf{I}_m, B_{x+\Delta x, s, \alpha+\Delta \alpha} \rangle|^2$.

As shown by Friedman, Hastie, and Tibshirani (2000) [34], the adaboost method is closely related to logistic additive regression. Let $p(\mathbf{I})$ be the distribution of positive examples, $q(\mathbf{I})$ be the distribution of negative examples, and ρ be the prior probability or proportion of positives, then the class probability and the likelihood ratio are linked by

$$\Pr(+|\mathbf{I}) = \frac{\rho p(\mathbf{I})}{\rho p(\mathbf{I}) + (1 - \rho)q(\mathbf{I})} = \frac{1}{1 + [(1 - \rho)/\rho][q(\mathbf{I})/p(\mathbf{I})]}. \quad (8)$$

The learning and inference of exponential family model such as the active basis model and the logistic regression model or adaboost are driven by two different criteria. The learning of the exponential family model, or generative learning, is based on maximizing the average of log-likelihood ratio, $\log[p(\mathbf{I})/q(\mathbf{I})]$, over positives. The learning of logistic regression, or discriminative learning, is based on maximizing the average of the log of class probability, $\log \Pr(\text{class label}|\mathbf{I})$ (or some margin criterion), over both positives and negatives. From Equation (8), we see that the likelihood ratio $p(\mathbf{I})/q(\mathbf{I})$ and the class probability $\Pr(+|\mathbf{I})$ are linked by a non-linear transformation, where $\Pr(+|\mathbf{I})$ saturates at 1 as $p(\mathbf{I})/q(\mathbf{I}) \rightarrow \infty$. This suggests that the likelihood criterion focuses on typical examples inside the classification boundary, whereas the logistic regression focuses on marginal examples that are close to the classification boundary. We propose to study the effect of this difference, both empirically and theoretically. The logistic regression is related to the partial likelihood [70], which is less efficient than the full likelihood. Here, the efficiency is not only about estimating the parameters, but also about selecting basis elements or weak classifiers. We propose to study the issue of efficiency in both estimation and selection.

Latent variables vs features. In $\text{MAX1}_m(x, \alpha) = \max_{(\Delta x, \Delta \alpha) \in A(\alpha)} |\langle \mathbf{I}_m, B_{x+\Delta x, s, \alpha+\Delta \alpha} \rangle|^2$, the active basis learning treats the local maximization as inferring the activities, which are latent variables in the generative model. Here the max-out of these latent variables can be considered as an approximation to integrating them out [46]. The perturbed basis element $B_{x_i+\Delta x_{m,i}, s, \alpha_i+\Delta \alpha_{m,i}}$ explains away part of \mathbf{I}_m , and inhibits nearby basis elements from explaining the same part of \mathbf{I}_m due to the generative linear regression structure. However, in adaboost, $\text{MAX1}_m(x, \alpha)$ is simply treated as a feature, and inhibition is done through reweighing. We propose to study the effect of this difference.



Figure 8: In each block: the left plot displays the active basis template and the right plot displays the adaboost template (red bars illustrate the features of the form $\text{MAX1}_m(x, \alpha) > c$, and blue bars $\text{MAX1}_m(x, \alpha) < c$). They are learned from the same sets of positive examples. Numbers of positive examples: 30 horses, 9 deer, 12 cows, 33 butterflies. Negative examples: 400 patches cropped from the two natural images in Figure (2) at different resolutions.

We have conducted some experiments showing that the active basis model compares favorably to the adaboost classifier in terms of testing results when sample size is relatively small [74]. Perhaps what interests us more is the comparison of the templates learned. Figure (8) displays active basis templates and corresponding adaboost templates. Each active basis template is in general cleaner than the adaboost template. Our limited experience suggests that increasing the number of positive examples does not always make the adaboost template cleaner, possibly

because it is affected more by the marginal examples close to the classification boundary. We propose to investigate this issue more thoroughly.

