A Novel OLS Algorithm for Training RBF Neural Networks with Automatic Model Selection

Peng Zhou
Institute for Pattern Recognition & artificial Intelligence
Huazhong University of Science and Technology
Wuhan, 430074, China
zhoupengsirok@163.com

Dehua Li ,Hong Wu and Feng Chen
Institute for Pattern Recognition & artificial Intelligence
Huazhong University of Science and Technology
Wuhan, 430074, China
lgf1007@126.com

***Abstract*—Orthogonal Least Squares（OLS）algorithm has been extensively used in basis selection problems for RBF networks, but it is unable perform model selection automatically because the user is required to specify the tolerance ρ, which is relevant to noises and will be difficult to implement in the real system. therefore, a generic criterion that defines the optimum number of its basis function is proposed. In this paper, Not only is the Bayesian information criteria (BIC) method incorporate into the basis function selection process of the OLS algorithm for assigning its approprite number, but also we develop a new method to optimize the widths of Gaussian functions in order to improve the generalization performance, The augmented algorithms are employed to the Radial Basis Function Neural Networks (RBFNN) to compare its performance for known and unknown noise nonlinear dynamic systems, Experimental results show the efficacy of this criterion and the importance of a proper choice of basis function widths.**

Keywords- orthogonal least squares; radial basis function networks; Bayesian information criteria; kernel widths

I. INTRODUCTION

Radial Basis Function Neural Networks (RBFNN), which mimics the function of the human brain, has been employed in various applications involving classification or function approximation. Due to its tolerance of noise, ambiguity, distortedness and incompleteness of data from the real world; it is superior to statistical approaches [1]. The radial basis function network offers a viable alternative to the two-layer neural network which is linear in the parameters by fixing all RBF centers and nonlinearities in the hidden layer. Thus the hidden layer performs a fixed nonlinear transformation with no adjustable parameters and it maps the input space onto a new space. The output layer then implements a linear combiner on this new space and the only adjustable parameters are the weights of this linear combiner, which can determined by the linear least squares (LS) or Recursive Least Squares (RLS) method. The nonlinearity of an RBF network can be chosen from a few typical nonlinear functions. obviously, the choice of functions has not a critical influence on the performance of RBFNN and the opinion can also be justified by MJD Powell [2]. The performance of RBFNN is mainly dependent of the chosen centers and corresponding standard deviation (widths of Gaussian functions). A common learning algorithm for RBFNN is based on first choosing randomly some data points as radial basis function centers and then using singular value decomposition to solve for the weights of the network. Such a procedure has several drawbacks, in particular, an arbitrary selection of centers is clearly unsatisfactory, and this resulting RBFNN often either perform poorly or have a large size. The k-means clustering proposed by Moody and Darken can be applied to allocate the centers of the Gaussian functions in the input space reasonably. However the clustering problem has in general many local minima and also the local data distributions mapped onto each of the k means are no stationary due to the fact that the cluster region boundaries are shifting [3]. The orthogonal least squares (OLS) method proposed by S. Chen[4] can be employed as a forward regression procedure to select a suitable set of centers (repressors) from a large set of candidates and the ill-conditioning problems occurring frequently in random selection of centers can automatically be avoided, but the algorithm have two serious limitations. The first one is that the tolerance ρ must be pre-known, the sum of error reduction ratio monotonically increases with increasing number of the chosen centers, and thus cannot detect a correct number of the hidden layer. However, a bad estimation of it cannot be sufficient to guarantee good generalization performance. There is no theoretical guide available for solving the problem, except for the heuristic method. The other limitation is that the formulation implies that the widths of Gaussian function are uniform. Though they maybe are less important than the positions of Gaussian kernels, some special case apparently deviates from many practical situations.

In this paper, by combing Bayesian information criteria with an efficient forward subset selection procedure, selecting the appropriate number of the hidden layer in an RBF networks is performed automatically, moreover, a new submodel to optimize the widths of Gaussian functions is developed. The remainder of the paper is organised as follow. Section 2 presents the RBF network architecture and understands how it works. The major approaches adopted to train this model is discussed, One of them, the novel OLS algorithm , is detailed in Section 3, the other one , a new method on the kernel widths, is discussed in Section 4.The

978-1-4244-7237-6/10/$26.00 ©2010 IEEE

section 5 shows functional approximation experiments, our conclusions are exposed in the section 6.

