Adjusting Active Basis Model by Regularized Logistic Regression

Ruixun Zhang
Department of Statistics and Probability
School of Mathematical Sciences
Peking University

Mentor: Prof. Ying Nian Wu
Supervisor: Zhangzhang Si
Dept. of Statistics
University of California, Los Angeles

Abstract

Active basis model is a generative model seeking a common wavelet sparse coding of images from the same object category, where the images share the same set of selected wavelet elements, which are allowed to perturb their locations and orientations to account for shape deformations. This work applies discriminative methods to adjust λ's of selected basis elements, including logistic regression, SVM and AdaBoost. Results on supervised learning show that discriminative post-processing on active basis model improves its classification performance in terms of testing AUC. Among the three methods the L2-regularized logistic regression is the most natural one and performs the best.

1 Methods

We use active basis model [1] to learn a template of size 80 ($B_1,\ldots,B_{80}$), with local normalization of filter response, and then adjust λ's of selected basis elements based on MAX1 scores after sigmoid transformation using discriminative criteria. After a great dimension reduction by active basis (from about 1 million features down to only 80), the computation is fast for discriminative methods.

We use generative model for unsupervised learning in the presence of hidden variables (unknown subcategories, locations, poses, scales, and perturbations). Then

we re-estimate λ's by fixing the inferred hidden variables in learning as well as selected basis elements. As a consequence of generative model, we fit a flat logistic regression. With hidden variables given and basis elements selected, the learning becomes supervised.

Learning in active basis model corresponds to full likelihood p(image, class) under conditional independence assumption where we only learn from positives, while the logistic regression corresponds to partial likelihood p(class | image) without conditional independence assumption where we use both positives and negatives. The logistic regression helps correct the conditional independence assumption in generative model.

1.1 Logistic regression. We use logistic regression from liblinear [2, 3] with L2-regularization. The model is

$$P(y = \pm 1) = \frac{1}{1 + \exp(-y(b + \boldsymbol{\lambda}^T \mathbf{x}))}$$

where y is the label of an image (0 or 1), **x** are selected (by Active Basis model) MAX1 scores after sigmoid transformation, $\boldsymbol{\lambda}$ is the regression coefficient and *b* is the intercept term. The loss function is

$$\frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{\lambda} + C\sum_{i=1}^{l}\log(1 + e^{-y_i\boldsymbol{\lambda}^T\mathbf{x}_i})$$

where the intercept term is included in the regularization term.

However, we want L2-regularization (corresponding to a Gaussian prior) without the intercept term, so we modified the codes slightly to make loss function:

$$\frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{\lambda} + C\sum_{i=1}^{l}\log(1 + e^{-y_i(\boldsymbol{\lambda}^T\mathbf{x}_i + b)})$$

In the training process, each image has equal data weight 1. Also, classification performance is not too sensitive w.r.t the tuning parameter C when C is small. So C is set to 0.01 in the experiment.

Modifications in the code (Lin, personal communication): In the following 3 functions:

- l2r_lr_fun::fun
- l2r_lr_fun::grad

- l2r_lr_fun::Hv

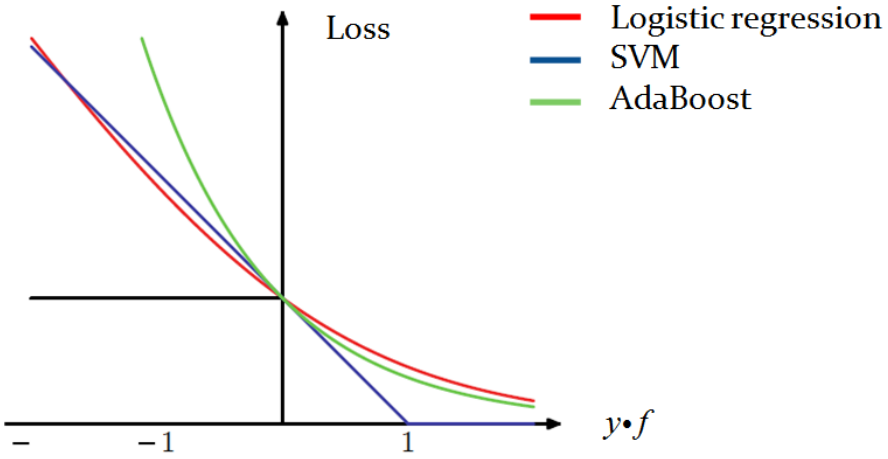
replace i<w_size with i<w_size-1

1.2 SVM. We use SVM [4] from SVM-light [5] with bias term and linear kernel. Classification performance is not sensitive at all w.r.t the tuning parameter C, which is then set to 1 in the experiment.

