

# NSF Annual Report DMS 0707055

## Activities and Findings

### 1. Describe the major research and education activities of the project

The goal of our research is to develop statistical models and model-based algorithms for representing and recognizing visual patterns in natural images.

This is the second year of the proposed research. The following are our activities.

(1) *Active basis model.* We continue our work on active basis model for object recognition. We have made the following progress. (a) We show that the active basis model can be a realistic generative model that can synthesize images of objects. (b) We develop algorithm for learning active basis model from training images where the objects appear at unknown locations and scales. (c) We develop local learning algorithm for learning active basis model locally in the high-dimensional image ensemble. (d) We develop an animated basis model for modeling dynamic action patterns.

Efforts in this aspect lead to three papers. A 60-page paper is accepted by *International Journal of Computer Vision (IJCV)*. Another paper is accepted by *Statistical Science*. A third paper is accepted by International Conference on Computer Vision.

(2) *Learning mixed template of sketches and textures.* We study the connection between the active basis model for objects and the Markov random field model that we have previously developed for textures. We develop a model that naturally combines sketch variables for shapes and texture variables for appearances. Efforts in this aspect lead to two papers. One is accepted by IEEE Conference on Computer Vision and Pattern Recognition. One is under review by *Pattern Recognition Letters*.

(3) *Reproducibility pages.* We continue to develop the reproducibility pages for our papers. The following are some important ones.

<http://www.stat.ucla.edu/~ywu/AB/ActiveBasisMarkII.html>

<http://www.stat.ucla.edu/~zzsi/ActiveBasis.html>

[http://www.stat.ucla.edu/~zzsi/mixed\\_template.html](http://www.stat.ucla.edu/~zzsi/mixed_template.html)

The first page is the major one. It contains the source code for all the 10 sets of experiments in our IJCV paper. We have also completely re-written the source code to make it more efficient and user friendly.

(4) *Supervision of graduate students.* Three graduate students, Nicole Chen, Kent Shi, and Zhangzhang Si received support from this grant and have been working under the supervision of the PI (Wu) and co-PI (Zhu). Kent Shi has graduated with a Ph.D. degree and is now working at Yahoo.

(5) *Summer school and visiting student.* The co-PI (Zhu) organized a China-US-France Summer School on Machine Learning, Statistics, and Computer Vision, from June 30 to July 11, 2008, at the Lotus Hill Institute in Hubei, China. The workshop was very well attended. The PI (Wu) served as the mentor of Yunlong He for the UCLA CSST (cross-disciplinary scholars in science and

technology) program during the summer of 2008. Yunlong He was a undergraduate student from University of Science and Technology of China. He will start his Ph.D. study in Georgia Tech.

## Describe the major findings resulting from these activities

### I. Image synthesis by active basis model

The active basis model is in the form of linear composition of elongated and oriented Gabor wavelet elements at selected locations and orientations. Intuitively, each wavelet element is a “stroke” for sketching the object, and these “strokes” together form a template. In the active basis model, each selected wavelet element is allowed to slightly shift its location and orientation to account for the deformations in object shapes. Therefore, the active basis model can be considered a deformable template. We can learn such an active basis from training images, and then use the learned model to detect and recognize the object in the testing image.

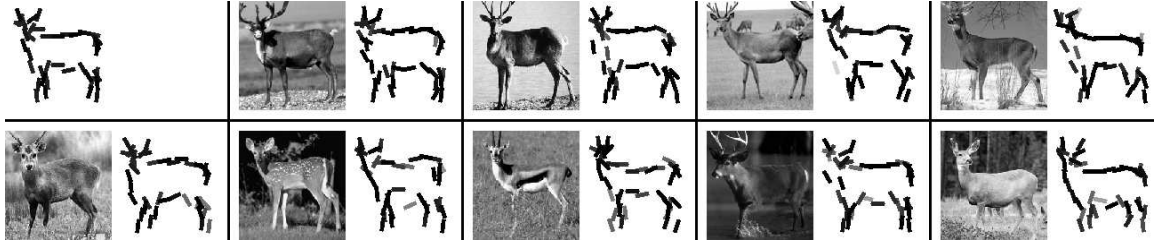


Figure 1: Learning active basis model. The first plot is the learned active basis, where each elongated and oriented wavelet element is represented by a bar of the same length and at the same location and orientation. Each wavelet element is allowed to locally shift its location and orientation to account for shape deformations. In each of the remaining blocks, the image on the left is the training image. The plot to the right of it displays the deformed active basis matched to the training image. The learning of the active basis is based on the maximum likelihood criterion.