II. THE RADIAL BASIS FUNCTION NETWORK

On the whole, a RBFNN is any network that has radial symmetric activation functions. The output of a hidden neuron is a function of the distance between an input vector and the centre of the function. The schematic of the RBFNN with n inputs and a scalar output is depicted in Fig.1. Given an input vector of dimension p, $x \in \Re^p$, the output of the network is described by (1)

$$f(x) = w_0 + \sum_{i=1}^{m} w_i \varphi_i(\| x - c_i \|) \tag{1}$$

Where w_i (i=0,1…m) are the network weights, $x \in \Re^p$ is the input vector, $\|.\|$ denotes the Euclidean norm, $c_i \in \Re^p$, $1 \le p \le m$, are the basis function centers, $\varphi(.)$ is a given function from $\Re^+$ to $\Re$, in this work the Gaussian exponential function was used(2),where σ_i defines the width of the receptive field

$$\varphi_i : \Re^p \to \Re \quad \varphi_i(x) = \exp(-\frac{\| x - c_i \|^2}{\sigma_i^2}) \tag{2}$$

To understand how this works, it is essential to view the RBFNN as a special case of the linear regression models; the corresponding geometric interpretation is best revealed by the following matrix form

$$d = p\theta + E \tag{3}$$

Where

$$\begin{bmatrix} d(1) \\ d(2) \\ \vdots \\ d(N) \end{bmatrix} = \begin{bmatrix} p_1(1) & p_2(1) & \cdots & p_M(1) \\ p_1(2) & p_2(2) & \cdots & p_M(2) \\ \vdots & \vdots & \ddots & \vdots \\ p_1(N) & p_2(N) & \cdots & p_M(N) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix} + \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{bmatrix}$$

The regressor vectors p_i form a set of basis vectors that are $\varphi_i(.)$, the solution $\hat{\theta}$ satisfies the condition that $p\hat{\theta}$ be the projection of d onto the space spanned by these basis vectors , the E is a white noise process representing for example the observation noises. It is apparent that a fixed center c_i with a given nonlinearity $\varphi(.)$ corresponds to a repressor p_i, and then the problem of how to select a suitable set of RBFNN centers from the data set can be regarded as an example of how to select a significant subset from a given candidate set of regressors.

III. THE NOVEL OLS ALGORITHM

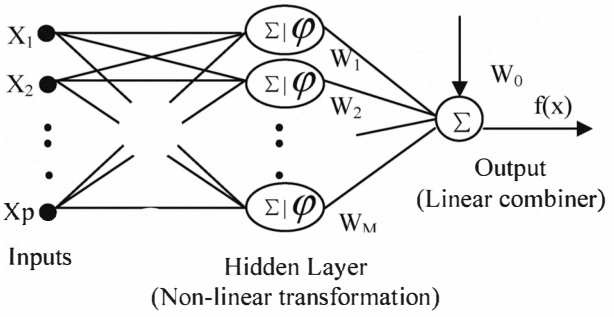

Figure 1. schemetic of RBFNN.

It is possible to interpret the hidden layer configuration of a RBF network as subset selection problem. Some approaches have various shortcomings in the literature. The common and the most crucial one is that the number of basis functions has to be given a priori. However,the orthogonal least squares algorithm (OLS) presented by Chen et al. [5] is not, but its tolerance ρ must be pre-known and fixed, which ideally should be very close to the noise to noised-signal ratio $\sum n_k^2 / \sum d_k^2$ [6](where the quantity d_k^2 is known from the measured data d , an appropriate estimation of n_k^2 that is the energy of noises from the d), therefore, for real-world applications, it is necessary to consider some heuristic procedure to rationally explore this search space, to overcome the drawbacks, two proposals for dealing with this problem are presented in the section.