1.3 AdaBoost. We use AdaBoost [6] which is exactly the same as in experiment 3.

1.4 Logistic regression VS other methods

The following figure [10] shows loss functions of the above-mentioned methods.



Generally there are 2 common ways to add regularization: L1-regularization and L2-regularization. Friedman [10] points out L1-regularization is preferred when the goal is to find a sparse representation, but since in our case basis elements have been already selected in generative learning, we want to use L2-regularization for smoothness instead of sparsity.

Furthermore, L2-regularized logistic regression is similar to SVM, and L1-regularized logistic regression is similar to AdaBoost [8, 9, 10, 13]. This is shown in the above figure, where the cost functions of the three methods are similar for 0-1 losses. Logistic regression is readily formulated in likelihood-based learning and inference, where the joint probability of data and label is trained towards good classification performance. While AdaBoost and SVM adopt a smartly designed cost function (exponential loss in the case of AdaBoost, and margin in the case of SVM), instead of the generic probability form. So we say logistic regression is more natural than the other two methods. In our experiment, we find that logistic regression consistently perform the best.

2 Classification Experiment

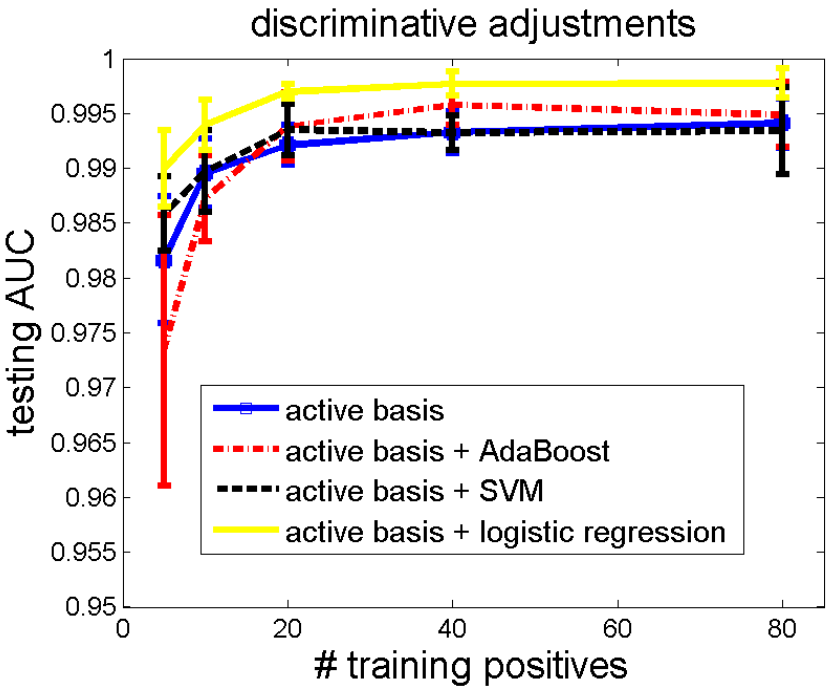
General learning problem should be unsupervised, but I work on supervised learning as a starting point.

2.1 Dataset. For the **head_shoulder** dataset, we tried 4 methods for classification:

- active basis with template size 80,
- active basis + adjustment by logistic regression,
- active basis + adjustment by SVM,
- active basis + adjustment by AdaBoost.

The following figure shows several positive examples in the head_shoulder dataset.

2.2 Results. Template size 80, training negatives 160, testing negatives 471. In total, 5 repetitions (randomly split the data) * 4 methods * 5 numbers of positive training examples (5, 10, 20, 40, 80) are tested. Testing AUC is plotted below. Logistic regression is the only method consistently improved active basis model.

