

Learning Compositional Sparse Coding Models for Natural Images

PI: Ying Nian Wu, UCLA Department of Statistics

Project Summary

1. Intellectual merit.

Images of natural scenes are a type of *big data* that are bewilderingly rich in patterns and abundantly available. Developing statistical models and associated learning algorithms for natural images is of fundamental importance for computer vision, and more importantly, the endeavor has the potential to enrich our treasured collections of statistical models and expand the already vast reach of statistical methodologies.

The PI proposes to learn compositional sparse coding models for representing natural images. The proposed models are built upon the original sparse coding framework where there is a dictionary of basis functions often in the form of localized, elongated and oriented wavelets, so that each image can be represented by a linear combination of a small number of basis functions automatically selected from the dictionary. In our compositional sparse coding models, the representational units are groups of basis functions exhibiting recurring compositional patterns in terms of their spatial arrangements. These compositional patterns can be considered shape templates. We propose unsupervised methods for learning a dictionary of frequently occurring templates from training images, so that each image can be represented by a small number of templates automatically selected from the learned dictionary.

The following are potential contributions of the proposed research. (1) *Sparse and symbolic representation of high-dimensional data.* The proposed compositional sparse coding scheme translates a raw image of a large number of pixel intensities into a small number of templates, thus generating a symbolic representation of the image data. Such symbolic representations can be crucial for image understanding and classification. (2) *New ground beyond Lasso and group Lasso.* In the $p \gg n$ regression setting, the sparsity or structured sparsity such as group sparsity enables us to infer the regression coefficients of the predictor vectors or regressors by Lasso or group Lasso, assuming that the candidate regressors and the candidate groups are given. The proposed research goes beyond inferring coefficients. It amounts to learning a dictionary of candidate groups from the data. It also amounts to selecting the regressors that are not only sparse but also highly patterned, where the patterns are unknown and are to be learned. (3) *New objects in the sparse-land.* In wavelets sparse coding framework (sometimes referred to as sparse-land), the dictionary of wavelets are often called atoms. The proposed research is to discover composite structures formed by atoms, which lead to much sparser and more meaningful representations than atomic decompositions. (4) *New hierarchical and spatial models.* The proposed models can be viewed as hierarchical models that seek to model the spatial arrangements of the selected basis functions.

2. Broader impacts.

The proposed activities will strengthen the educational and research program in the Department of Statistics, UCLA. Topics related to the proposed research can be incorporated into related graduate courses. The wealth of data and code that have been posted on the reproducibility webpages have proven useful for training both undergraduate and graduate students. Such webpages will continue to be enhanced and developed. The proposed research will support graduate students. The proposed activities will provide innovative training by mixing students from different academic backgrounds, namely, mathematics-statistics and engineering-computer science. The PI plans to organize workshops to promote generative modeling and learning for vision, and to facilitate the integration of research themes in statistics, applied mathematics, and computer science.

1 Introduction

1.1 Motivation and objective

We are living in the exciting era of *big data*, big not only in the amount of data but also in the dimensionality and complexity of the data. We statisticians are facing both tremendous challenges and thrilling opportunities in developing methods and theories for learning from the big data that arise from different disciplines. One type of big data that are bewilderingly rich in patterns and abundantly available are images of natural scenes [60]. The objective of the proposed research is to develop statistical models and associated learning algorithms for representing patterns in natural images. This endeavor is of fundamental importance for computer vision, more importantly, it shall also enrich our treasured collections of statistical models and expand the already vast reach of statistical methodologies.

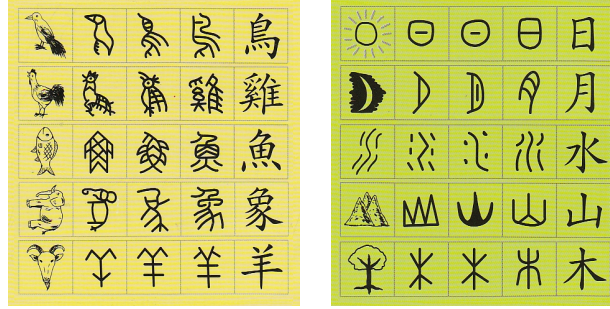


Figure 1: Chinese characters evolved from representations of natural images of objects and scenes [40]. In each row, the first block shows a picture of the object, and the rest four blocks display the evolution of the corresponding Chinese character over time. Left panel: bird, chicken, fish, elephant and goat. Right panel: sun, moon, water, mountain and wood.

To be more specific, we propose to learn compositional sparse coding models for representing natural images. As illustrated by Figure 1, the ancient Chinese developed the early form of the Chinese characters as a coding scheme for representing natural images where each character is a pictorial description of a pattern. The early pictorial form then gradually evolved into the form that is in use today. The system of Chinese characters can be considered a compositional sparse code: each natural image can be described by a small number of characters selected from the dictionary, and each character is a composition of a small number of strokes (the strokes become clearer in the more evolved form of the Chinese characters in Figure 1).

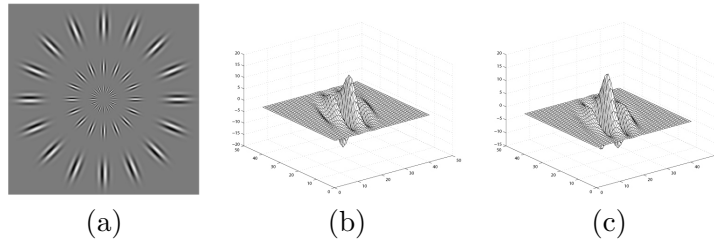


Figure 2: The Gabor wavelets are Gaussian modulated sine and cosine waves. They can serve as basis functions that can be linearly combined to represent natural images. (a) A sample of Gabor wavelets at different locations, orientations, and scales. (b) A Gabor sine wavelet. (c) A Gabor cosine wavelet. The Gabor wavelets can be truncated to have finite support (and length).

The compositional sparse coding models that we propose to develop can be viewed as mathematical realizations of the coding system of the Chinese characters. In our proposed models, each

“stroke” is a linear basis function such as a Gabor wavelet [13] (see Figure 2 for an illustration), and the images are represented by linear combinations of these basis functions. Each “character” is a compositional pattern or a shape template formed by a group of selected basis functions. We propose unsupervised learning methods for learning the frequently occurring templates from training images, so that each training image can be represented by a small number of templates automatically selected from the learned dictionary of templates.

Finding sparse representations of high-dimensional data is of fundamental importance for understanding and analyzing the data. Our proposed compositional sparse coding scheme translates a raw image of a large number of pixel intensities into a small number of templates, thus facilitating the signal to symbol transition and giving rise to a symbolic representation of the image data that is much sparser and more meaningful than the wavelets representation. Our preliminary experiments show that our method is capable of learning meaningful compositional sparse code. Experiments also show that the learned templates can be useful for image classification. For example, they serve as meaningful “visual words” for the so-called “bag-of-words” classification scheme.

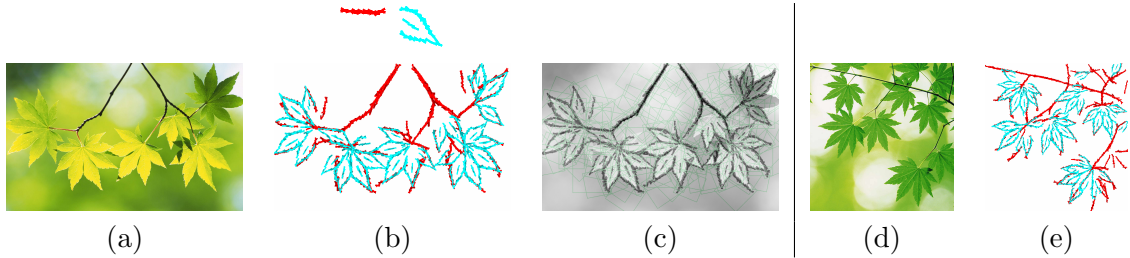


Figure 3: Unsupervised learning of compositional sparse code (a,b,c) and using it for recognition (d,e). Each Gabor wavelet (17×17 pixels in size) is illustrated by a bar with the same location, orientation and length. (a) Training image of 480×768 pixels. (b) Above: 2 compositional patterns (twig and leaf) in the form of shape templates learned from the training image. The bounding box of each template is 100×100 pixels. The numbers of wavelets in the twig and leaf templates are 22 and 40 respectively, and the numbers are automatically determined. Below: Representing the training image by spatially translated, rotated, scaled and deformed copies of the 2 templates. (c) Superposing the deformed templates on the original image. Green squared boxes are bounding boxes of the templates. (d) Testing image. (e) Representation (understanding) of the testing image by the 2 templates.