Figure 2: Object detection and recognition by the learned active basis model. The image on the left is the testing image. The image on the right is superposed with the deformed active basis that achieves the best matching to the image data, where template matching is measured by the likelihood score.

Figures (1) and (2) show an example, where a template of deer is learned from 9 training images. In order to detect and recognize the deer in the testing image, we can scan the template over the testing image and deform the template to match the image data.

The active basis model is a statistical model that can generate image intensities. Such a model is often called a generative model in machine learning literature. The generative model defines a likelihood function, so that both learning and detection can be based on maximizing the likelihood function.

In machine learning community, there has been on-going debate as to whether we need to develop generative model of the observed data for the purpose of classification or detection, or whether it is sufficient to simply learn the conditional or posterior probability of class label given the observed data. The latter approach is often called discriminative. The discriminative approach is more direct, while the generative approach appears to be an overkill that is more than necessary. Some renowned researchers such as Vapnik strongly argue for the discriminative approach. Vapnik sometimes gets very philosophical on this issue, calling for a paradigm shift to the so-called “anti-scientific” approach where the goal is not to find simple explanations of the data but to approximate the decision rules which can often be complex.

However, we believe that generative model is necessary for a highly specialized field such as vision that deals with highly specialized data, that is, natural images. Intuitively, the generative model tells us things like “what a deer looks like,” while the discriminative approach tells us things like “how to recognize a deer.” Although the two appear to be similar, there is subtle but fundamental difference. Specifically, a generative model such as active basis model can be written as an exponential family model relative to a reference distribution that is a marginal approximation to the distribution of natural images. Such an exponential family translates to a discriminative model such as a logistic regression model for predicting class labels, such as ‘deer’ versus “anything else.” However, the learning of these two models are based on two different criteria. The learning of the exponential family model is based on the likelihood ratio, whereas the learning of the logistic regression is based on class probability (or some margin-based criterion). For a typical positive example, such as a deer image, a large change in likelihood ratio only translates to a small change in class probability, which is close to 1. Therefore, the likelihood ratio tends to be more sensitive than the class probability in selecting the optimal dimensions or features that characterize the typical positive examples, especially when the sample size is small. For the detection and recognition purpose, the likelihood ratio also tends to be more sensitive in detecting the typical examples in testing images. This can be very relevant in unsupervised learning, which involves inferring unknown locations, scales, and poses of the objects. A large change in class probability often occurs when the example is marginal, where the class probability is close to  $1/2$ . So the discriminative approach tends to focus on marginal examples instead of explaining typical examples. For instance, in adaboost, the marginal examples receive more and more weights. In SVM, the marginal examples serve as support vectors.

In our previous work, we have shown that the active basis model can learn more meaningful template and achieve better classification performance than adaboost when the sample size is relatively small (e.g., less than 100). During the past year, we also investigate the issue whether the active basis model can indeed generate realistic images of objects. For this purpose, we fit the active basis model to images from various categories, and we then synthesize images by the linear

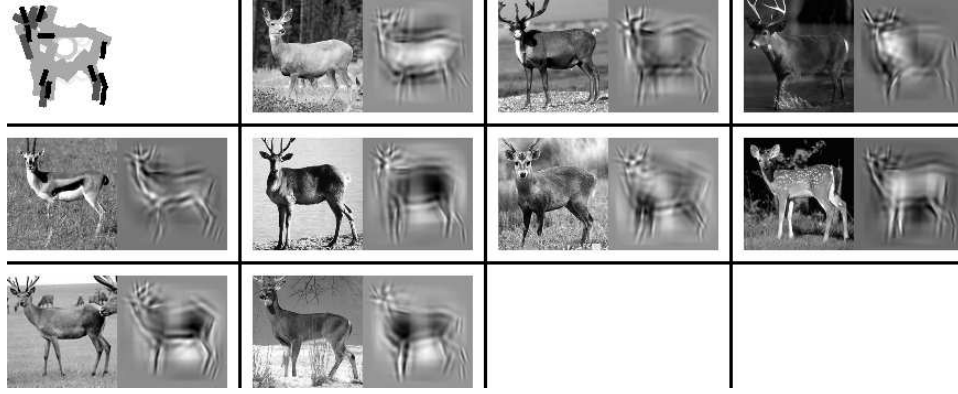


Figure 3: The first block shows the learned active basis consisting of 50 selected Gabor and DoG elements. The smaller Gabors are illustrated by darker bars. The DoG elements are illustrated by circles. The remaining blocks display the original images and the corresponding reconstructed images by the active basis model. The image size is  $102 \times 100$ .

