International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:2, 2012

An Optimal Feature Subset Selection for Leaf Analysis

N. Valliammal and S.N. Geethalakshmi

***Abstract*—**This paper describes an optimal approach for feature subset selection to classify the leaves based on Genetic Algorithm (GA) and Kernel Based Principle Component Analysis (KPCA). Due to high complexity in the selection of the optimal features, the classification has become a critical task to analyse the leaf image data. Initially the shape, texture and colour features are extracted from the leaf images. These extracted features are optimized through the separate functioning of GA and KPCA. This approach performs an intersection operation over the subsets obtained from the optimization process. Finally, the most common matching subset is forwarded to train the Support Vector Machine (SVM). Our experimental results successfully prove that the application of GA and KPCA for feature subset selection using SVM as a classifier is computationally effective and improves the accuracy of the classifier.

***Keywords*—**Optimization, Feature extraction, Feature subset, Classification, GA, KPCA, SVM and Computation

I. Introduction

ONE of the upcoming research areas is the need for the development of automatic plant recognition system such as CAP-LR (Computer Aided Plant Leaf Recognition) [11],[12]. Botanists need a computer-aided tool without human interaction to study and identify leaves instead of holding a plant encyclopedia. Huge Volumes of biological information are now providing on-line access to hundreds and thousands of images of specimens, helping to digitize the complete Specimen Collection of the leaf images [3]. Such a system will return within seconds the top matching species, along with supporting data that describes about textual descriptions and high resolution type specimen images just by feeding into the computer the photograph of a leaf specimen [5]. By using our system, a botanist in this field can quickly search the entire collections of plant species within seconds which earlier took hours together. Plant species identification is a process in which each individual plant should be correctly assigned to descending series of groups of related plants, as based on common characteristics. Currently, plant taxonomy methods still adopt traditional classification method such as morphologic anatomy, cell biology and molecular biological approaches. The traditional method is carried out by botanists which tends to be time consuming and less efficient. Furthermore, it is a troublesome task.

However, due to the rapid development in computer technologies nowadays, there are a new opportunities to improve the ability of plant species identification such as designing a convenient and automatic recognition and classification system of plants.

Assistant Professor Authors are with Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641 043, valli.p.2008@gmail.com, sngeethalakshmi@yahoo.com

Feature selection is one of the major tasks in classification problems. Most of these features are either partially or completely irrelevant or redundant to the classified target [8]. In advance, to discriminate among the classes, these features will not provide sufficient information. It is also infeasible to include all possible features in the processes of classifying the patterns and objects [1],[2],[3]. The discriminate features have to be carefully extracted from the image and the extracted features are used to train the classifiers. The optimal features subset is selected to increase the matching accuracy based on the performance of the classifiers [4],[13],[15]. Reducing the dimensions of the feature space not only reduces the computational complexity, but also increases estimated performance of the classifiers. An optimal approach with a combination of GA and KPCA are used to select the feature subset and is forwarded to train the SVM classifier [9],[14] to successfully classify the plant and tree leaves. However, performing these two steps separately might result in a loss of information relevant to classification tasks. Recently, several approaches [8], [9], [10], [11], [12] for joint feature selection and SVM construction have been proposed. Extending previous work on feature selection and classification, this paper proposes a convex framework for jointly learning optimal features using GA and KPCA and further classifying using SVM. Our approach proves computationally effective by improving the classification performance. The experimental results prove the effectiveness and superiority of this method.

The rest of the paper is organized as follows; Section 2 describes the Overview of the proposed optimal approach. Section 3 discusses about the various features considered for reduction. Section 4 explains in detail about the subset selection through GA and KPCA. Section 5 explains the classification accuracy through SVM. Section 6 shows the Experimental results and comparisons and finally, the conclusion is summarized in section 7 with references.

II. Problem in Optimal Feature Subset Selection

Plant leaves are approximately two-dimensional in nature and the shape of plant leaf is one of the most important features for characterizing various plants species. Therefore, it is necessary for us to develop an easy and automatic method that can correctly discriminate and recognize leaf shapes of different species. Some research work has been done on this problem. Two basic approaches to shape analysis exist: region based and boundary-based (contour-based). Region based systems typically use moment descriptors9 that include geometrical moments, Zernike moments and Legendre moments8. Boundary-based systems use the contour of the objects and usually give better results for images that are

distinguishable by their contours. According to plant recognition, the second most vital interpretation element is color. By combining texture with the shape feature discrimination capability of the method can be improved. Most of the existing plant recognition methods are based on both the global shape feature for leaves. The goal of this paper is to illustrate the value of feature selection in combining features from different data models, and to demonstrate the potential difficulties of performing feature selection in small sample size situations, due to the curse of dimensionality. The effectiveness of optimal feature subset selection techniques are implemented.