Figure 3 displays an example from our preliminary experiments. We assume that the dictionary of basis functions is given (this assumption is to be relaxed in the proposed work, where we propose to learn basis functions as well as their grouping patterns), and they are Gabor wavelets centered at a dense collection of locations and tuned to a collection of scales and orientations (see Figure 2). In Figure 3, each Gabor wavelet is illustrated by a bar at the same location and with the same length and orientation as the corresponding wavelet. The wavelets are well connected and form clear templates. Figure 3.(a) displays the training image. (b) displays a mini-dictionary of 2 compositional patterns of wavelets learned from the training image. Each compositional pattern is a template formed by a group of a small number of wavelets at selected locations and orientations. The learning is unsupervised in the sense that the image is not labeled or annotated. The number of templates in the dictionary is automatically determined by a BIC-like criterion. The 2 templates are displayed in different colors, so that it can be seen clearly how the spatially translated, rotated, scaled and deformed copies of the 2 templates are used to represent the training image, as shown in (b). In (c), the templates are overlaid on the original image, where each green squared box is the bounding box of the template. In our current implementation, we allow limited amount of overlap between the bounding boxes of the templates. The templates learned from the training image can be generalized to testing image, as shown in (d) and (e).

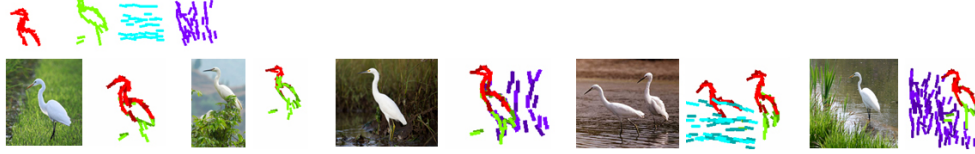


Figure 4: Unsupervised learning of 4 compositional patterns (templates) from 20 training images. The bounding box of each template is 100×100 .

Figure 4 shows another example from our preliminary experiments, where the learned patterns repeat across different images.

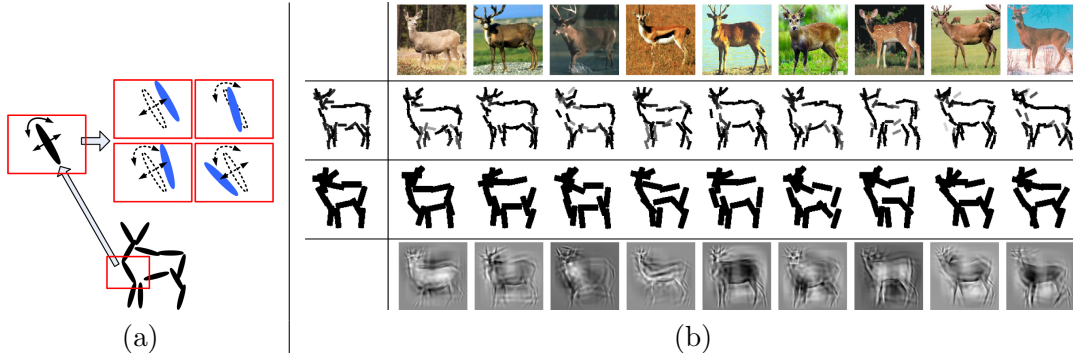


Figure 5: (a) An active basis model is a composition of a small number of basis functions, each is a Gabor wavelet in our current experiments and is illustrated by a bar with the same location, orientation and length. Each basis function can perturb its location and orientation. (b) Supervised learning of active basis model from aligned images. The first row displays the 9 training images. The second row: the first plot is the nominal template formed by 50 selected basis functions. The rest of the plots are the deformed templates matched to the images. The third row: the same as the second row, except that the scale of the Gabor wavelets is about twice as large, and the number of wavelets is 14. The last row displays the reconstruction of each training image by linear combination of 100 selected and perturbed basis functions at multiple scales.

We represent each compositional pattern of basis functions by an active basis model developed by the PI and collaborators [68], while being supported by prior NSF grants. An active basis model is a composition of a small number of basis functions automatically selected from a given dictionary. The selected basis functions are allowed to perturb their locations and orientations so that the linear basis formed by the group of selected basis functions become active and the active basis can be viewed as a deformable template [2]. Figure 5 illustrates the basic idea of the active basis model and the supervised learning of the model from aligned images.

1.2 Statistical foundations and contributions

New ground beyond Lasso and group Lasso. Recent years have witnessed an explosion of research activities on sparsity and structured sparsity [4, 33]. In the $p \gg n$ regression setting [8], the most popular tools for variable selection under sparsity or group sparsity are Lasso [62] (see also [4, 3, 5, 21, 35, 63, 73, 79, 80] etc. for related work on theories and methods) and group Lasso [70] (see also [4, 74]). Related non-convex methods include SCAD [23] and MCP [72] etc. In the language of $p \gg n$ regression, each image can be viewed as a response vector, and each basis function in the dictionary can be viewed as a predictor vector or a regressor. In Lasso and group Lasso, the candidate regressors and the candidate groups of regressors are given. In our proposed work, we seek to learn the candidate groups of regressors from the training data. From a variable

selection perspective, our method amounts to selecting the regressors that are not only sparse but also highly patterned, where the patterns are unknown and are to be discovered from the data. Thus the proposed research has the potential to break new ground in the exciting area of sparsity and structured sparsity.

New objects in sparse-land. The equivalence of Lasso in harmonic analysis and signal processing literature is basis pursuit [10], where the basis functions are often called atoms, and the sparse coding by the selected basis functions is called atomic decomposition [15, 18]. The basis functions can either be learned [46, 1] or designed, such as Gabor wavelets [13], edgelets [17], wedgelets [16], ridgelets [6], curvelets [7], and beamlets [34] etc. This atomic sparse coding framework is sometimes referred to as sparse-land. Our proposed work seeks to learn composite structures formed by the atoms, following the compositionality principle [27, 77]. These composite structures lead to much sparser and more meaningful representations than atomic decompositions. If the basis functions are atoms, then their compositions may be viewed as molecules.

New hierarchical and spatial models. The proposed models can be viewed as hierarchical models. At the bottom layer, the images are represented by linear combinations of the dictionary of basis functions. In the literature of Bayesian variable selection [28, 11, 47], the coefficients are often assumed to be independent super-Gaussian distributions (or mixtures of Gaussian distributions). Our proposed models correct the independence assumption by modeling the spatial patterns formed by the basis functions with non-zero coefficients.

Comparison with deep learning. The proposed models (including the multi-layer version in subsection (4.5)) bear some similarity to deep learning [30], especially those under sparsity constraint [36, 39, 61, 71]. The difference is that our method seeks to learn patterned sparsity in the linear regression. The representational units in our models are sparse compositions of selected basis functions, where sparsity is directly built into the representational units and is achieved by a shared variable selection scheme (or a corresponding penalty term as in subsection (4.1)). As a result, our representational units are more explicit and interpretable, which is more in line with statistical thinking.

To conclude, nowadays the boundaries between statistics, machine learning and signal analysis are much more blurred than before, and what the PI proposes to do is really *nothing but statistics*.

2 Background: sparse coding model and active basis model

This section reviews the original sparse coding model and the active basis model in order to fix the notation and set the stage for the proposed research.

2.1 Olshausen-Field model: learning the dictionary of regressors

Olshausen and Field [46] proposed that the role of simple cells in primary visual cortex is to infer sparse representations of natural images. Let $\{\mathbf{I}_m, m = 1, \dots, M\}$ be a set of training image patches (e.g. 12×12), which are two-dimensional functions defined on a certain image domain. The Olshausen-Field model seeks to represent these images by

$$\mathbf{I}_m = \sum_{i=1}^N c_{m,i} B_i + U_m, \quad (1)$$

where $(B_i, i = 1, \dots, N)$ is a dictionary of basis functions defined on the same domain as \mathbf{I}_m . We assume that the basis functions are normalized to have unit ℓ_2 norm. $c_{m,i}$ are the coefficients, and U_m is the unexplained residual image. N is often assumed to be greater than the number of pixels in \mathbf{I}_m (e.g. $N = 2 \times 12 \times 12$), so the dictionary is said to be over-complete or redundant. The basis functions in this redundant dictionary can afford to be very specific so that the number

of coefficients ($c_{m,i}, i = 1, \dots, N$) that are non-zero (or significantly different from zero) is assumed to be small (e.g., less than 10) for each image \mathbf{I}_m .