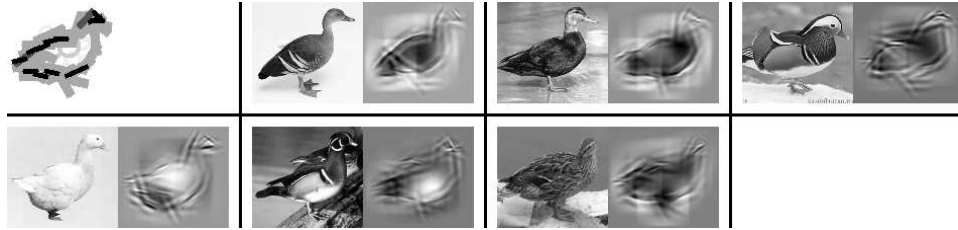


Figure 4: The first block shows the learned active basis consisting of 40 selected Gabor and DoG elements. The remaining blocks display the original  $100 \times 110$  images and the corresponding reconstructed images.

combinations of the wavelet elements. These wavelet elements are selected from a dictionary of Gabor wavelet elements at different scales, locations, and orientations. The dictionary also includes different of Gaussian (DoG) wavelets at different scales and locations. Figures (3) and (4) show some examples. Clearly, the active basis model captures some important features of the observed images, and the model is reasonably realistic.

## II. Learning active basis model from non-aligned images

If the objects in the training images are roughly aligned, like those in Figure (1), the learning can be accomplished by a simple shared sketch algorithm that we have developed. However, it is often the case that objects in the training images appear at unknown locations and scales. We develop a EM-like algorithm for learning active basis model from such training images. The algorithm only needs to know the bounding box of the object in the first image. Then it iterates between a detection step and a supervised learning step. It usually converges in a few iterations.

Figure (5) shows some examples, where each Gabor wavelet element is symbolically represented