III. Overview and Framework of The Optimal Approach

The selection of the optimal features subset and the classification has become an important methodology in the field of Leaf classification. There are many approaches available for optimal feature subset selection and leaf classification such as Swarm Optimization, Ant Colony Optimization, Bee optimization, GA, KPCA, Hidden Markov Models (HMM), Bayesian networks, SVM and Dynamic Time Warping etc. Among these approaches SVM classifier has proven to be a powerful tool for solving problems of prediction, classification and pattern recognition. Such systems can achieve greater accuracy than HMM based systems and can handle low quality and noisy data. Our optimal approach is proved efficiently through experimental results. Figure 1 explains the framework for the classification system.

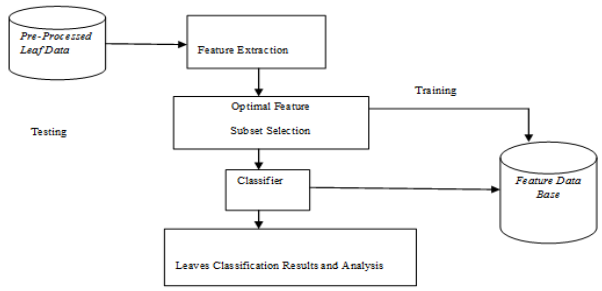

Fig. 1 Frame work for Optimal Classification system

A. Feature Extraction

The main work of a leaf classification system is to extract common features among the images belonging to the same class of the data set and consequently indexing them. This method is applied to capture visual content of images for indexing and retrieval. The great variability of shapes and size of the leaves for plants and trees makes the recognition task difficult. Feature extraction methodologies analyse leaf images to extract the most prominent features that are representative of the various classes of objects [6], [7]. The below stated features are considered for extraction.

1. Shape Features

The shape features used for identification of leaves are given in the table I,

TABLE 1
Leaf Shape Features

S.No.	Feature Name	Feature Calculation
1.	Rectangularity	$R = A_{Leaf} / A_{Bounbox}$
2.	Circularity	$C = \mu_r / \delta_r$
3.	Sphericity	$S = r_i / r_c$
4.	Eccentricity	$E = E_A / E_B$
5.	Axis ratio	$A = D_{\min} / D_{\max}$
6.	Diameter	$D = sqrt(4 * area / pi)$
7.	Complexity	$CM = \sqrt{perimeter} / area$
8.	Perimeter	$P = 2l + 2w$

2. Texture Features

The texture features considered for Extraction are shown in the following table II,

TABLE II
Leaf Texture Features

S.No.	Feature Name	Feature Calculation
1.	Energy	$Energy = \sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2$
2.	Entropy	$Entropy = \sum_{i,j=0}^{N-1} \left(P_{ij}\right) P_{ij}$
3.	Correlation	$correlation = \sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2}$
4.	Contrast	$Contrast = \sum_{i,j=0}^{N-1} P_{ij} \left(i-j\right)^2$
5.	Homogeneity	$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+\left(i-j\right)^2}$
6.	Sum of Squares	$SOS = \sum_i \sum_j \left(i-\mu\right)^2 P(i,j)$
7.	Inverse Different Moment	$IDM = \sum_i \sum_j \frac{1}{1+\left(i-j\right)^2} P(i,j)$
8.	Angular Second Moment	$ASM = \sum_i \sum_j \left(P_{ij}\right)^2$

3. Color Features

The color features used for Extraction are shown in the following table III,

TABLE III
LEAF COLOR FEATURES

S.No.	Feature Name	Feature Calculation
1.	Mean	$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} P_{ij}$
2.	STD	$\sigma = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{ij}-\mu\right)^2}$
3.	Skewness	$\theta = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{ij}-\mu\right)^3}{MN\sigma^3}$
4.	Kurtosis	$\gamma = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left(P_{ij}-\mu\right)^4}{MN\sigma^4}$

IV. Optimal Feature Subset Selection

Feature selection is an important task that allows the determination of the most relevant features for pattern recognition. The extracted features are normalized or reduced by selecting appropriate features to improve the classification accuracy [15]. A good feature selection falls inside the following specifications.

- Faster training and better generalization.
- Removes redundant leaf images.
- Focuses recognition to a small set of properties.
- Displays the final classified outcome.

In this paper, we describe about the selection of optimal set of features using GA and KPCA, which provide the discriminating information to classify the leaf patterns and increases the matching accuracy. In our approach, Feature subset selection is performed in two stages.