The dictionary of basis functions ($B_i, \forall i$) can be learned from the training images $\{\mathbf{I}_m, m = 1, \dots, M\}$ by minimizing

$$\sum_{m=1}^M \left[\left\| \mathbf{I}_m - \sum_{i=1}^N c_{m,i} B_i \right\|^2 + \lambda \sum_{i=1}^N S(c_{m,i}) \right] \quad (2)$$

jointly over ($B_i, \forall i$) and ($c_{m,i}, \forall m, i$), where $S()$ is a sparsity inducing penalty function, and λ is a regularization parameter. Interestingly, the learned (B_i) resemble Gabor wavelets in Figure 2! In the language of regression, \mathbf{I}_m is a response vector (e.g., 144-dimensional), and each B_i is a regressor. So (2) enables us to learn a dictionary of candidate regressors from a training sample of response vectors.

Geometric attributes. One may assume that the basis functions in the dictionary are spatially translated, rotated and dilated versions of one another [48], so that each B_i can be written as $B_{x,s,\alpha}$, where x is the location (a two-dimensional vector), s is the scale, and α is the orientation. We call such a dictionary self-similar, and we call (x, s, α) the geometric attribute of $B_{x,s,\alpha}$. Given the self-similar dictionary ($B_{x,s,\alpha}, \forall (x, s, \alpha)$), with (x, s, α) properly discretized, we seek to encode \mathbf{I}_m by

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_{m,i}, s_{m,i}, \alpha_{m,i}} + U_m, \quad (3)$$

where $n \ll N$ is a small number. $(x_{m,i}, s_{m,i}, \alpha_{m,i}, i = 1, \dots, n)$ form a spatial point process.

2.2 Active basis model for shared sparse coding of aligned image patches

The active basis model was proposed by Wu et al. [68] for modeling deformable templates formed by basis functions.

Suppose we have a set of training image patches $\{\mathbf{I}_m, m = 1, \dots, M\}$. This time we assume that they are defined on the same bounding box, and the objects in these images come from the same category. In addition, these objects appear at the same location, scale and orientation, and in the same pose in the images. See Figure 5 for 9 image patches of deer. We call such image patches aligned.

The active basis model is of the following form

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} + U_m, \quad (4)$$

where $\mathbf{B} = (B_{x_i, s, \alpha_i}, i = 1, \dots, n)$ form the nominal template of an active basis model (sometimes we simply call \mathbf{B} an active basis template). Here we assume that the scale s is fixed and given. $\mathbf{B}_m = (B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}, i = 1, \dots, n)$ is the deformed version of the nominal template \mathbf{B} for encoding \mathbf{I}_m , where $(\Delta x_{m,i}, \Delta \alpha_{m,i})$ are the perturbations of the location and orientation. The perturbations are introduced to account for shape deformation. Both $\Delta x_{m,i}$ and $\Delta \alpha_{m,i}$ are assumed to vary within limited ranges.

In the language of linear regression, the multiple response vectors $\{\mathbf{I}_m\}$ share the common set of regressors. This is the setting of multi-task learning or support union regression [45, 41, 38], except that the regressors are subject to perturbations in their geometric attributes.

2.3 Shared variable selection: least squares setting

Given the dictionary of basis functions (regressors) $\{B_{x,s,\alpha}, \forall x, s, \alpha\}$, the learning of the active basis model from the aligned image patches $\{\mathbf{I}_m\}$ involves the selection of B_{x_i,s,α_i} and the inference of its perturbed version $B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}$ in each image \mathbf{I}_m . So the problem amounts to select a common set of regressors (up to perturbations) shared by multiple response vectors. We call the learning as supervised, because the bounding boxes of the objects are given and the images are aligned. See Figure 5 for an illustration of the learning results.

The shared matching pursuit algorithm is a generalization of matching pursuit [43]. It is a greedy algorithm that seeks the maximal reduction of the following squared loss in each iteration:

$$\sum_{m=1}^M \|\mathbf{I}_m - \sum_{i=1}^n c_{m,i} B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}\|^2. \quad (5)$$

[0] Initialize $i \leftarrow 0$. For $m = 1, \dots, M$, initialize the residual image $U_m \leftarrow \mathbf{I}_m$.

[1] $i \leftarrow i + 1$. Select $(x_i, \alpha_i) = \arg \max_{x,\alpha} \sum_{m=1}^M \max_{\Delta x, \Delta \alpha} |\langle U_m, B_{x+\Delta x, s, \alpha+\Delta \alpha} \rangle|^2$, where $\max_{\Delta x, \Delta \alpha}$ is the local maximum pooling within the small ranges of $\Delta x_{m,i}$ and $\Delta \alpha_{m,i}$.

[2] For $m = 1, \dots, M$, given (x_i, α_i) , infer the perturbations by retrieving the arg-max in the local maximum pooling of step [1]: $(\Delta x_{m,i}, \Delta \alpha_{m,i}) = \arg \max_{\Delta x, \Delta \alpha} |\langle U_m, B_{x_i+\Delta x, s, \alpha_i+\Delta \alpha} \rangle|^2$. Let $c_{m,i} \leftarrow \langle U_m, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}} \rangle$, and let $U_m \leftarrow U_m - c_{m,i} B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}$.

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

The local max pooling in step [1] is hypothesized as the function of complex cells in primary visual cortex [51].

2.4 Statistical modeling: non-Gaussian exponential family model

Orthogonality. Because of the arg-max explaining-away in step [2], the basis functions selected for each deformed template $\mathbf{B}_m = (B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}, i = 1, \dots, n)$ usually have little correlation with each other. For computational and modeling convenience, we may assume that these selected basis functions are orthogonal to each other, so that the coefficient can be obtained by projection: $c_{m,i} = \langle \mathbf{I}_m, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}} \rangle$. We write $C_m = (c_{m,i}, i = 1, \dots, n)$. In practice, we allow small overlap between the selected basis functions in each \mathbf{B}_m . The orthogonality assumption shall be relaxed in the proposed work.

Density substitution. The algorithm guided by (5) implicitly assumes that the residual U_m is Gaussian white noise. This assumption can be questionable because the unexplained background in the image may contain salient structures such as edges. A better assumption is to assume that U_m follows the same distribution as that of natural images. More precisely, the distribution of \mathbf{I}_m given the deformed template $\mathbf{B}_m = (B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}}, i = 1, \dots, n)$, i.e., $p(\mathbf{I}_m | \mathbf{B}_m)$, is obtained by modifying the distribution of natural images $q(\mathbf{I}_m)$ in such a way that we only change the distribution of $C_m = (c_{m,i} = \langle \mathbf{I}_m, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}} \rangle, i = 1, \dots, n)$ from $q(C_m)$ to $p(C_m)$, while leaving the conditional distribution of U_m given C_m unchanged. Here $p(C_m)$ and $q(C_m)$ are the distributions of C_m under $p(\mathbf{I}_m | \mathbf{B}_m)$ and $q(\mathbf{I}_m)$ respectively. Specifically, $p(\mathbf{I}_m | \mathbf{B}_m) = q(\mathbf{I}_m)p(C_m)/q(C_m)$. Such a density substitution scheme was first used in projection pursuit density estimation [26], see also [50, 37, 78, 67].

For computational simplicity, we further assume that $(c_{m,i} = \langle \mathbf{I}_m, B_{x_i+\Delta x_{m,i},s,\alpha_i+\Delta\alpha_{m,i}} \rangle, i = 1, \dots, n)$ are independent given \mathbf{B}_m , under both p and q , so $p(\mathbf{I}_m | \mathbf{B}_m) = q(\mathbf{I}_m) \prod_{i=1}^n p_i(c_{m,i})/q(c_{m,i})$, where $q(c)$ is assumed to be the same for $i = 1, \dots, n$ because $q(\mathbf{I}_m)$ is translation and rotation invariant. $q(c)$ can be pooled from natural images in the form of a heavy-tailed histogram.

Exponential family model. For parametric modeling, we assume the following exponential family model $p_i(c) = p(c; \lambda_i)$, where $p(c; \lambda) = \exp\{\lambda h(|c|^2)\}q(c)/Z(\lambda)$. $h(r)$ is a function of the response $r = |c|^2$ that saturates for large r . Specifically, we assume that $h(r) = \xi[2/(1 + e^{-2r/\xi}) -$

1]. $Z(\lambda)$ is the normalizing constant. $\mu(\lambda) = \mathbb{E}_\lambda[h(r)] = \int h(r)p(c; \lambda)dc$ is the mean parameter. Both $Z(\lambda)$ and $\mu(\lambda)$ can be computed beforehand from natural images.