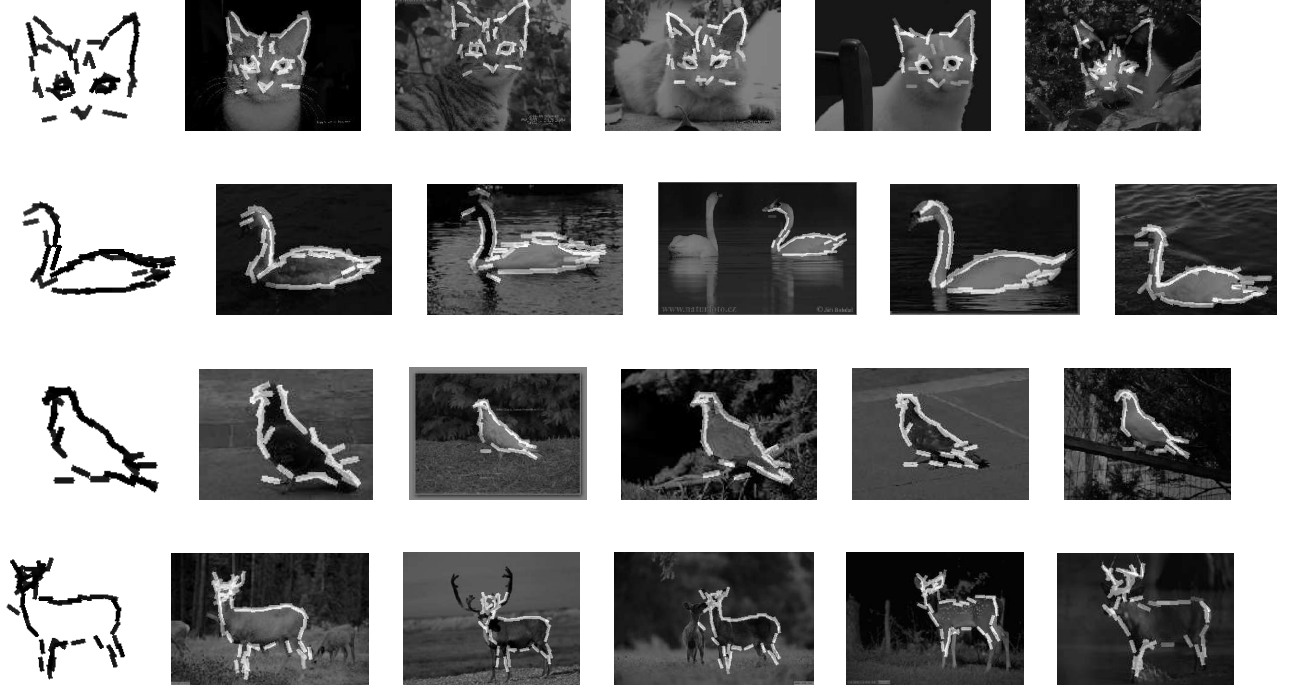


Figure 5: The bounding box of the first image is given. (1) Cats: The size of the bounding box is  $136 \times 140$ . The number of elements in the active basis is 60. (2) Swans: The bounding box is  $129 \times 178$ . Number of elements is 50. (3) Pigeons: The bounding box is  $103 \times 129$ . Number of elements is 30. (4) Deers: The bounding box is  $143 \times 149$ . Number of elements is 50.

by a bar.

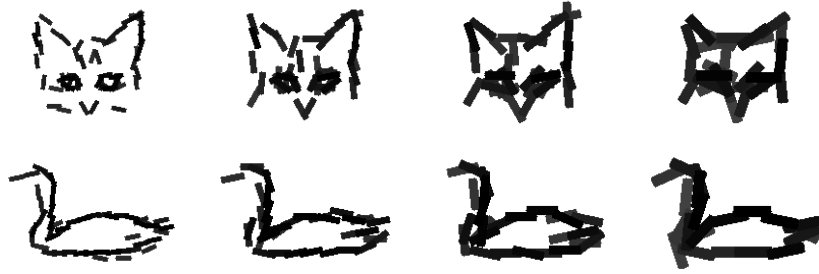


Figure 6: Multi-scale templates. The lengths of the Gabor wavelets are 17, 25, 33, 39 respectively. Cat: Number of elements at the lowest scale is 60. The numbers of elements are inverse proportional to the scales. Swan: Number of elements at the lowest scale is 50.

We can also learn templates at different scales. Figure (6) displays two examples of multi-scale templates.

We also experiment with the algorithm where we use the whole image of the first example to initialize the algorithm, so that there is no need to know the bounding box of the object in any

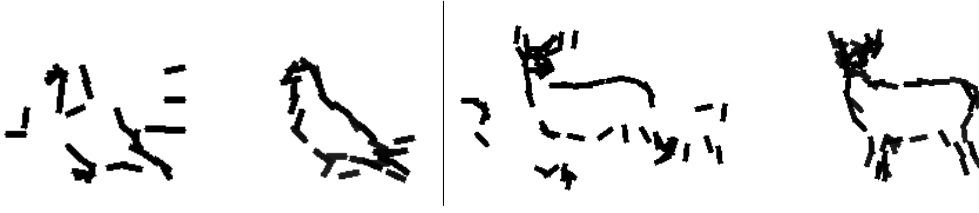


Figure 7: In each example, the first template is the starting template. The second template is learned after 5 iterations. The number of elements of the active basis is 30 in the left example, and 60 in the right example.

image. Figure (7) shows two examples.

Our method can also be used to learn part-templates. For articulated objects or objects with large deformations, we can model them as compositions of part-templates.

### III. Local learning of active basis model

Our argument to support generative model in Part I is based on a crucial assumption, that is, the generative model is realistic. If this is not the case, then the performance of the generative model can deteriorate badly. In comparison, the discriminative approach assumes much less than the generative approach, and often shows much more robust performance.

While a single active basis model may fail to represent the whole sample of training images, it is nonetheless possible to pave the whole training sample with multiple active basis models. Intuitively, we may learn multiple templates for the training images, where each template only represent a small subset of examples. This idea can be casted as a mixture model, although it is unnecessary for each mixture component to form a separate cluster. For instance, if the training images are cars at multiple view points, then each template may represent images at one view points. These templates together pave the whole range of view points.

Viewed in the high dimensional space of training images, each generative model spans a local portion of the image ensemble, although the concept of locality is closely connected to the particular generative model to be learned. We may also consider the training images as residing on a “manifold.” Each generative model represents a local “patch” that can be embedded into the “manifold.” Or each generative model abstracts local dimensions (possibly non-linear) of the “manifold.” Therefore, we may learn an active basis model around each training image. Then we can trim and combine such locally learned active basis models to pave the whole training sample.