A. Automated model selection

Chen's Learning is stopped once the terminating criterion is satisfied. Actually we find that Bayesian information criteria (BIC) method always detect a correct number of the hidden units in noise nonlinear-dynamic systems in all attempt, the experimental results on it are given in the paper. The BIC is a classical model selection criterion and has the generic form of

$$BIC = -2\log lik + (\log N)p \tag{4}$$

Where in the case of linear models p is the dimensionality of imputes. Under the Gaussian model, Schwarz's[7] procedure can be written as:

$$BIC(M) = N\log(\sigma^2) + M\log(N) \tag{5}$$

Where σ is the standard deviation of training set and N is the number of it, M denotes rank of the model and here is number of basis function. An RBF network with too few basis function givens poor generalization on new data, on the contrary, an RBF network with too many basis function also yields poor predictions since it is maybe fits the noises in the training data, A small number of basis function yields a high bias and a low variance, whereas a large number of basis functions yields a low bias but high variance estimator. The best generalization performance is obtained via a compromise between number of basis function and variances; the minimal value of BIC is a desired trade-off.

B. Inverse of orthogonal matrix product

An important matrix dealing with forward selection of RBF networks is Inverse of orthogonal matrix product ϕ . Hence, one can write:

$$\phi_k = 1 / R_k^T R_k \tag{6}$$

Where R_{k+1} is orthogonal matrix, k denotes the column. To make the (k+1)th column orthogonal to each of the k previously orthorgonalized columns, we must construct new matrix R_{k+1}=[R_{k+1} v_{max}] (v_{max} is a new orthogonal vector which is selected latently) and compute ϕ_{k+1}. Therefore, the augmentation of ϕ is implemented as follow [8,6]:

$$\alpha = \phi_k R_k^T v_{\max} \; ; \; e = v_{\max} - \phi_k \alpha \; ; \; \beta = e^T e \tag{7}$$

$$\phi_{k+1} = \begin{bmatrix} \phi_k + \alpha\alpha^T / \beta & -\alpha / \beta \\ -\alpha^T / \beta & 1/\beta \end{bmatrix} \quad (8)$$

C. Complete OLS

The Bayesian information criteria (BIC) method offers a simple and effective mean of seeking a subset of significant regressors in forward-regression manner. The augmentation of ϕ enables a efficient reduction in running time. We will use the classical Gram-Schmidt scheme as an example. The regressor selection main procedure is summarized as follows

At the first step, for $1 \le i \le M$, compute

$$[Q]_1^{(i)} = ((p_1^{(i)})^T d)^2 / ((p_1^{(i)})^T p_1^{(i)}) \quad (9)$$

Find $[Q]_1^{(i_1)} = \max\{[Q]_1^{(i)}, 1 \le i \le M\}$ And select

$$R_k = w_1 = w_1^{i_1} = p_{i_1} \quad \phi_k = 1/(R_k^T R_k) = 1/(w_1^T w_1)$$

$$E_k = d - R_k(\phi_k(R_k^T d)) \; ; S_k^2 = E_k^T E_k / N \; ; BIC_k = NInS_k^2 + kInN \quad (10)$$

At the kth step, where $k \ge 2$ for $1 \le i \le M$,$i \ne i_1$,…, $i \ne i_{k-1}$,compute

$$\alpha_{jk}^{(i)} = \phi_j R_j^T p_i, i \le j \le k \quad (11)$$

$$w_k^{(i)} = p_i - \sum_{j=1}^{k-1} \alpha_{jk}^{(i)} R_j \quad (12)$$

$$[Q]_k^{(i)} = ((w_k^{(i)})^T d)^2 / ((w_k^{(i)})^T w_k^{(i)}) \quad (13)$$

Find $[Q]_k^{(i_k)} = \max\{[Q]_k^{(i)}, 1 \le i \le M, i \ne i_1, ..., i \ne i_{k-1}\}$, And select $v_{\max} = w_k^{i_k}$, Compute

$$R_{k+1} = [R_k v_{\max}] \quad (14)$$

By using (7), (8), and (14) it results that

$$E_{k+1} = d - R_{k+1}(\phi_{k+1}(R_{k+1}^T d)) \; ; S_{k+1}^2 = E_{k+1}^T E_{k+1} / N$$

$$BIC_{k+1} = NInS_{k+1}^2 + (k+1)InN \quad (15)$$

The procedure is terminated at the (k+1)th step when

$$BIC_k < BIC_{k+1} \quad (16)$$

A subset model containing k hidden units is given. The Bayesian information criterion(BIC) is an important instrument in balancing the accuracy and the complexity of the final network. In many signal processing applications, it makes the whole selection procedure simple and efficient. Furthermore there is no puzzle of the toleranceρ, because it is hard to find but the Bayesian information criteria (BIC) algorithm only need the residual energy and don't want to know the noise to noised-signal ratio.