A. Feature selection based on GA

For the first stage the parameters of all the shape, texture and color features are optimized through GA. The normal process for searching the features is computationally expensive and therefore GA is used as a search algorithm [9]. GAs have been successfully applied to feature selection in an effort to reduce the number of features needed while improving the classification accuracy. Compared with other optimization methods, GA has the following advantages: (i) encoding feature: GA takes certain genetic encoding as the object of the operation leading to solve all kinds of complicated optimization problem uniformly (ii) strong robustness: GA finds the optimal way of population search so that it can search the whole solution space more effectively (iv) performs probabilistic search where every genetic operation is executed on the stochastic case resulting in the improved ability to skip the local optimal. With these advantages, GA is a popular algorithm for complicated and difficult optimization problems. There are four major preparatory steps needed in solving a problem with GA (i) initialization, determine the population size (ii) the fitness measure, which is the foundation of evaluating and selecting the individuals (iii) three genetic operators, which are the search mechanisms to form the new population from generation to generation and (iv) some evolutionary parameters, such as the mutation rate , the crossover rate (mutation and crossover are done with certain probabilities) and the maximum of iteration.

B. Evaluation Function

Selecting an appropriate evaluation function which produces the fitness of each individual in the population is essential. GAs then use this feed-back to bias the search process so as to provide an improvement in the population's average fitness. The natural representation for the feature selection problem is the binary string of length N to indicate the presence or absence of each of the N possible features. The following figure 2 shows the framework for optimal approach feature subset selection using GA and KPCA.

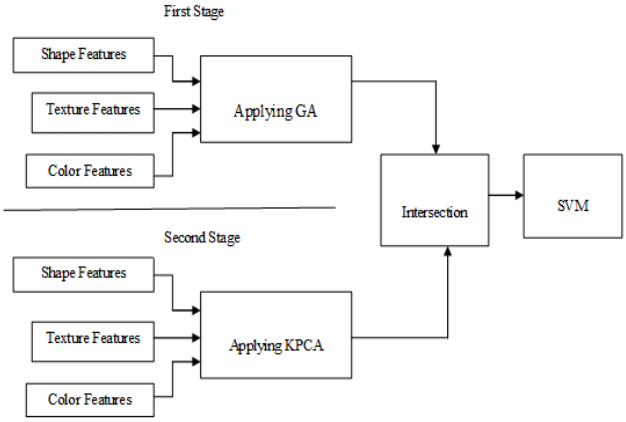

Fig. 2 Proposed Optimal Subset Selection for Leaf Analysis

The evaluation function selects the appropriate feature set which is solely based on the performance of the classification process. In our system, the specified twenty features in the first phase are reduced to seven features. These selected features are reduced through the optimization of GA and forwarded to the intersection operator path which is finally routed to SVM. The following table IV gives the reduced subset value based on GA.

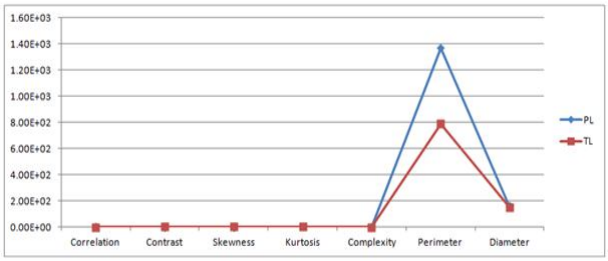

Fig. 3 Subset Selection using GA for Leaf Analysis

TABLE IV
SUBSET SELECTION USING GA FOR PLANT AND TREE LEAF

S.No.	Feature Name	PL	TL
1	Correlation	9.41E-01	1.33E-02
2	Contrast	2.68E-02	9.67E-01
3	Skewness	6.52E-01	9.55E-01
4	Kurtosis	1.43E+00	1.91E+00
5	Rectangularity	20616	18072
6	Perimeter	1.37E+03	7.91E+02
7	Diameter	1.62E+02	1.52E+02

C. Feature selection by KPCA

KPCA is an independent nonlinear feature selection method, which performs the mapping into the feature space F with kernel functions and uses a linear analysis algorithm to discover patterns in the nonlinear kernel-defined space. Kernel PCA is a non-linear extension of the PCA in a kernel-defined feature space making use of the dual representation.[9]

There are several methods for computing the principal components of a symmetric matrix. The choice depends on the properties of the matrix and on how many components one is seeking. The features optimized through KPCA are shown in table 5 and correspondingly in Figure 4. By applying KPCA, 20 features are optimized to a reduction of 11 features.