Knowledge representation. Figure (8) also raises a deeper question about how to represent the knowledge of visual patterns in general. Should we represent them in terms of classifiers or in terms of generative models? A classifier depends on the definition of “negative examples,” which may change according to the situation. For instance, sometimes we want to separate a horse from the background image, but sometimes we want to separate a horse from a deer. A generative model targeting typical examples is less affected by the definition of “negative examples.” We propose to study the effect of negative examples.

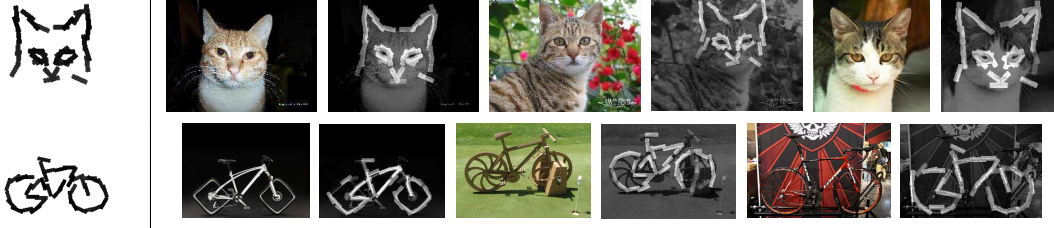


Figure 9: Learning from images where the objects appear at unknown locations and scales. The cat face template is learned from 9 images (3 shown here). The bike template is learned from 7 images (3 shown here).

Unsupervised learning. We also propose to compare the active basis model with discriminative approaches in the so-called “unsupervised learning” (which should be more appropriately called “less supervised learning”). For a generative model, unsupervised learning can be naturally carried out by incorporating latent variables such as unknown locations and poses into the model. The learning can still be guided by maximum likelihood, which can be achieved by the EM-type algorithms [17] [49] [29] [59].

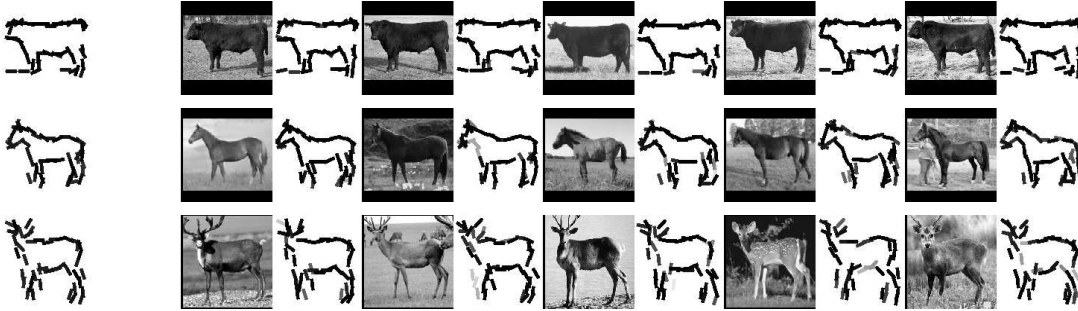


Figure 10: Templates learned from a mixed training set of images of different objects.

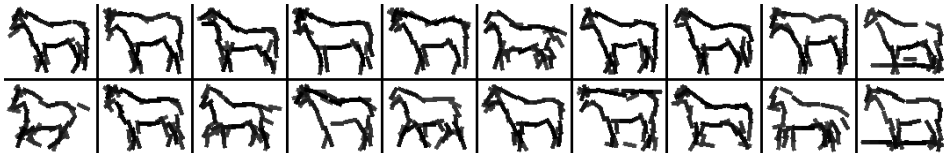


Figure 11: Horse templates of different poses learned from 900+ images by an EM-like local learning scheme.

For instance, Figure (9) shows the learning of the active basis templates when the objects have unknown locations and scales in the training images. Figure (10) shows the learning from a training set that mixes images from three categories.

Comparison with SVM. We have developed an EM-like local learning scheme which learns an active basis template around each image. We can then select from such locally learned templates to obtain a subset of “exemplars” or “prototypes.” Figure (11) shows the horse templates at different poses learned from 900+ images by such a scheme.

The locally learned models appear to be tantalizingly related to the kernel functions of SVM [15], except that such local models are not hand-designed, and may have more generalization power than the kernels. It is unclear whether we can use SVM to select the “exemplars” for classification. We propose to investigate this issue.