The next section will discuss the width of regressor vectors *p* since the various value of it gives rise to different relative energy of regressors. Obviously, this offers the advantage of taking the distribution variations of the data into account and guarantees a natural overlap between Gaussian kernels.

IV. The Width of Gaussian Functions

Intuitively, it is assumed that there is a negative correlation between numbers of basis function and the width of Gaussian functions if the data were uniformly distributed in the input space, leading to a uniform distribution of centroids. Unfortunately most real-life problems show non-uniform data distributions. In this section we give a variable width method to solve the problem.

A. local widths

To calculate the local widths, a pilot density estimate is first computed as[9]

$$p(x_i) = \sum_{j \ne i} k((x_i - x_j)/\sigma_0)/(n\sigma_0^d) \quad (17)$$

Where σ_0 is a manually specified global width and d is the dimension of the data space. Based on eq. local widths are calculated as:

$$\sigma(x_i) = \sigma_0[\lambda / p(x_i)]^\alpha \quad (18)$$

Where $p(x_i)$ is the estimated density at point xi, α is the sensitivity parameter, a number satisfying $0 \le \alpha \le 1$ (a suggest value for α is 1/2[10]),λ is a constant which is by default assigned to be geometric mean of $\{p(x_i)\}|i=1…N$. It can be written as

$$\log \lambda = n^{-1} \sum \log(p(x_i)) \quad (19)$$

However, the local width is influenced by the choice of the proportionality constant λ, which changes the relative value of it, if $p(x_i) < \lambda$,$\sigma(x_i)$ increases relative to σ_0 implying more smoothing for the point x_i, for data points that $p(x_i) < \lambda$, the local width becomes narrower. A particular choice of $\sigma(x_i)$ is able to perform much better than fixed-widths methods, as they offer a greater adaptability to the data.

B. global widths

In the subsection we suggest the differential evolution (DE) algorithm for the computation of the Gaussian function global widths based on an exhaustive search; the purpose is to obtain optimization of them. Currently, there are several variants of DE. The particular variant used throughout this investigation is the DE/rand/1 bin scheme. The stable value of BIC is a fitness which is used to measure candidate optimality.

V. Computational Experiments

Two tasks, to explore the effectiveness of our novel OLS algorithm, were considered in the section: the Hermit polynomial and a linear combination of six Gaussian functions. Both are well-known problems recognized as benchmark tasks in the study of different neural network architectures. In this subsection, a set of experiments was carried out in order to evaluate the method under different noise levels and several training sets with distinct characteristics.

To perform the experiments, all the parameters of DE are set as follows: population size, N_p=30; differential amplification factor, F=0.5; crossover probability constant, C_r=0.1; strategy, DE/rand/1/bin (classical version of DE); termination criterion, $|\sigma_0(k+1) - \sigma_0(k)|/\sigma_0(k) < 0.001$(k is the Iterations of DE).

A. Hermit polynomial

In the section, the results for approximating the Hermit Polynomial are presented. The Hermit Polynomial is given by the following equation:

$$F(x) = 1.1(1-x+2x^2)\exp(-x^2/2) \quad (20)$$

A random sampling of the interval [-4 4] is used in obtaining 400 input-output data for the training set. The output set can be modeled adequately as

$$y_k = F(x_k) + fe_k \quad (21)$$

Where f describes the noise amplifying factor, which is a constant, e_k denotes the system noise, and it meet the norm distribution with mean value 0 and covariance 1, the y_k (output set) is a fitting set and the value of $F(x_k)$ is used to form the testing set. The former is used in the selection and the latter is used to validate the selected network. The model test output $\hat{y}_k$ and the test error rate

$$MSE = (\hat{y}_k - F(x_k))^2 / N \quad (22)$$

In order to verify the effectiveness of the novel OLS algorithms for different noise Hermit systems, it can be interesting to compare it with the original OLS (Chen et al.1991). To do so, the noise factor f is uniformly sampled in the interval (0.03, 0.3), and also both the fixed-width OLS algorithm with the tolerance ρ(classical method) and the variable width OLS algorithm with the Bayesian information criteria (proposed method) were applied to this data set. Figure 2,3,4,5 show the results from the two algorithms. With increasing of noise factor (f) the learning MSE of the two methods becomes worse and worse, but comparatively the proposed method still provide slightly lower approximation error for us and the number of hidden units is slightly less under different noise levels. But the global widths searched by DE algorithm are almost identical for them.