The above scheme is reasonable for the following two reasons. First, unlike the models in physics which can generalize to a vast range, a generative model in vision can only be expected to generalize within a limited range. So the generative model has to be local. Second, it is possible to learn a generative model locally because of efficiency of likelihood-based learning, as we discussed in Part I. Intuitively, learning a generative model by maximum likelihood is like “grasping” the high-density portion of the data, and such generative learning is very suitable for identifying informative local dimensions of the “manifold” formed by the training images.

We develop a EM-like iterative procedure for local learning of generative models.

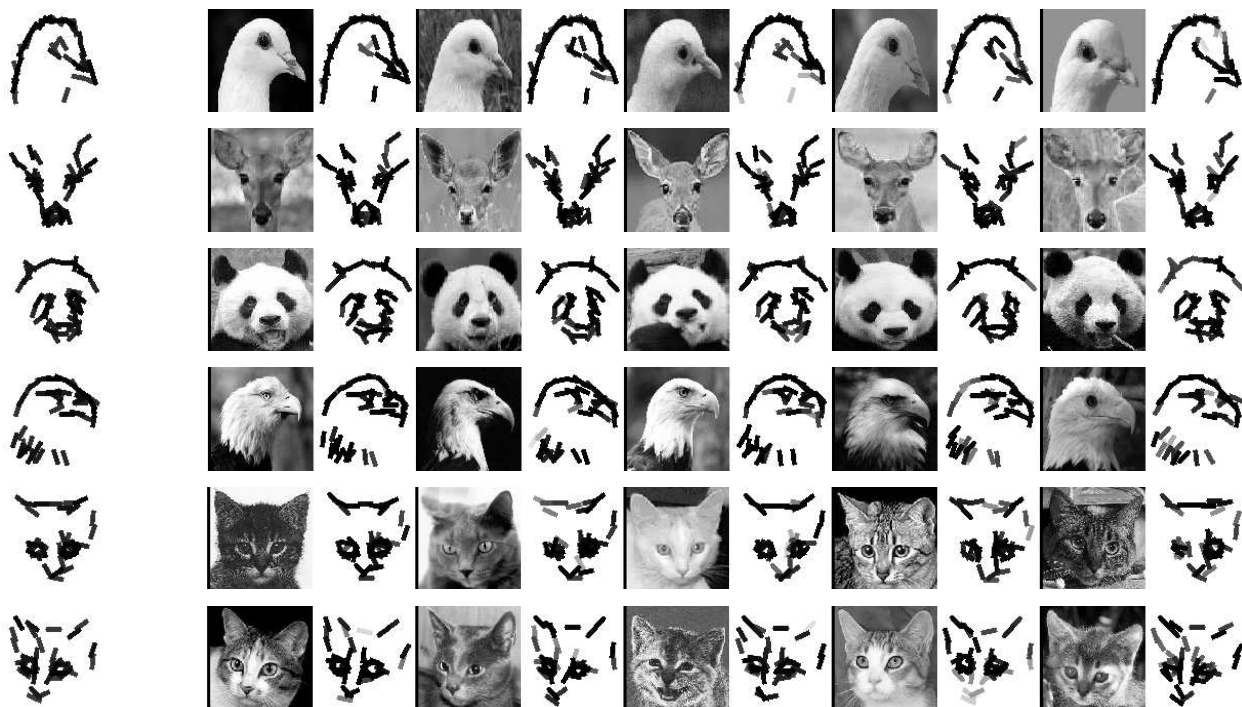


Figure 8: Some locally learned templates and their neighboring examples.

Figure (8) shows an example of local learning from images of animal heads. The figure displays some locally learned templates and their neighboring examples. The algorithm does not know beforehand which image belongs to which type of animal head. But the algorithm is able to find local clusters by itself. Some of the local clusters are distinct, but some are not. For example, the last two clusters represent the cat heads turned to slightly different directions.

The local learning of generative model can be contrasted against the discriminative approach such as adaboost and SVM. Unlike adaboost, which selects a fix set of features with a fixed set of weights globally, the local learning abstracts the dimensions locally. Unlike SVM which selects a set of support vectors and generalize locally by kernel functions, the local learning generalizes by generative models with much reduced dimensions that are learned from a local neighborhood.