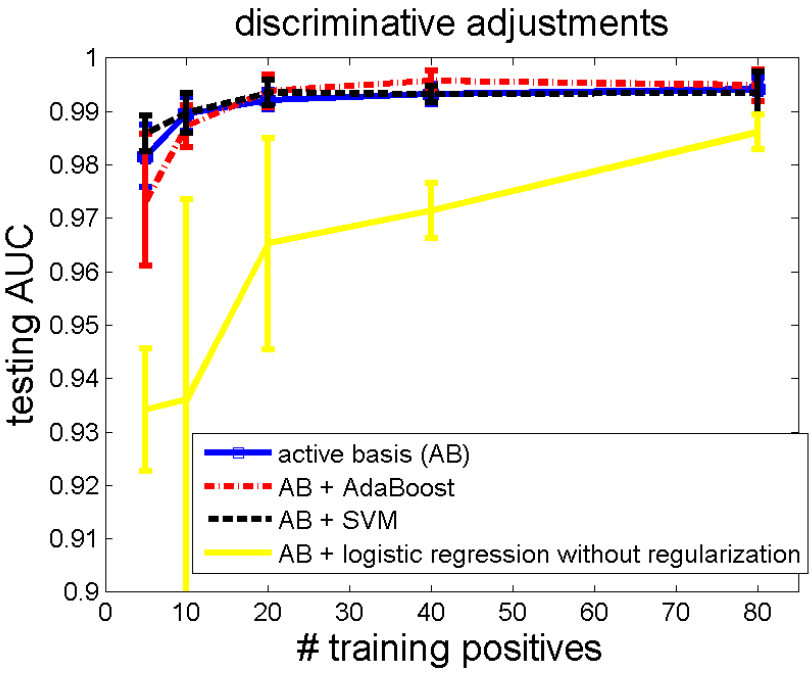


2.3 Computing time. The feature selection step in active basis model has greatly reduced the dimension of an image from thousands of pixels to several basis elements (in this case 80), so the computing time for discriminative adjustments is short. The table below shows the time of one active basis learning, and one logistic adjustment after SUM1 and MAX1 step.

Intel Core i5 CPU, RAM 4GB, 64bit windows		
# pos	**Learning time (s)**	**LR time (s)**
5	0.338	0.010
10	0.688	0.015
20	1.444	0.015
40	2.619	0.014
80	5.572	0.013

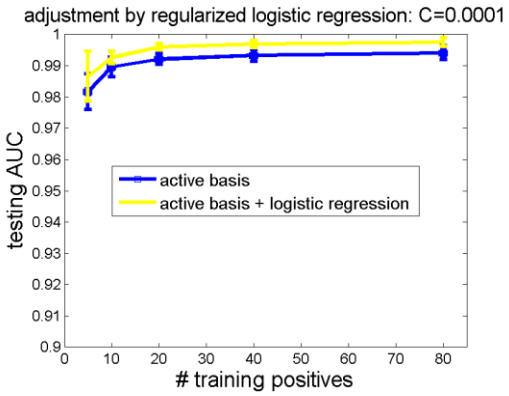
3 Sensitivity of Tuning Parameter

We have tried logistic regression without regularization on the feature selected by active basis model. But the performance keeps worse than pure active basis model, even after we re-weight the sample during learning. The following figure shows the testing AUC, where logistic regression is from MATLAB.

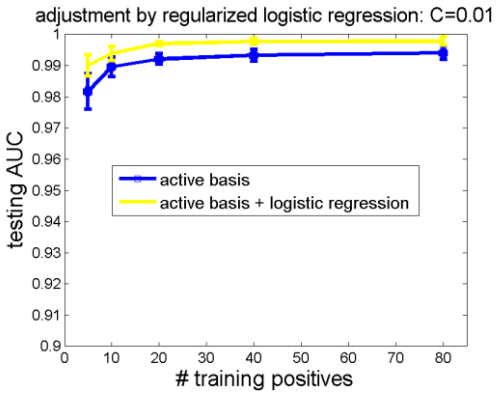


We doubt that logistic regression suffers from overfitting, which is the very motivation of adding a L2-regularization term. In order to verify this, we test different tuning parameters for L2-regularized logistic regression. See figure below.