The locally learned models can also be used to initialize the EM algorithm for fitting a mixture model. We propose to conduct experiments to study the classification performance of the mixture of active basis models, similar to [4].

Fraction of missing information. Recently researchers in machine learning and computer vision have also developed discriminative methods for unsupervised learning, such as multiple instance learning [68] and latent SVM [28] [76] [79]. We propose to compare our generative methods with these. A key concept in the EM-type learning algorithms is the fraction of missing information [17] caused by the latent variables. We propose to study the corresponding concept in discriminative approach for unsupervised learning.

Another critical issue is the initialization of the learning algorithm. The active basis model can be trained on a single image, which often gives a reasonable initialization. We propose to investigate this issue.

Novelty and significance. The classical Fisherian philosophy of parsimonious models, understanding oriented and likelihood-based learning is being seriously challenged by the modernistic Vapnikian philosophy of complex world, prediction oriented and margin-based learning [66]. The proposed activity has the potential to enhance our understanding of this profound issue both empirically and theoretically.

2.3 A unified model of both shape and texture

Adaptive null hypothesis against variable selection. Recall the active basis model in the form of orthogonal decomposition $\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{m,i} + U_m = C_m \mathbf{B}_m + U_m$, where $p(\mathbf{I}_m | \mathbf{B}_m) = p(C_m)q(U_m | C_m) = q(\mathbf{I}_m) \prod_{i=1}^n p_i(c_{m,i})/q(c_{m,i})$. The reference distribution $q(\mathbf{I}_m)$ is introduced to account for the strong edges in the residual U_m , and this leads to the likelihood ratio $p_i(c_{m,i})/q(c_{m,i})$ for selecting the regressor B_i . $p_i(c)$ is pooled from $\{\langle \mathbf{I}_m, B_{m,i} \rangle, m = 1, \dots, M\}$, where $B_{m,i} = B_{x_{m,i}, s, \alpha_{m,i}}$, with $x_{m,i} = x_i + \Delta x_{m,i}$, and $\alpha_{m,i} = \alpha_i + \Delta \alpha_{m,i}$. It serves as the alternative hypothesis. $q(c)$ is pooled from the two natural images in Figure (2). It serves as the null hypothesis.

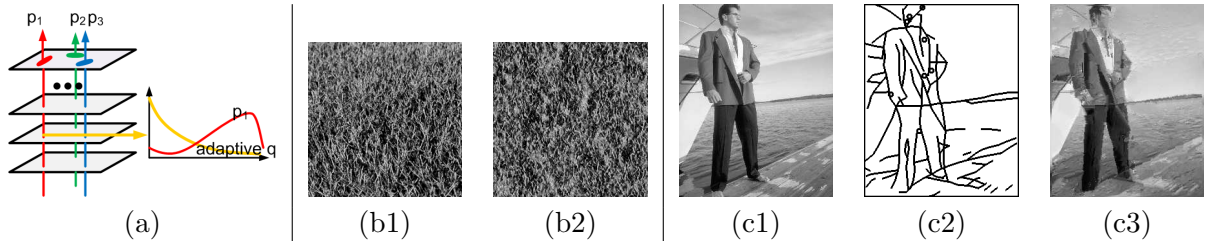


Figure 12: (a1) Spatially pooled marginal histogram as adaptive background. (b1) Observed grass image. (b2) Synthesized by matching spatial statistics. (c1) Observed image. (c2) Sketch. (c3) Synthesized by coupling sketch and texture.