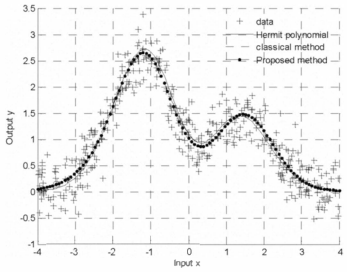

Figure 2. Behavior of the classical and proposed algorithm on a typical execution on the Hermit polynomial.

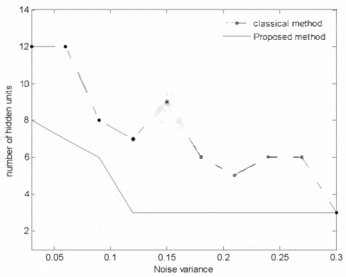

Figure 3. comparison of number(hidden units) between the classical and proposed method for Hermit polynomial under different noise levels.

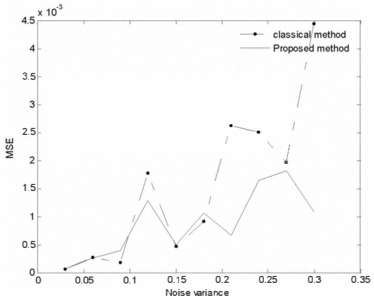

Figure 4. comparison of MSE between the classical and proposed method for Hermit polynomial under different noise levels.

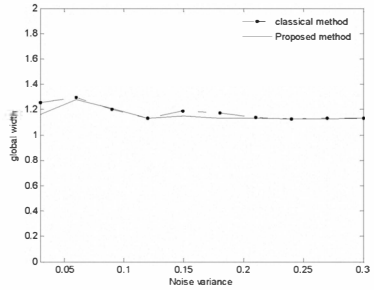

Figure 5. comparison of global widths between the classical and proposed method for Hermit polynomial under different noise levels.

It can be seen from four figures that the two algorithm techniques achieved the same excellent generalization performance but the proposed algorithm obtained a slightly smaller size than the classical method under different noise levels. It should be emphasized that the tolerance ρ is critically important for the classical method but will be difficult to implement in the real system, so it has more learning parameters that require tuning. However, Bayesian information criteria (BIC) is employed to the selection of centers for the proposed method, as it enables the selection procedure to be automatically terminated without the need for the user to specify the tolerance ρ, and thus it is easier to tune and computationally more efficient than the classical algorithm. It is also worth emphasizing that the proposed method adopts the basic idea of variable widths and augment it with a local widths strategy to obtain a minimal RBFNN network. The local widths strategy offers a greater adaptability to the data and guarantees a natural overlap between Gaussian kernels, and thus it needs less number of hidden units and maximize the generalization ability of the network. For all those problems, the proposed approach produced a network with fewer hidden neurons than the classical method with the same or smaller approximation errors.

B. A linear combination of six Gaussian functions

Consider the nonlinear function which results from linear combinations of six Gaussian units.

$$y(x) = \exp[-\frac{(x_1-0.3)^2+(x_2-0.2)^2}{0.01}] + \exp[-\frac{(x_1-0.7)^2+(x_2-0.2)^2}{0.01}]$$
$$+ \exp[-\frac{(x_1-0.1)^2+(x_2-0.5)^2}{0.02}] + \exp[-\frac{(x_1-0.9)^2+(x_2-0.5)^2}{0.02}]$$
$$+ \exp[-\frac{(x_1-0.3)^2+(x_2-0.8)^2}{0.01}] + \exp[-\frac{(x_1-0.7)^2+(x_2-0.8)^2}{0.01}]$$

Theoretically we can say that to approximate this function, the novel OLS algorithm should give rise to only six hidden neurons. To test this, training samples are randomly chosen in the interval (0, 1) and the observations are corrupted by noise with mean 0 and variance 0.3. By one or several iterations of the DE algorithm, the results of the classical and proposed method are obtained. Figure 6 shows the results of applying the two algorithms to this problem. It is obvious that the OLS of fixed-width method finally sets to 8 hidden units instead of the desired 6 hidden units and the novel OLS algorithm detected a correct number of hidden neurons(the desired centers (□) and the centers produced by the classical method (o)). Thus the latter did not overfits and so the network realized by the method is minimal. Moreover, to evaluate the stable performance of the two algorithms, we fix the global widths acquired just and then make the their procedures run ten times in different random data sets but variance still is 0.3. It can be seen from figure 7, 8 that our algorithm achieves the same or a lower approximation error and less number of hidden units, which is always 6, and thus has a good stability.