Work in (I), (II), (III) is reported in a 60-age paper accepted by *International Journal of Computer Vision*, and a paper accepted by *Statistical Science*.

#### IV. Learning mixed template of sketches and textures

The active basis model captures the edges and also regional contrasts in images of objects. But it does not capture the surface properties such as smoothness and textures. We develop a generative model for mixed template that integrates both sketch variables for shapes and texture variables for appearances. Both the sketch variables and texture variables are based on responses of Gabor wavelets. The sketch variables are in the form of local maxima of Gabor responses,

whereas the texture variables are in the form of local averages of Gabor responses. In fact, the sketch variables and the texture variables form a natural couple, where the strengths of sketch variables are measured against the corresponding texture variables. Such a unified treatment of sketches and textures is elegant and is shown to improve the classification performance in a number of experiments. The work is reported in a paper accepted by CVPR (Computer Vision and Pattern Recognition) 2009, a peer-reviewed top conference on pattern recognition.

## **V. Learning animated basis for action modeling**

The active basis model is developed for static images. It can be naturally extended to model dynamic video sequences, such as horse running and bird flying. We call the resulting model the animated basis model, which is a sequence of active basis models ordered in time. In each active basis model, each constituent wavelet element is endowed with a speed along its normal direction. The model can be learned from training video sequences, by a method that is very similar to that in Part II, except that we also need to align the templates over time by dynamic programming. The work is reported in a paper accepted by ICCV (International Conference of Computer Vision) 2009, a peer-reviewed top conference on computer vision.

## **2. Describe the opportunities for training and development provided by your project**

(1) The research in this project represents the state of art in statistical modeling, learning and computing in computer vision. It provides opportunities for students to learn modern statistics while working on an important scientific problem.

(2) The work requires considerable amount of programming. It provides opportunities to train students in programming in matlab and C. All the source code has been posted on the reproducibility pages that we have constructed.

(3) Students receive training in conducting critical review of the current literature, and in communicating and presenting their work. Zhangzhang Si will present the work on learning mixed template in CVPR. Benjamin Yao will present the work on animated basis in ICCV.

(4) Students receive training in reproducible research. Currently in the field of computer vision, reproducibility is in general not enforced, which we consider a major limitation that impedes the real progress of the field. We have tried very hard to persuade students to construct a reproducibility page for each paper they submit or publish, where all the figures, plots, tables, and numbers reported in the paper should be accompanied by the data, code, and readme files that reproduce them. Now even in the on-going projects, we sometimes communicate with each other by such pages.

## **3. Describe outreach activities your project has undertaken**

(1) The PI (Wu) served as an opponent in the thesis defense of David Gustavssonin of the Department of Computer Science at the University of Copenhagen in June 2009. Part of the thesis is based on the PIs' previous work.

(2) The PI (Wu) gave a discussion to the paper by Storlie, Lee, Hannig and Nychka on multiple



target tracking, during the annual symposium of the International Chinese Statistical Association, held in June 2008, New Jersey. The discussion has been published in *Statistica Sinica*.

(3) The PI (Wu) served on the program committee of the First International Workshop on Stochastic Image Grammars, held on June 2009, Miami, Florida. The co-PI (Zhu) served on the organizing committee of this workshop.

(4) The PI (Wu) served as an associate editor for *Statistica Sinica*.

(5) The PI (Wu) was invited to serve as an associate editor for *JASA* applications and case studies, starting in January 2010.

(6) The PI (Wu) served on an NSF panel in June 2009.

(7) The co-PI (Zhu) organized a China-US-France Summer School on Machine Learning, Statistics, and Computer Vision, from June 30 to July 11, 2008, at the Lotus Hill Institute in Hubei, China.

(8) The co-PI (Zhu) was awarded the J. K. Aggarwal Prize by the International Association for Pattern Recognition, and gave a speech during the 19th International Conference on Pattern Recognition.

(9) The PI (Wu) served as the mentor of Yunlong He for the UCLA CSST (cross-disciplinary scholars in science and technology) program during the summer of 2008. Yunlong He was a undergraduate student from University of Science and Technology of China. He will start his Ph.D. study in Georgia Tech.

## Publications and Products

### 1. What have you published as a result of this work?

Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2009) Learning active basis model for object detection and recognition. *International Journal of Computer Vision*, accepted.