The log-likelihood is

$$l(\{\mathbf{I}_m\} \mid \mathbf{B}, \{\mathbf{B}_m\}) = \sum_{m=1}^M \sum_{i=1}^n \left[\lambda_i h(\langle \mathbf{I}_m, B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} \rangle) - \log Z(\lambda_i) \right]. \quad (6)$$

2.5 Shared variable selection: maximum likelihood setting

We revise the shared matching algorithm in subsection (2.3) in order to maximize the log-likelihood (6) instead of minimizing the squared loss (5) as in subsection (2.3).

[0] Initialize $i \leftarrow 0$. For each m , initialize the response maps $R_m(x, \alpha) \leftarrow \langle \mathbf{I}_m, B_{x, s, \alpha} \rangle$ for all (x, α) .

[1] $i \leftarrow i + 1$. Select $(x_i, \alpha_i) = \arg \max_{x, \alpha} \sum_{m=1}^M \max_{\Delta x, \Delta \alpha} h(|R_m(x + \Delta x, \alpha + \Delta \alpha)|^2)$, where $\max_{\Delta x, \Delta \alpha}$ is again local maximum pooling.

[2] For $m = 1, \dots, M$, given (x_i, α_i) , infer the perturbations by retrieving the arg-max in the local maximum pooling of step [1]: $(\Delta x_{m,i}, \Delta \alpha_{m,i}) = \arg \max_{\Delta x, \Delta \alpha} |R_m(x_i + \Delta x, \alpha_i + \Delta \alpha)|^2$. 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 $\text{corr}(B_{x, s, \alpha}, B_{x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}) > \epsilon$. Then compute λ_i by solving the maximum likelihood equation $\mu(\lambda_i) = \sum_{m=1}^M h(|c_{m,i}|^2)/M$.

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

In step [2], the arg-max basis function inhibits correlated basis functions to enforce the approximate orthogonality.

3 Proposed model and algorithm, with preliminary results

3.1 Compositional sparse coding model: grouping the regressors

In this subsection, we strive to write down the proposed model in a form that is analogous to the Olshausen-Field model (3), by using compactified notation.

Compactified notation. As the first step of this exercise of compactification, let us slightly generalize the active basis model by assuming that the template may appear at location X_m in image \mathbf{I}_m , then we can write the representation in the following form:

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_{X_m + x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}} + U_m = C_m \mathbf{B}_{X_m} + U_m, \quad (7)$$

where $\mathbf{B}_{X_m} = (B_{X_m + x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}, i = 1, \dots, n)$ is the deformed template spatially translated to X_m , $C_m = (c_{m,i}, i = 1, \dots, n)$, and $C_m \mathbf{B}_{X_m}$ is defined to be $\sum_{i=1}^n c_{m,i} B_{X_m + x_i + \Delta x_{m,i}, s, \alpha_i + \Delta \alpha_{m,i}}$. Here we no longer assume that the training images $\{\mathbf{I}_m\}$ are aligned.

\mathbf{B}_{X_m} explains the part of \mathbf{I}_m that is covered by \mathbf{B}_{X_m} . For each image \mathbf{I}_m and each X_m , we can define the log-likelihood ratio:

$$l(\mathbf{I}_m \mid \mathbf{B}_{X_m}) = \log \frac{p(\mathbf{I}_m \mid \mathbf{B}_{X_m})}{q(\mathbf{I}_m)} = \sum_{i=1}^n \left[\lambda_i \max_{\Delta x, \Delta \alpha} h(|\langle \mathbf{I}_m, B_{X_m + x_i + \Delta x, s, \alpha_i + \Delta \alpha} \rangle|^2) - \log Z(\lambda_i) \right] \quad (8)$$

As the next step of this compactification exercise, in addition to spatial translation and deformation, we can also rotate and scale the template. So a more general version of (7) is $\mathbf{I}_m = C_m \mathbf{B}_{X_m, S_m, A_m} + U_m$, where X_m is the location, S_m is the scale, and A_m is the orientation of the spatially translated, rotated, scaled and deformed template. The scaling of the template can be implemented by changing the resolution of the original image. We adopt the convention that whenever the notation \mathbf{B} appears in image representation, it always means the deformed

template, where the perturbations of the basis functions can be inferred by local max pooling. The log-likelihood ratio $l(\mathbf{I}_m | \mathbf{B}_{X_m, S_m, A_m})$ can be similarly defined as in (8).

Compactified representation. Now suppose we have a dictionary of T active basis templates, $\{\mathbf{B}^{(t)}, t = 1, \dots, T\}$, where each $\mathbf{B}^{(t)}$ is a type of compositional pattern of basis functions. Then we can represent the image \mathbf{I}_m by K_m templates that are spatially translated, rotated, scaled and deformed copies of these T types of templates in the dictionary:

$$\mathbf{I}_m = \sum_{k=1}^{K_m} C_{m,k} \mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})} + U_m, \quad (9)$$

where each $\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_k)}$ is obtained by translating the template of type t_k , i.e., $\mathbf{B}^{(t_k)}$, to location $X_{m,k}$, dilate it to scale $S_{m,k}$, rotate it to orientation $A_{m,k}$, and deform it to match \mathbf{I}_m .

If the K_m templates do not overlap, the log-likelihood is $\sum_{m=1}^M \sum_{k=1}^{K_m} \left[l(\mathbf{I}_m | \mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}) \right]$. In order to control model complexity, we attach a penalty γ to each basis function used for representing the image \mathbf{I}_m . γ can be interpreted as the cost for coding the perturbations of each basis function from the MDL perspective [52]. It can also be viewed from the BIC perspective [54] as compensating for the fact that we max out the perturbations instead of integrating them out. The penalized log-likelihood is

$$\sum_{m=1}^M \sum_{k=1}^{K_m} \left[l(\mathbf{I}_m | \mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}) - n^{(t_{m,k})} \gamma \right], \quad (10)$$

where $n^{(t)}$ is the number of basis functions in $\mathbf{B}^{(t)}$. The penalty γ enables us to determine when to stop the shared matching pursuit algorithm in supervised learning, so that the number of basis functions in a template can be automatically determined.

Connection with group Lasso. In the representation (9), each $\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}$ is a group of basis functions (or regressors), and the K_m groups are to be selected from the collection of groups that correspond to all possible translated, rotated, scaled and deformed versions of the compositional patterns in the dictionary. The situation is very similar to that of group Lasso [70], which is also about selecting groups of variables from all possible candidate groups. Our work goes beyond the group Lasso scenario in that the collection of groups is unknown, and we learn a dictionary of compositional patterns of these groups from training images. This dictionary then defines a large collection of candidate groups by spatial translation, rotation, scaling and deformation.

3.2 Preliminary version of the proposed unsupervised learning algorithm

The preliminary version of our proposed learning algorithm seeks to maximize the log-likelihood (10). It is an iterative algorithm where each iteration consists of two steps.

Step (I): Group selection — Image encoding by template matching pursuit. Suppose we are given the current dictionary $\{\mathbf{B}^{(t)}, t = 1, \dots, T\}$. Then for each \mathbf{I}_m , the template matching pursuit process seeks to represent \mathbf{I}_m by sequentially selecting a small number of templates (groups of regressors) from the dictionary. Each selection seeks to maximally increase the penalized log-likelihood (10).

[I.0] Initialize the maps of template matching scores for all (X, S, A, t) : $\mathbf{R}_m^{(t)}(X, S, A) \leftarrow l(\mathbf{I}_m | \mathbf{B}_{X, S, A}^{(t)}) - n^{(t)} \gamma$. This can be accomplished by first rotating the template $\mathbf{B}^{(t)}$ to orientation A , and then scanning the rotated template over the image zoomed to the resolution that corresponds to scale S . Initialize $k \leftarrow 1$.

[I.1] Select the spatially translated, rotated, scaled and deformed template by finding the global maximum of the response maps: $(X_{m,k}, S_{m,k}, A_{m,k}, t_{m,k}) = \arg \max_{X, S, A, t} \mathbf{R}_m^{(t)}(X, S, A)$.

[I.2] Let the selected arg-max template inhibit overlapping candidate templates. Let D be the side length of the bounding box of the selected template $\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}$, then for all (X, S, A, t) , if X is within a distance ρD from $X_{m,k}$, then set the response $\mathbf{R}_m^{(t)}(X, S, A) \leftarrow -\infty$. Currently we set $\rho = .4$, so we allow only limited overlap between the templates. We shall relax this requirement in the proposed work and pursue more rigorous methods for group selection.

[I.3] Stop if all $\mathbf{R}_m^{(t)}(X, S, A, t) \leq 0$. Otherwise let $k \leftarrow k + 1$, and go to [I.1].

The penalty γ enables us to determine when to stop the template matching pursuit process, so that the number of templates K_m for encoding \mathbf{I}_m is automatically determined.