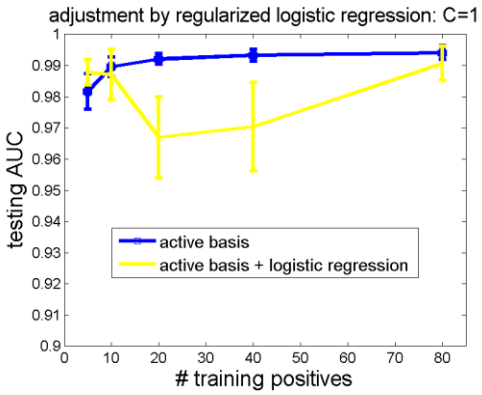
C=0.0001 (regularization is high)



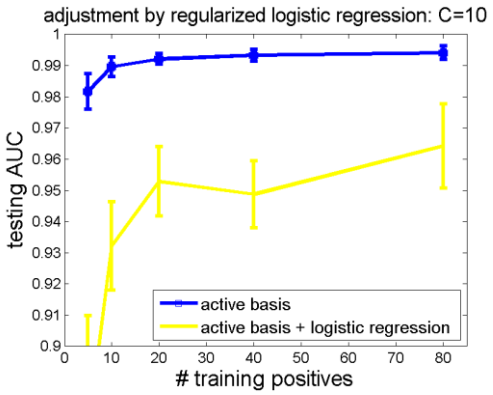
C=0.01



C=1



C=10 (almost no regularization)

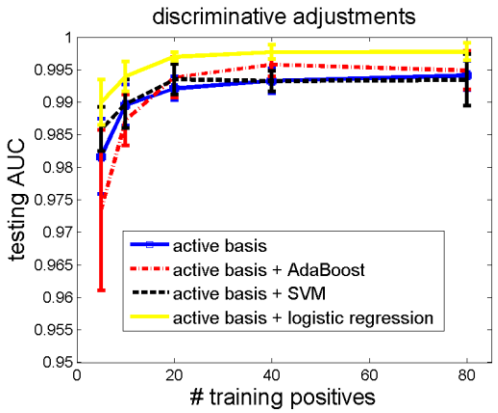


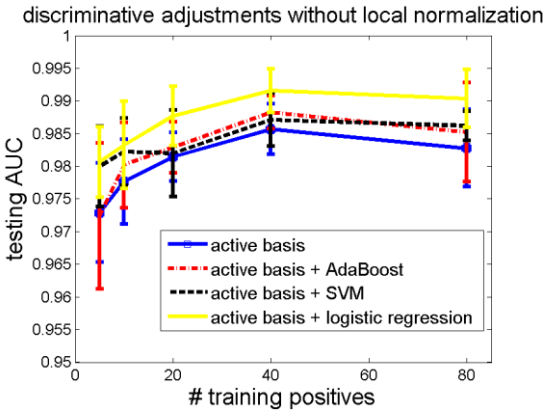
The yellow line is the testing AUC after adjustment by logistic regression. Conclusions are:

- Smaller tuning parameters give better classification performances. Because small tuning parameters imply high level of regularization, this result provides evidence of overfitting in logistic regression without regularization.
- Testing AUC remains stable when tuning parameter is 0.01 or less. In other words, the model is not sensitive to tuning parameter when it is small enough. Therefore in experiments we just set C = 0.01.

4 With or without local normalization

Currently local normalization for filter response is included. The following 2 experiments compare local normalization with no local normalization. It is clear that local normalization helps classification a lot.