Si, Z., Gong, H., Zhu, S. C., and Wu, Y. N., (2009) Learning active basis models by EM-type algorithms. *Statistical Science*, in press.

Si, Z., Gong, H., Wu, Y. N., and Zhu, S. C. (2009) Learning mixed template for object recognition. *Proceedings of Computer Vision and Pattern Recognition*.

Zhu, S. C., Shi, K., Si, Z., and Wu, Y. N. (2009) Learning explicit and implicit visual manifolds by information projection. *Pattern Recognition Letter* (under review).

Yao, B. and Zhu, S. C. (2009) Learning deformable action templates from crowded videos. *Proceedings of International Conference on Computer Vision*.

Yang, X., Wu, T. and Zhu, S. C. (2009) Evaluating information contributions of bottom-up and top-down processes. *Proceedings of International Conference on Computer Vision*.

### 2. What web sites or other internet sites have you created?

We have constructed the following reproducibility pages for our work.

(1) The reproducibility page on active basis for the paper that has been accepted by *International Journal of Computer Vision*. The source code is in matlab and C.

<http://www.stat.ucla.edu/~ywu/AB/ActiveBasisMarkII.html>

(2) The reproducibility page on active basis for the paper that has been accepted by *Statistical Science*.

<http://www.stat.ucla.edu/~zzsi/ActiveBasis.html>

(3) The reproducibility page on mixed template for the paper that has been accepted by CVPR 2009.

[http://www.stat.ucla.edu/~zzsi/mixed\\_template.html](http://www.stat.ucla.edu/~zzsi/mixed_template.html)

(4) The reproducibility page on animated basis for the paper that has been accepted by ICCV 2009, under construction.

### **3. What other specific products (databases, physical collections, educational aids, software, instruments and the like) have you developed?**

The co-PI (Zhu) has been maintaining a comprehensive database on image parsing. The following is the webpage.

<http://www.imageparsing.com/FreeDataOutline.html>

This is a major effort led by the co-PI (Zhu). The goal is to create challenging image databases and benchmarks with high quality human annotated ground truth. Such a database can be very useful for learning statistical models for object patterns, and for testing the performance of such models.

The data can be downloaded from the webpage mentioned above, after registration.

## **Contributions**

### **Contributions within discipline**

*Contributions to statistics and vision.*

Vision is a phenomenon that is still far from being well understood. At the core of this phenomenon is how the knowledge of the wide variety of visual patterns are represented and learned, and how such representations can be used for the purpose of recognition. It is believed by many that the patterns can be represented in the form of generative models, and the learning and recognition can be guided by the likelihoods of the generative models. However, after decades of research, there still have not been generative models that can synthesize realistic image patterns and that can be easily computed in training and recognition.

Our work on active basis model is an attempt to develop such a model. We show that this model is capable of synthesizing reasonably realistic image patterns. We show that the model can be learned from images where the objects may appear at different locations and scales. We also show that the model can be learned locally and can be used to capture local dimensions of the image manifold. The inferential algorithm of the model is closely related to the functions of simple and complex cells in primary visual cortex.

The active basis model is a statistical model. Although the model is based on a linear structure of wavelet expansion, it is highly non-Gaussian in the distributions of wavelet coefficients and is highly non-linear because of the perturbations in the locations and orientations of the wavelet elements. The selection of basis elements is based on a coupling of matching pursuit and projection pursuit. The active basis model is an advance in the art of statistical modeling.

### **Contributions to other disciplines**

*Contribution to neuroscience.* The active basis model is based on mathematical theories of simple and complex cells in primary visual cortex. The recognition algorithm is based on a cortex-like structure of sum and max maps. This computational architecture is biologically plausible, and enables researchers in neuroscience to ask relevant questions and design experiments to investigate visual cortex beyond primary visual cortex. The PI (Wu) has had several interesting conversations with Tai Sing Lee, a neuroscientist in CMU, on how to test the plausibility of the active basis model.

*Contribution to wavelets and harmonic analysis.* The active basis model is a direct consequence of applying wavelet sparse coding on a set of training images from the same object category. In the past, the wavelet models are often guided by the generic principle of sparsity. Although such a generic principle has immense scope of applications, sparsity alone is clearly inadequate for representing the wide varieties of patterns in natural scenes. The active basis model is a simple but important step forward, where sparsity is replaced by more specific prior distributions so that the wavelet model can be used to represent more concrete patterns.