Step (II): Shared variable selection — Dictionary re-learning by shared matching pursuit. For each $t = 1, \dots, T$, we re-learn $\mathbf{B}^{(t)}$ (by selecting a small set of basis functions or regressors) from all the image patches that are currently covered by $\mathbf{B}^{(t)}$.

[II.0] Image patch cropping. For each \mathbf{I}_m , go through all the selected templates $\{\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}, \forall k\}$ that encode \mathbf{I}_m . If $t_{m,k} = t$, then crop the image patch of \mathbf{I}_m (at the resolution that corresponds to $S_{m,k}$) covered by the bounding box of the template $\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}$.

[II.1] Template re-learning. Re-learn template $\mathbf{B}^{(t)}$ from all the image patches covered by $\mathbf{B}^{(t)}$ that are cropped in [II.0], with their bounding boxes aligned. The learning is accomplished by the shared matching pursuit algorithm of subsection (2.5).

Random initialization. The learning algorithm is initialized by learning each $\mathbf{B}^{(t)}$ from image patches that are randomly cropped from $\{\mathbf{I}_m\}$, so these initial templates are rather meaningless. Meaningful templates emerge very quickly after a few iterations.

Figure 6 illustrates the learning of the maple leaf template from the training image shown in Figure 3. Figure 6.(a) traces the template of maple leaf learned over the first 7 iterations of the learning algorithm. (b) shows the process of shared matching pursuit for learning this template in the last (10th) iteration, where the constituent basis functions are sequentially added.



Figure 6: (a) Template of leaf learned in the first 7 iterations of the unsupervised learning algorithm. (b) In each of iteration, the shared matching pursuit process selects the basis functions sequentially to form each template. The sequence shows the process selecting 1, 3, 5, 10, 20, 30, 40 wavelets to form the leaf template in the last (10th) iteration.

Dictionary size. In order to determine the number of templates in the dictionary, T , we adopt an adjusted BIC-like criterion [54, 25]:

$$\beta \sum_{m=1}^M \sum_{k=1}^{K_m} \left[l(\mathbf{I}_m | \mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}) - n^{(t_{m,k})} \gamma \right] - \frac{1}{2} \sum_{t=1}^T n^{(t)} \log \sum_{m=1}^M K_m, \quad (11)$$

where β is a ratio that discounts the overlap between the selected templates $\{\mathbf{B}_{X_{m,k}, S_{m,k}, A_{m,k}}^{(t_{m,k})}\}$. We currently define it as the number of pixels actually covered by the selected templates and the sum of the numbers of pixels of these templates.

3.3 Preliminary results on unsupervised learning

Figure 7 shows an example of selecting the number of templates T in the dictionary. The first image is the training image. The remaining four blocks display the learned dictionaries as well as the representations of the training image using the learned dictionaries. The numbers of templates in the dictionaries are respectively 1, 2, 3, and 4. Just as in Figures 3 and 6, each basis function or wavelet is illustrated by a bar at the same location and orientation, and with the same length

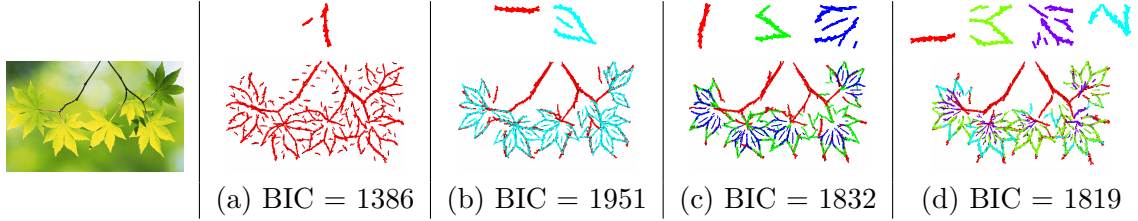


Figure 7: The adjusted BIC computed for different numbers of templates (1-4) in the dictionaries. The size of templates is 100×100 . The allowed range of scale change is $\{.8, 1, 1.2\}$ of the original image. The templates are allowed full range of rotation. The maximal number of basis functions in each template is 40 and the actual number is automatically determined.

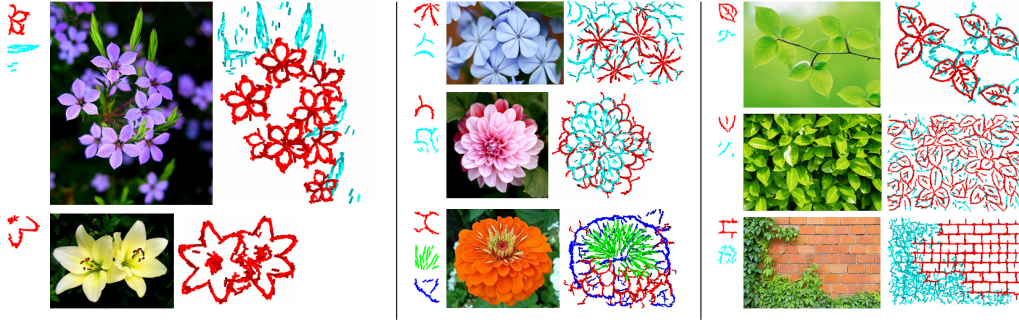


Figure 8: flowers, leaves and ivy wall. Parameters are the same as in Figure 7.

as the corresponding wavelet. All the templates are of the size 100×100 pixels. We also display the adjusted BIC criterion for each learned dictionary. Figures 8 to 10 show more examples of representing natural images.

Better “words” for classification. The learned templates can serve as the visual words in the “bag of words” scheme [12] for image classification. Our preliminary results show that the learned templates achieve better classification performance than the codebooks learned based on the popular SIFT features [42]. Our current classification results is close to the state of art for the Caltech 101 data set [24].

More results and details of preliminary experiments can be found in our reproducibility page: <http://www.stat.ucla.edu/~ywu/ABC/ABC.html>

4 Planned research activities

We shall continue to experiment with the model and algorithm described in the previous section. The PI proposes to investigate the following issues. (1) The effect of the penalty term γ on controlling the model complexity. (2) Selection of the template size and image resolution, in addition to the dictionary size. (3) Integration of image segmentation into dictionary learning.

We shall also explore more rigorous learning methods without making simplified assumptions such as approximate orthogonality of the deformed templates and limited overlap between the selected templates.

4.1 Planned activity 1: two-way group Lasso for patterned variable selection

An interesting feature of the unsupervised learning algorithm in the previous section is that both the image encoding step (I) and the dictionary learning step (II) are generalizations of matching pursuit. In fact, the iterative learning algorithm can be viewed as highly patterned variable selection, *except that the patterns are unknown!*

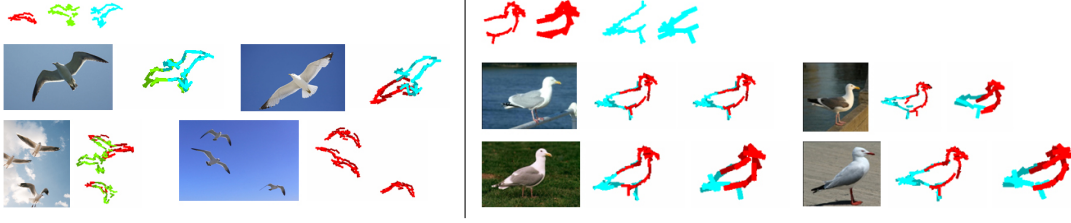


Figure 9: Seagulls flying (number of training images is 20) and standing (number of training images is 11). For the standing seagulls experiment, we learn multi-scale templates at two different scales.

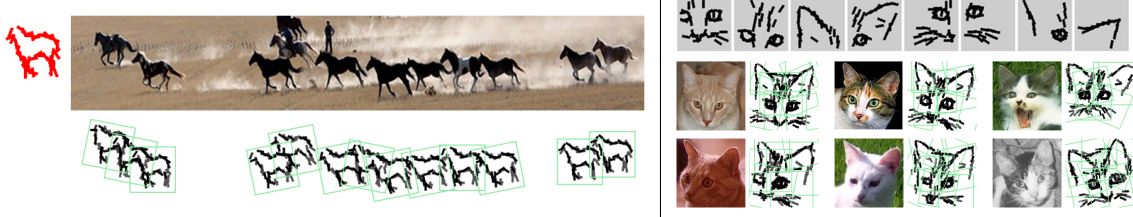


Figure 10: Horse; cat faces (number of training images is 89).