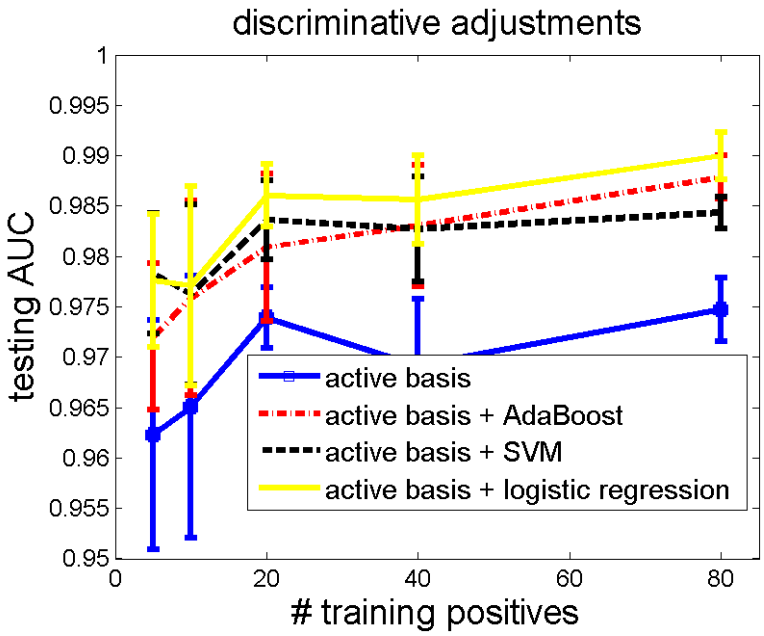




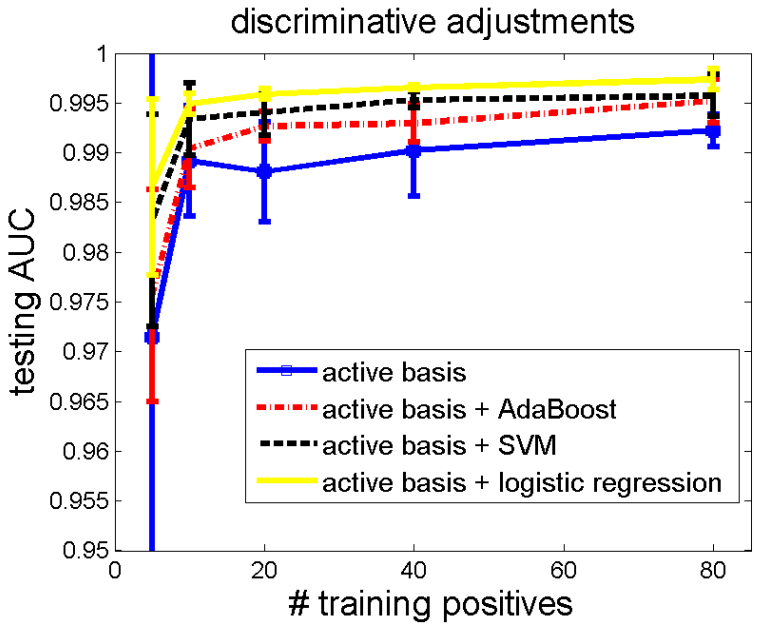
5 Experiments on More Datasets

We repeat the classification experiment for other datasets. The following figures show similar results as in head_shoulder.

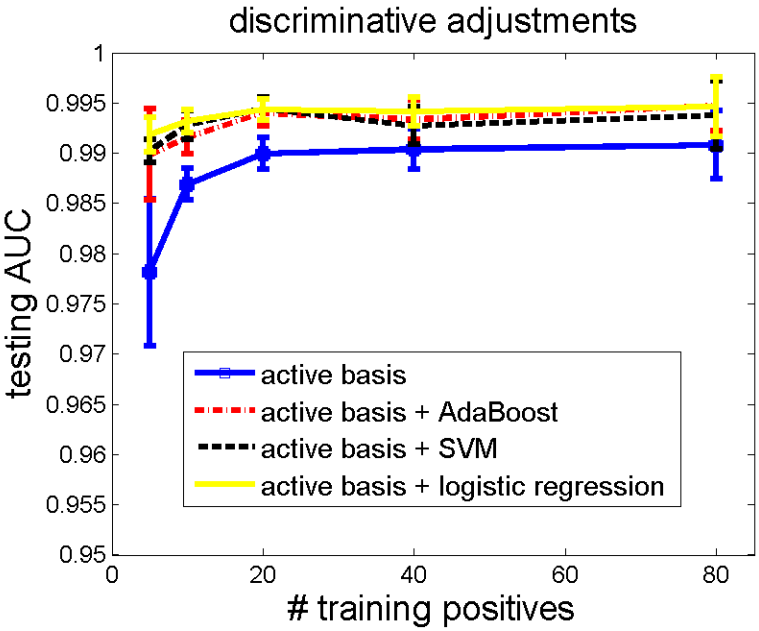
5.1 Horse. Template size 80, training negatives 160, testing negatives 471.