VI. CONCLUSION

The crucial question of how to select radial basis function centers and widths from the data points has been investigated and a learning strategy based on the novel OLS has been developed for the better construction of RBF neural networks in different noise nonlinear-dynamic systems, the results show that the proposed algorithm learning procedure indeed offers much better performance over the classical orthogonal least squares method.

ACKNOWLEDGMENT

The authors would like to thank the anonymous receives for their helpful suggestions, and also thank Hong Wu for the help in writing.

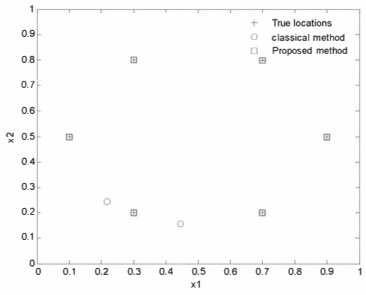

Figure 6. Number and positions of hidden unit centers achieved by two methods for the static function approximation problem.

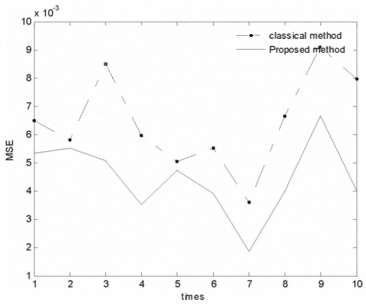

Figure 7. comparison of MSE between the classical and proposed method for A linear combination of six Gaussian functions with noise variance 0.3.

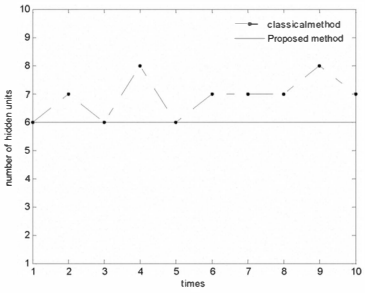

Figure 8. comparison of number (hidden units) between the classical and proposed method for A linear combination of six Gaussian functions with noise variance 0.3.

REFERENCES

[1] C. Hung, Y. Kim and T. Coleman, "A comparative study of radial basis function neural networks and wavelet neural networks in classification of remotely sensed data," 5th Biannual World Automation Congress,2002.

[2] M.J.D. Powell, "Radial basis functions for multivariate interpolation: a review," in Algorithms for the approximation, J. C. Mason and M. G. Cox (Eds.), Oxford University Press, 1987,143-167.

[3] C. Darken and J. Moody, "Fast Adaptive K-Means Clustering: Some Empirical Results," Proceedings of the IEEE IJCNN Conference, IEEE Press, Piscataway, New Jersey.2, 1990,233-238.

[4] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," Int. J. Contr., 50(5),1989,1873-1896.

[5] S. Chen, C.F.N. Cowan, P.M. Grant, Orthogonal least squares learning algorithm for radial basis funciton networks , IEEE Trans. Neural Networks 2(2) ,1991, 302-309.

[6] P.Zhou, Dehua Li, H. Wu, et. al. "A Comparative Study of Radial Basis Function Neural Networks in dynamic clustering algorithm" Proceedings of the SPIE, Vol. 7496, 2009, pp.749612-8.

[7] G.Schwarz, Estimating the dimension of a model, Ann. Stat. 6(2),1978,461-464.

[8] X. f. Wu and Y. h. Lao, "A new algorithm for model structure determination and parameters estimation of nonlinear dynamic systems," ACTA AUTOMATICA SINICA, 18(4), 1992,385-392.

[9] D.Comaniciu,V.Ramesh,and P.Meer,"The Variable Bandwidth Mean Shift and Data-Driven Scale Selection," Proc. Eighth int'l Conf. Computer Vision,vol.1,July,2001,pp.438-445.

[10] B.W.Silverman ,Density Estimation for Statistics and Data Analysis, New York: Chapman and Hall, 1986.