For simplicity, let us assume that the templates can only spatially translate (without rotation and dilation). Let us also ignore the perturbations of the basis functions at this stage. Then according to our compositional sparse coding model, each image \mathbf{I}_m is a linear combination of K_m templates:

$$\mathbf{I}_m = \sum_{k=1}^{K_m} C_{m,k} \mathbf{B}_{X_k}^{(t_k)} + U_m = \sum_{k=1}^{K_m} \sum_{i=1}^{n^{(t_k)}} c_{m,i,X_k}^{(t_k)} B_{X_k+x_i^{(t_k)},s,\alpha_i^{(t_k)}} + U_m, \quad (12)$$

where template k is a spatially translated version of a template of type t_k in the dictionary $\{\mathbf{B}^{(t)}, t = 1, \dots, T\}$. Each template $\mathbf{B}^{(t)}$ is a group of $n^{(t)}$ basis functions (regressors): $\mathbf{B}^{(t)} = (B_{x_i^{(t)},s,\alpha_i^{(t)}}, i = 1, \dots, n^{(t)})$. If we spatially translate $\mathbf{B}^{(t)}$ to location X , then $\mathbf{B}_X^{(t)} = (B_{X+x_i^{(t)},s,\alpha_i^{(t)}}, i = 1, \dots, n^{(t)})$. Recall that we fix the scale s .

Compared to the original sparse coding model $\mathbf{I}_m = \sum_{i=1}^{n_m} c_{m,i} B_{x_i,s,\alpha_i} + U_m$, (12) is highly patterned variable selection: the selected regressors form K_m groups that exhibit T types of recurring compositional patterns in their spatial arrangements.

We can solve this patterned variable selection problem by a highly stylized Lasso, which we call two-way group Lasso:

$$R(\{c_{m,t,X+x,\alpha}, \forall m, t, X, x, \alpha\}) = \sum_{m=1}^M \|\mathbf{I}_m - \sum_{t=1}^T \sum_{X \in D_m} \sum_{(x,\alpha) \in D_0} c_{m,t,X+x,\alpha} B_{X+x,s,\alpha}\|^2 \quad (13)$$

$$+ \gamma_0 \sum_{m=1}^M \sum_{t=1}^T \sum_{X \in D_m} \sum_{(x,\alpha) \in D_0} |c_{m,t,X+x,\alpha}| \quad (14)$$

$$+ \gamma_1 \sum_{m=1}^M \sum_{t=1}^T \sum_{X \in D_m} \sqrt{\sum_{(x,\alpha) \in D_0} c_{m,t,X+x,\alpha}^2} \quad (15)$$

$$+ \gamma_2 \sum_{t=1}^T \sum_{(x,\alpha) \in D_0} \sqrt{\sum_{m=1}^M \sum_{X \in D_m} c_{m,t,X+x,\alpha}^2} \quad (16)$$

In (13), D_m is the domain of image lattice of \mathbf{I}_m , which contains the locations of all the candidate templates (groups). D_0 contains all the candidate basis functions (regressors) in each template,

where each basis function is indexed by its location x and orientation α (scale s is fixed). D_0 is centered at origin. (14) is a generic sparsity term for the selected basis functions. (15) is the group sparsity that encourages encoding each \mathbf{I}_m by a small number of templates (groups). The grouping (inside the square root sign) is within the same translated template. (16) is the shared variable selection sparsity that encourages each template in the dictionary to have a small number of basis functions. The grouping is across different images and locations that are explained by the same template in the dictionary. It is multi-task learning or support union regression [45, 41, 38].

After minimizing R (with $\gamma_0, \gamma_1, \gamma_2$ carefully tuned), the sparsity pattern caused by (15) produces encoding of \mathbf{I}_m by a small number of templates, similar to template matching pursuit. The sparsity pattern caused by (16) gives us the T templates in the dictionary, similar to shared matching pursuit. Of course, in real applications, we shall also add back the rotation and scaling of the templates, as well as the perturbations of the basis functions.

We shall study the algorithmic issues regarding the minimization of R . By using the squared loss in (13), we assume Gaussian white noises for U_m . For this assumption to approximately hold, we need to whiten \mathbf{I}_m before learning. We shall also consider the non-Gaussian models as in subsections (2.4) and (2.5).

4.2 Planned activity 2: Bayesian variable selection for patterned sparsity

We shall also study the model (12) in the framework of Bayesian variable selection [28, 11, 47]. To that end, we need to define two sets of indicators for variable/group selection. One is $\delta_{m,X}^{(t)}$ for group selection. If $\delta_{m,X}^{(t)} = 1$, then the template $\mathbf{B}^{(t)}$ is active at location X in image \mathbf{I}_m . Otherwise $\delta_{m,X}^{(t)} = 0$. The other set of indicators is $\eta_{x,\alpha}^{(t)}$ for shared variable selection. If $\eta_{x,\alpha}^{(t)} = 1$, then the basis function $B_{x,s,\alpha}$ is included in template $\mathbf{B}^{(t)}$. Otherwise $\eta_{x,\alpha}^{(t)} = 0$. We still have the linear regression model as in (13):

$$\mathbf{I}_m = \sum_{t=1}^T \sum_{X \in D_m} \sum_{(x,\alpha) \in D_0} c_{m,t,X+x,\alpha} B_{X+x,s,\alpha} + U_m, \quad (17)$$

where the coefficient $c_{m,t,X+x,\alpha} = 0$ if either $\delta_{m,X}^{(t)} = 0$ or $\eta_{x,\alpha}^{(t)} = 0$. Otherwise, if both $\delta_{m,X}^{(t)} = 1$ and $\eta_{x,\alpha}^{(t)} = 1$, then the coefficient has a prior distribution $c_{m,t,X+x,\alpha} \sim N(0, \tau^2)$. We assume the white noise model for residual: $U_m(x) \sim N(0, \sigma^2)$ i.i.d. The prior distribution for $\delta_{m,X}^{(t)}$ is Bernoulli(ρ_1). The prior distribution for $\eta_{x,\alpha}^{(t)}$ is Bernoulli(ρ_2). The posterior of $(\delta_{m,X}^{(t)}, \eta_{x,\alpha}^{(t)}, c_{m,t,X+x,\alpha})$ can be sampled by MCMC. $\delta_{m,X}^{(t)}$ gives us the image encoding by templates, similar to template matching pursuit. $\eta_{x,\alpha}^{(t)}$ gives us the basis functions in each templates, similar to shared matching pursuit.

We shall study the design of efficient MCMC for Bayesian posterior sampling, for instance, by modifying the PI's recent work [9]. We shall also study non-Gaussian models for U_m and $c_{m,t,X+x,\alpha}$ as in subsections (2.4) and (2.5).

4.3 Planned activity 3: active factor analysis — learning regressors

In the active basis model, we assumed that the dictionary of the candidate basis functions (regressors) is given. The PI proposes to learn basis functions from the training data, and we call it active factor analysis (or active component analysis).

We first consider the supervised learning of a single active basis model from aligned images. The model is $\mathbf{I}_m = \sum_{i=1}^n c_{m,i} B_i(\Delta_{m,i}) + U_m$, where $\mathbf{B} = (B_i, i = 1, \dots, n)$ is a set of unknown basis functions, and $B_i(\Delta_{m,i})$ denotes the basis function obtained by perturbing the location and orientation of B_i by $\Delta_{m,i}$, which encodes the perturbation of B_i in image \mathbf{I}_m and which is

restricted to limited range. We assume locality and sparsity directly: each B_i is locally supported and centered at x_i , where $(x_i, i = 1, \dots, n)$ form a regular grid in the image domain, with n small. We can learn $\mathbf{B} = (B_i)$ by minimizing $\sum_{m=1}^M \|\mathbf{I}_m - \sum_{i=1}^n c_{m,i} B_i(\Delta_{m,i})\|^2$ jointly over $(c_{m,i}, \Delta_{m,i})$ and $\mathbf{B} = (B_i)$. The computational scheme of K-SVD [1] may be employed.

We can also learn a dictionary of active basis models $\{\mathbf{B}^{(t)} = (B_i^{(t)}, i = 1, \dots, n), t = 1, \dots, T\}$ in unsupervised fashion by minimizing

$$\sum_{m=1}^M \|\mathbf{I}_m - \sum_{t=1}^T \sum_{X \in D_m} \sum_{i=1}^n c_{m,t,X,i} B_i^{(t)}(X, \Delta_{m,t,X,i})\|^2 + \gamma \sum_{m=1}^M \sum_{t=1}^T \sum_{X \in D_m} \sqrt{\sum_{i=1}^n c_{m,t,X,i}^2}, \quad (18)$$

where $B_i^{(t)}(X, \Delta_{m,t,X,i})$ is to spatially translate the i -th unknown basis function in template t , i.e., $B_i^{(t)}$, to location X and perturb it by $\Delta_{m,t,X,i}$. Again we can treat this problem in the Bayesian variable selection framework by introducing the selection indicator $\delta_{m,X}^{(t)}$.