Although the heavy-tailed and highly non-Gaussian $q(c)$ gives a reasonable description of the background U_m , it is not fitted to U_m specifically, and is not the most relevant null hypothesis. A more appropriate $q(c)$ should be pooled from \mathbf{I}_m specifically around x_i at orientation α_i . That

is, $q(c)$ should be pooled from $\{\langle \mathbf{I}_m, B_{x,s,\alpha_i} \rangle, x \in W(x_i) \}$, where $W(x_i)$ is a local neighborhood or local window around x_i . We call this distribution $q_{m,x_i,s,\alpha_i}(c)$. Figure (12.a) illustrates the idea, where p_i is pooled vertically over multiple images at a fixed location (subject to local perturbations), and q is pooled horizontally on a particular image over a window around this fixed location. Such a q serves as an adaptive background for describing U_m . It is the adaptive null hypothesis against p_i . In other words, we should score $\langle \mathbf{I}_m, B_{x_m,i,s,\alpha_{m,i}} \rangle$ against its surrounding peers $\{\langle \mathbf{I}_m, B_{x,s,\alpha_i} \rangle, x \in W(x_i) \}$ (like comparing a Harvard student's SAT score with his or her classmates, instead of the whole population of students).

Interestingly, such adaptive marginal distributions (or histograms) have been used for describing textures. In particular, the two PIs, together with Mumford, develop a Markov random field model for textures based on such texture statistics [85]. The resulting Markov random field model is simply an exponential family model with such texture statistics being the sufficient statistics. Markov random fields [6] [35] and wavelet sparse coding [47] [21] seem to be a world apart, even though both enjoy wide popularity in image modeling. These two classes of models come together in the active basis model.

As an example, in Figure (12), (b1) shows a texture image, and (b2) is a randomly sampled image that matches the above-mentioned marginal histograms extracted from (b1). Clearly, (b2) shares the same texture pattern as (b1) even though they are two different images. As another example, images (c1), (c2) and (c3) illustrate the modeling of an image using both sketch and texture [39] [40].

The Markov random field models can in general be justified by maximum entropy (or minimum Kullback-Leibler divergence) principles [53] [1] [85] [72]. The density substitution scheme and the exponential tilting scheme in Subsection (1.4) also follow the same principle. The maximum entropy $q(\mathbf{I})$ that reproduces the marginal distributions $q(c)$ pooled from generic natural images has been studied by the co-PI and Mumford [82].

Unified treatment of shape and texture. Recall that in Subsection (1.4), $p_i(c)$ is further parametrized by $p(c; \lambda_i)$ as an exponential tilting of $q(c)$ pooled from generic natural images. λ_i is estimated by $\hat{\lambda}_i = \mu^{-1}(\sum_{m=1}^M h(|\langle \mathbf{I}_m, B_{m,i} \rangle|^2)/M)$, where the averaging is along the vertical arrow in Figure (12.a). We can model the adaptive $q_{m,x_i,s,\alpha_i}(c)$ in exactly the same way. Let

$$h_{x,s,\alpha}(\mathbf{I}_m) = \frac{1}{|W(x)|} \sum_{x' \in W(x)} h(|\langle \mathbf{I}_m, B_{x',s,\alpha} \rangle|^2) \quad (9)$$

be the locally and spatially pooled average, where $|W(x)|$ is the number of pixels in the local window $W(x)$ around x , and the averaging is along the horizontal arrow in Figure (12.a). We can model $q_{m,x_i,s,\alpha_i}(c) = q(c; \lambda_{m,x_i,\alpha_i})$, which is also an exponential tilting of the generic $q(c)$. Then we can estimate $\hat{\lambda}_{m,x_i,\alpha_i} = \mu^{-1}(h_{x_i,s,\alpha_i}(\mathbf{I}_m))$. This gives us a unified treatment of shape and texture. Both are based on the same Gabor filter responses, and both are modeled by the same form of exponential family models. They form an inseparable couple of alternative and null hypotheses, where object shape pops out against the background texture.

A model coupling shape and texture. The texture statistics $h_{x,s,\alpha}(\mathbf{I}_m)$ leads to the adaptive null hypothesis against $p_i(c)$. Meanwhile, $\{h_{x,s,\alpha}(\mathbf{I}_m), \forall \alpha, s\}$ also provide useful description of the local texture pattern around x . Such spatial statistics can be very important for modeling textures in natural scenes, as well as textured objects such as zebra, giraffe, cheetah, leopard, tiger, etc. We propose the following model that couples both shape and texture:

$$p(\mathbf{I}_m) = q(\mathbf{I}_m) \prod_{j=1}^K \frac{p(h_{x_j,s,\alpha_j}(\mathbf{I}_m))}{q(h_{x_j,s,\alpha_j}(\mathbf{I}_m))} \prod_{i=1}^n \frac{p(c_{m,i}; \lambda_i)}{p(c_{m,i}; \lambda_{m,x_i,\alpha_i})}, \quad (10)$$

which contains K almost non-overlapping patches of textures for “painting” the background, and n almost non-overlapping strokes for “sketching” the foreground shape against the “painted”

background. The model approximately follows the form $p(\text{textures})p(\text{sketches} \mid \text{textures})$. In model (10), we score $c_{m,i}$ against the adaptive $p(c_{m,i}; \lambda_{m,x_i,\alpha_i})$ fitted to the texture statistics $h_{x_i,s,\alpha_i}(\mathbf{I}_m)$. This approximately takes care of the conditioning in $p(\text{sketches} \mid \text{textures})$.

The selection of texture statistics $(h_{x_j,s,\alpha_j}, j = 1, \dots, K)$ also follows the density substitution scheme, where $p(h_{x_j,s,\alpha_j}(\mathbf{I}_m))$ is the distribution of $h_{x_j,s,\alpha_j}(\mathbf{I}_m)$ estimated from training images, and $q(h_{x_j,s,\alpha_j}(\mathbf{I}_m))$ is obtained under the reference model $q(\mathbf{I})$. We propose to study the mathematical forms and the parametrizations of these two distributions of local spatial statistics. We also propose to develop the learning algorithm for selecting these statistics.

Our recent work [58] shows that this coupled model improves the testing results for classification tasks over the original active basis model, see also [84]. We can also extend the model to dynamic textures [22] and action templates [77].

Novelty and significance. Modeling residuals in linear regression as the adaptive null hypothesis against variable selection is a new way of thinking about linear regression, and brings together wavelets sparse coding and spatial statistics to form a unified and more powerful model.

2.4 Prior NSF support related to this proposal

Our work on the active basis model has mainly been supported by the NSF grant DMS-0707055. The results have already been extensively reviewed in previous sections. So in the following, we shall only briefly list related publications.

DMS-0707055: 07/2007- 06/2010, *From information scaling to regimes of statistical models of natural image patterns*, PI Y. N. Wu and Co-PI S. C. Zhu, \$390K. This project studies the connection between different regimes of visual patterns through the scaling of images. Main results include: (1) Analysis of the change of entropy rate and inferential uncertainty in the scaling process [72]. (2) Active basis model for patterns in the mid-entropy regime [73] [74] [59]. (3) Developing model and algorithm for ChIP-Chip data [80].

IIS-0713652: 07/2007-06/2010, *Large scale object recognition and ground truth representation by stochastic image grammar*, PI S. C. Zhu and co-PI Y. N. Wu, \$450K. This project studies stochastic grammar in the form of and-or graphs [83] for image modeling. Main results include: (1) Supervised learning of the and-or graph for object models [55]. (2) Scheduling of top-down and bottom-up computations [71]. (3) Applications to face modeling [62], aerial image understanding [54], and image to text [78]. (4) Learning animated templates [77]. (5) Learning image manifolds [84].

3 Educational activities and broader impacts

The two PIs and Prof. Alan Yuille are jointly running the Center for Image and Vision Science at UCLA. The proposed activities will provide training opportunities for students to apply statistical theories and methodologies to vision.

For the past two years, the PI participated in the Cross-disciplinary Scholars in Science and Technology (CSST) program jointly sponsored by UCLA and elite universities in China to train undergraduate students during their summer visits to UCLA. The PI has worked with two undergraduate students Yulong He (in 2008) and Zhuoliang Kang (in 2009) on various aspects of the active basis model. The reproducibility webpage has proven to be very useful for the training.

The co-PI organized China-US-France Summer School on Machine Learning, Statistics and Vision in the summer of 2008 in China. It was enthusiastically attended by students from many universities in China. The co-PI also co-organized the First International Workshop on Stochastic Image Grammar in the summer of 2009.

The PIs propose to continue such efforts to promote the training of students and to facilitate the integration of research themes in statistics, applied mathematics, and computer science.