5.2 Guitar. Data is from Caltech 101 [12]. template size 80, training negatives 160, testing negatives 855.

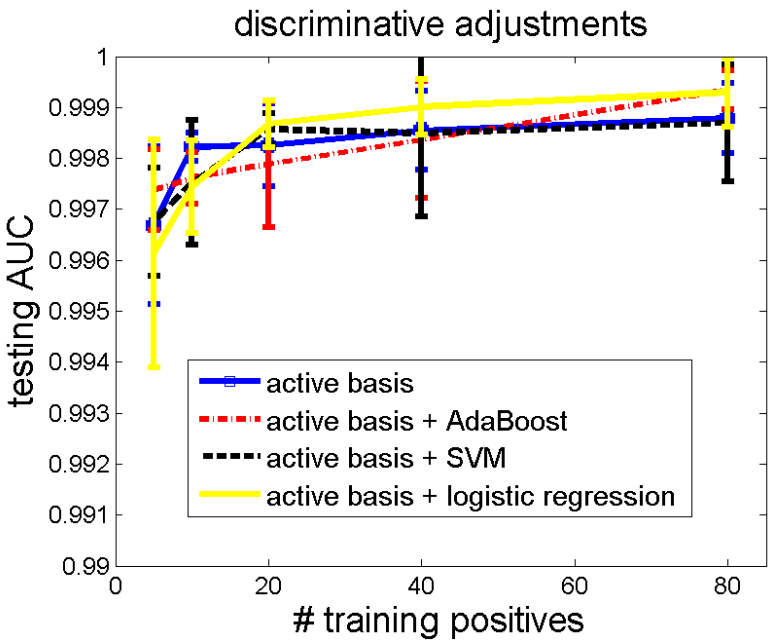


5.3 Motorbike. Data is from Caltech 101 [12]. template size 80, training negatives 160, testing negatives 855.



5.4 Butterfly. Template size 80, training negatives 160, testing negatives 471.

Since testing AUC of the butterfly dataset is already over 99.5% for active basis, it is hard for discriminative methods to further improve performance. However, logistic regression is still the best one.



6 More Comments

This project works on supervised learning as a starting point. Its value lies in the promising future for extending to unsupervised learning, rather than merely high performances.

For unsupervised learning, general picture remains the same. We apply generative learning by active basis because it is good at discovering hidden variables, and then discriminative adjustment to tighten up the parameters and improve classification performances.

7 Acknowledgements

Thanks to my mentor Prof. Ying Nian Wu and PhD fellow Zhangzhang Si, for their instructions to my project, as well as to my self-developing. I have learned a lot from the project. It is a fantastic summer for me. Also thanks to Dr. Chih-Jen Lin for his liblinear software package and his detailed suggestions about how to adjust the software for our experiment.

References

[1] 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.*

[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research.*

[3] Lin, C. J., Weng, R.C., Keerthi, S.S. (2008). Trust Region Newton Method for Large-Scale Logistic Regression. *Journal of Machine Learning Research.*

[4] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer.

[5] Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning,* B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press.

[6] Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences.*

[7] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision.*

[8] Rosset, S., Zhu, J., Hastie, T. (2004). Boosting as a Regularized Path to a Maximum Margin Classifier. *Journal of Machine Learning Research.*

[9] Zhu, J. and Hastie, T. (2005). Kernel Logistic Regression and the Import Vector Machine. *Journal of Computational and Graphical Statistics.*

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Elements of Statistical Learning; Data Mining, Inference, and Prediction.* New York: Springer.

[11] Bishop, C. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

[12] L. Fei-Fei, R. Fergus and P. Perona. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE. CVPR, Workshop on Generative-Model Based Vision.*

[13] Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Ann. Statist.*