4.4 Planned activity 4: combining generative and discriminative learning



Figure 11: Generative vs discriminative learning. In each block, the template on the left is learned generatively from positive images by shared matching pursuit. The template on the right is learned discriminatively from both positive and negative images by Lasso-logistic regression, where each template consists of selected basis functions with positive coefficients λ_i .

In supervised learning, we can also train the active basis model, i.e., selecting the basis functions $\mathbf{B} = (B_{x_i,s,\alpha_i}, i = 1, \dots, n)$, by discriminative learning. Let S_+ be the set of positive training images, and S_- be the set of negative training images, and let $y_m \in \{+1, -1\}$ be the class label of image \mathbf{I}_m . Then according to the non-Gaussian exponential family model in subsection (2.4), $p(y_m | \mathbf{I}_m)$ follows a logistic regression with coefficients (λ_i) and variables $(\max_{\Delta x, \Delta \alpha} h(|\langle \mathbf{I}_m, B_{x_i+\Delta x, s, \alpha_i+\Delta \alpha} \rangle|^2))$. We can select $\mathbf{B} = (B_{x_i,s,\alpha_i})$ by fitting ℓ_1 -regularized (Lasso) logistic regression.

Figure 11 displays the templates learned by the shared matching pursuit algorithm (generative) and the Lasso-logistic regression (discriminative). An interesting observation is that the generative templates are much cleaner, but the discriminative templates give better classification performances on testing data.

We shall compare the discriminative and generative learning empirically and theoretically [20]. For instance, we shall examine the performance of discriminative learning in unsupervised setting. The PI also proposes to combine generative and discriminative learning in supervised setting by minimizing

$$\begin{aligned} R(\{\lambda_{x,\alpha}, c_{m,x,\alpha}\}) &= \sum_m \log(1 + \exp\{-y_m \sum_{x,\alpha} \lambda_{x,\alpha} \max_{\Delta x, \Delta \alpha} h(|\langle \mathbf{I}_m, B_{x+\Delta x, s, \alpha+\Delta \alpha} \rangle|^2)\}) \\ &+ \gamma_0 \sum_{m \in S_+} \|\mathbf{I}_m - \sum_{x,\alpha} c_{m,x,\alpha} B_{x+\Delta x, s, \alpha+\Delta \alpha}\|^2 \\ &+ \gamma_1 \sqrt{\lambda_{x,\alpha}^2} + \gamma_2 \sum_{m \in S_+} c_{m,x,\alpha}^2. \end{aligned} \quad (19)$$

It is again multi-task learning [45, 41, 38]. The last term in (19) ties the variable selection in discriminative and generative learning. We shall also study this combination in unsupervised setting.

4.5 Planned activity 5: hierarchical and spatial modeling



Figure 12: Supervised learning of template of tandem bike. We first learn overlapping part-templates, and then select some of them (with bounding boxes) according to their log-likelihood scores (indexed by color).

We can also learn multi-layer compositional sparse coding models, where each template is a deformable composition of part-templates, and each part-template itself is a deformable composition of a number of basis functions. Figure 12 illustrates the basic idea of supervised learning. We can combine it with template matching pursuit for unsupervised learning. We can also modify the two-way group Lasso objective function in subsection (4.1) or the Bayesian variable selection model in subsection (4.2) for learning hierarchical structures. For structural texture images, such as brick wall and ivy leaves, we shall also consider modeling the spatial arrangements of the templates (groups of regressors).

4.6 Planned activity 6: theoretical analysis

We shall study the performance of the two-way group Lasso in subsection (4.1) in terms of estimation accuracy [3, 5, 73] and more importantly variable selection accuracy [75, 63]. Most of Lasso analyses in the literature assume independent sub-Gaussian errors. We shall study dependent super-Gaussian errors, which is a more realistic assumption for U_m .

In a recent paper by Wei Biao Wu and the PI [66], we analyzed the performance of Lasso under dependent super-Gaussian errors by generalizing the Nagaev inequality [44] to the weighted and dependent case [65]. The original inequality is as follows. Let X_1, \dots, X_n be mean 0 independent random variables, and $S_n = \sum_{i=1}^n X_i$. Further assume that X_i has finite q -th moment, i.e., $\|X_i\|_q = [E(|X_i|^q)]^{1/q} < \infty$, $q > 2$, for $i = 1, \dots, n$. Let $\mu_{n,q} = \sum_{i=1}^n E(|X_i|^q)$. By Corollary 1.7 in Nagaev (1979) [44], for $x > 0$, the tail probability

$$P(|S_n| \geq x) \leq (1 + 2/q)^q \mu_{n,q}/x^q + 2 \exp\{-c_q x^2/\mu_{n,2}\}, \quad (20)$$

where $c_q = 2e^{-q}(q+2)^{-2}$. This is a very sharp inequality under polynomial moment condition. The PI plans to extend (20) to two-dimensional dependent case and use it to analyze the two-way group Lasso in subsection (4.1).

4.7 Timeline and reproducible research

Year 1: Finish coding and obtain initial experimental results. *Year 2:* Conduct extensive experiments on large data sets, and apply the learning results to applications such as object detection, classification, as well as image compression (due to space limit, we have not discussed these applications in detail, but we will definitely work on them). *Year 3:* Further extend the proposed methods and identify more applications, and try to obtain state of art performances.

The PI has been working hard on reproducible research. The reproducibility webpages contain a wealth of data and code. The PI will continue to adhere to the principle of reproducible research [19], by posting all the data, code and results on the reproducibility webpages.

5 Prior NSF support related to this proposal

NSF DMS: *Statistical Modeling and Learning in Vision*, 07/01/10 - 06/30/13, PI Wu, co-PI Zhu. Much of our work on active basis model and the preliminary results reported in this proposal are supported by this grant. The following are specific results. (1) Learning active basis model for object recognition [68, 55]. (2) Modeling geometric shape motifs and curves by active basis model [59, 32]. (3) Hybrid image templates that combine sketches and textures [57]. (4) Bayesian variable selection by stochastic matching pursuit [9]. (5) Learning compositional sparse coding for natural images [31].

NSF IIS: *Learning and Inference in And-Or Graphs for Image Understanding*, 08/01/10 - 07/31/13, PI Zhu, co-PI Wu. This grant supports work on learning hierarchical models in the form of and-or graphs and developing inference algorithms. The following are specific results. (1) Bottom-up and top-down inference processes in and-or graphs [64], as well as fast MCMC algorithms for computing multiple solutions in graphical models [49]. (2) Learning and-or templates [58] and animated pose templates [69] for object and action recognition. (3) Image parsing by stochastic scene grammar [76]. (4) Intractability in motion data [29].

6 Educational activities and broader impacts

The PI is part of the vision group in UCLA Department of Statistics. The other two faculty members of that group are Song-Chun Zhu and Alan Yuille.

The PI is currently supervising five Ph.D. students. Two of them are working with the PI directly on the proposed project. They mingle with the large number of students of Zhu and Yuille. The proposed activities will provide training opportunities for students from different backgrounds (math-statistics and CS-engineering) to learn, develop and apply modern statistical methodologies to big data analysis such as vision.

The PI has been teaching the graduate courses STAT 200AB and undergraduate courses STAT 100AB on probability and statistics. The ideas of templates and template matching scores provide intuitive examples of statistical models and likelihood scores. The PI will continue to channel the research results obtained under NSF support to his teaching. Zhu has been teaching STAT 232AB (cross-listed as CS262AB) on vision and imaging science. Part of our joint work has been incorporated into the courses. The proposed research shall help to further enhance these courses.

For the past few years, the PI has 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 undergraduate students Yulong He (in 2008, now at Georgia Tech), Zhuoliang Kang (in 2009, supervised by Yuille, now at USC), Ruixun Zhang (2010, now at MIT) and Shuhan Liang (2012, now a senior in Zhejiang University) on various aspects of the PI's research supported by NSF.

The PI has been working hard on reproducible research and has released a wealth of data and code on his reproducible webpages, such as <http://www.stat.ucla.edu/~ywu/ActiveBasis.html>. See more on data management plan.

The PI will be organizing a session on statistical modeling and learning in vision in the 2013 Spring Research Conference on Statistics in Industry and Technology, to be held from June 20 to 22, 2013 in UCLA. The PI plans to organize workshops on generative models in vision together with Zhu, in association with computer vision conferences.

The PI is currently serving as a co-PI in the ONR MURI project: *Knowledge Representation, Reasoning and Learning for Understanding Scenes and Events*. Led by Zhu, this project involves researchers from UCLA, Caltech, Berkeley, Stanford, MIT, and Brown. Each year, there is a meeting in UCLA where the PIs report on their progresses. This provides a good venue to expose the PI's research results, in addition to regular publications in journals and conferences.

References

- [1] M. Aharon, M. Elad, and A.M. Bruckstein. The K-SVD: an algorithm for designing of over-complete dictionaries for sparse representation, *IEEE Transactions On Signal Processing*, **54**, 4311-4322, 2006.
- [2] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, **86** 376-387, 1991.
- [3] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705-1732, 2009.
- [4] P. Buhlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.
- [5] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169-194, 2007.
- [6] E. J. Candes, and D. L. Donoho. Ridgelets: The key to high-dimensional intermittency? *Philosophical Transactions of the Royal Society A*, **357**, 2495-2509, 1999.
- [7] E. J. Candes and D. L. Donoho. Curvelets - a surprisingly effective nonadaptive representation for objects with edges. *Curves and Surfaces*, L. L. Schumakeretal. (eds), Vanderbilt University Press, 1999.
- [8] E. J. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, **35**, 313-351.
- [9] R. B. Chen, C. H. Chu, T. Y. Lai, and Y. N. Wu, Stochastic matching pursuit for Bayesian variable selection. *Statistics and Computing*, **21**, 247-259, 2011.
- [10] S. Chen, D. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, **20**, 33-61, 1999.
- [11] H. Chipman, M. Hamada, and C. F. J. Wu. A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, **39** 372-381, 1997.
- [12] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Workshop of European Conference on Computer Vision*, 2004.
- [13] J. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Society of America*, **2**, 1160-1169, 1985.
- [14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B*, **39**, 1-38, 1977.
- [15] D. L. Donoho. Sparse components of images and optimal atomic decomposition. *Constructive Approximation*, **17**, 353-382, 2001.
- [16] D. L. Donoho. Wedgelets: Nearly minimax estimation of edges. *The Annals of Statistics*, **27** 859-897, 1999.
- [17] D. L. Donoho and X. Huo. Combined image representation using edgelets and wavelets. *Wavelet Applications in Signal and Image Processing VII*, 1999.

- [18] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, **47**, 2845-62, 2001.
- [19] D. Donoho¹, A. Maleki, I. Rahman, M. Shahram¹, and V. Stodden. 15 years of reproducible research in computational harmonic analysis, 2008.
- [20] B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892-898, 1975.
- [21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression. *The Annals of Statistics*, **32**, 407499, 2004.
- [22] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, **9**, 1871-1874, 2008.
- [23] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348-1360, 2001.
- [24] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR Workshop*, 2004.
- [25] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611-631, 2002.
- [26] J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, **82**, 249-266, 1987.
- [27] S. Geman, D. F. Potter, and Z. Chi. Composition systems. *Quarterly of Applied Mathematics*, **60**, 707-736, 2002.
- [28] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of American Statistical Association*, **88**, 881-889, 1993.
- [29] H. F. Gong and S. C. Zhu. Intrackability : characterizing video statistics and pursuing video representations, *International Journal of Computer Vision*, **7**, 255-275, 2012.
- [30] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527-1554, 2006.
- [31] Y. Hong, Z. Z. Si, W. Z. Hu, S. C. Zhu and Y. N. Wu. Unsupervised learning of compositional sparse code for natural image representation, *Quarterly of Applied Mathematics*, under review, 2012.
- [32] W. Z. Hu, Y. N. Wu and S. C. Zhu. Image representation by active curves. *International Conference on Computer Vision*, 2011.
- [33] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Proceedings of the 26th International Conference on Machine Learning*, 2009.
- [34] X. Huo and D. L. Donoho. Applications of beamlets to detection and extraction of lines, curves and objects in very noisy images. *Nonlinear Signal and Image Processing*, 2001.
- [35] G. M. James, P. Radchenko, and J. Lv. DASSO: Connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society, B*, **71**, 127142, 2009.

- [36] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *ICCV*, 2009.
- [37] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.
- [38] M. Kolar, J. Lafferty, and L. Wasserman. Union support recovery in multi-task learning, *Journal of Machine Learning Research*, 2010.
- [39] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML*, 2009.
- [40] C. Lo, *Amazing Chinese Characters*, Panda Media Co., 2002.
- [41] K. Lounici, A. B. Tsybakov, M. Pontil, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. *Proceedings of the 22nd Conference on Learning Theory*, 2009.
- [42] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**, 91–110, 2004.
- [43] S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, **41**, 3397–3415, 1993.
- [44] S. V. Nagaev. Large deviations of sums of independent random variables. *Annals of Probability*, **7**:745–789, 1979.
- [45] G. Obozinski, M. J. Wainwright, and M. I. Jordan, Support union recovery in high-dimensional multivariate regression, *Annals of Statistics*, **39**, 1–47, 2011.
- [46] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609, 1996.
- [47] B. A. Olshausen and K. J. Millman. Learning sparse codes with a mixture-of-Gaussians prior. *Advances in Neural Information Processing Systems*, **12**, 841–847, 2000.
- [48] B. A. Olshausen, P. Sallee and M. S. Lewicki. Learning sparse image codes using a wavelet pyramid architecture. *Advances in Neural Information Processing Systems*, **13**, 887–893, 2001.
- [49] J. Porway and S. C. Zhu. C4: Computing Multiple Solutions in Graphical Models by Cluster Sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1713–1727, 2011.
- [50] S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 380–393, 1997.
- [51] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019–1025, 1999.
- [52] J. Rissanen. *Information and Complexity in Statistical Modeling*. Springer, 2007.
- [53] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, **35**, 1012–1030, 2007.
- [54] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464, 1978.
- [55] Z. Si, H. Gong, S. C. Zhu, and Y. N. Wu. Learning active basis models by EM-type algorithms. *Statistical Science*, **25**, 458–475, 2010.

- [56] Z. Si, and Y. N. Wu, Wavelet, active basis, and shape script — a tour in the sparse land. *ACM SIGMM International Conference on Multimedia Information Retrieval, Special session on Statistical Modeling and Learning for Multimedia*, 2010.
- [57] Z. Si and S. C. Zhu. Learning hybrid image template (HiT) by information projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 1354–1367, 2012.
- [58] Z. Si and S. C. Zhu. Learning and-or templates for object modeling and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, tentatively accepted, 2012.
- [59] Z. Si, H. Gong, S. C. Zhu, and Y. N. Wu. Learning active basis models by EM-type algorithms. *Statistical Science*, **25**, 458-475. 2010.
- [60] A. Srivastava, A. Lee, E. Simoncelli, and S. C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, **18**, 17-33, 2003.
- [61] A. Szlam, K. Kavukcuoglu, and Y. LeCun. Convolutional matching pursuit and dictionary training. arXiv:1010.0422
- [62] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, **58**, 267-288, 1996.
- [63] M. J. Wainwright. Sharp thresholds for noisy and highdimensional recovery of sparsity using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183-2202, 2009.
- [64] T. Wu and S. C. Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International Journal of Computer Vision*, **93**, 226–252, 2011.
- [65] W. B. Wu. Nonlinear system theory: another look at dependence. *Proceedings of National Academy of Science*, 102:14150–14154, 2005.
- [66] W. B. Wu and Y. N. Wu. Analyzing the Lasso with dependent errors using Nagaev-type inequalities, under review, 2012.
- [67] Y. N. Wu, C. Guo, and S. C. Zhu, From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, **66**, 81-122, 2008.
- [68] Y. N. Wu, Z. Si, H. Gong, and S. C. Zhu. Learning active basis model for object detection and recognition. *International Journal of Computer Vision*, **90**, 198-235, 2010.
- [69] B. Yao, Z. Liu, and S.C. Zhu. Animated pose templates for modeling and detecting human actions, under review, 2012.
- [70] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, B*, **68**, 49-67, 2006.
- [71] M. Zeiler, G. Taylor and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. *ICCV*, 2011.
- [72] C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942, 2010.
- [73] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, **37**, 2109-2144, 2009.

- [74] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**, 3468–3497, 2009.
- [75] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541-2567, 2006.
- [76] Y. B. Zhao and S.C. Zhu. Image parsing via stochastic scene grammar, *Neural Information Processing Systems (NIPS)*, 2011.
- [77] S. C. Zhu and D. B. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, **2**, 259–362, 2006.
- [78] S. C. Zhu, Y. N. Wu and D. B. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, **9**, 1627-1660, 1998.
- [79] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418-1429, 2006.
- [80] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, B*, **67**, 301-320, 2005.