TABLE V
SUBSET SELECTION USING KPCA FOR PLANT AND TREE LEAF

S.No.	Feature Name	PL	TL
1	Correlation	9.41E-01	9.67E-01
2	Contrast	2.68E-02	9.93E-01
3	Homogenity	9.87E-01	1.28E+00
4	Sum of Squares	1.43E+00	1.28E+00
5	Mean	1.34E+00	2.08
6	Std	1.76	9.55E-01
7	Skewness	6.52E-01	1.91E+00
8	Kurtosis	1.43E+00	18072
9	Rectangularity	20616	7.91E+02
10	Perimeter	1.37E+03	1.52E+02
11	Diameter	1.62E+02	1.33E-02

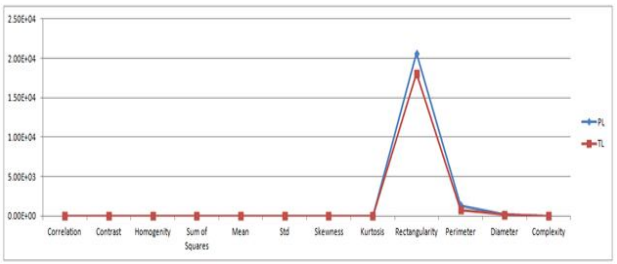

Fig. 4 Subset Selection using KPCA for Leaf Analysis

The optimized feature subset selection using combination of GA and KPCA after intersection are shown in table V and correspondingly in figure 5.

TABLE V
SUBSET SELECTION USING OPTIMAL APPROACH FOR PLANT AND TREE LEAF

S.No.	Feature Name	PL	TL
1	Correlation	9.41E-01	1.33E-02
2	Contrast	2.68E-02	9.67E-01
3	Skewness	6.52E-01	9.55E-01
4	Kurtosis	1.43E+00	1.91E+00
5	Rectangularity	20616	18072
6	Perimeter	1.37E+03	7.91E+02
7	Diameter	1.62E+02	1.52E+02

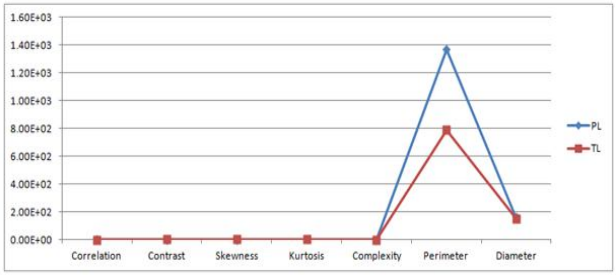

Fig. 5 Optimal approach for Leaf Analysis

V. CLASSIFICATION USING SVM

Support Vector Machines are gaining popularity because of their promising performance in classification and prediction. The success of SVM lies in suitable kernel design and selection of its parameters. SVM is theoretically well-defined and exhibits good generalization result for many real world problems. SVM is extended from binary classification to multiclass classification since many real-life datasets involve multiclass data. A special property of SVM is it simultaneously minimizes the empirical classification error and maximizes the geometric margin. So SVM is called as Maximum Margin Classifiers. The parameters of the SVM are tuned to improve the overall generalization performance [8],[11].

VI. EXPERIMENTAL RESULTS AND COMPARISON

In our work the shape features, texture features and color features are extracted from different plant and tree leaves. These features are separately optimized using GA and KPCA. To these optimized subsets intersection operation is performed and the results of which is forwarded to train SVM. Fifty image data set is taken and 20 features are extracted and it is reduced to 7 features for classification. 70% of images are used for training and 30% of images are used for testing purpose. Some sample dataset considered for classification is shown in the following figure 6.

Fig. 6 Sample set taken for Leaf Analysis

TABLE VI
CONFUSION MATRIX FOR SVM CLASSIFICATION

		Predicted Value	
		Plant Leaf	Tree Leaf
Actual	Plant Leaf	27	03
	Tree Leaf	6	19

The above table VI shows the confusion matrix for a binary classification using SVM classification. The target values are either Plant leaf or Tree leaf. The following can be computed from this confusion matrix.

1. The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} = 0.92$$

2. The *recall* or *true positive rat*e (*TP*) is the proportion of positive cases that were correctly identified, is calculated using the equation:

$$TP = \frac{d}{c+d} = 0.76$$

3. The *false positive rate* (*FP*) is the proportion of negative results that were incorrectly classified as positive, is calculated using the equation:

$$FP = \frac{b}{a+b} = 0.12$$

4. The *true negative rate* (*TN*) is defined as the proportion of negative results that were classified correctly, is calculated using the equation:

$$TN = \frac{a}{a+b} = 1.08$$

5. The *false negative rate* (*FN*) is the proportion of positives results that were incorrectly classified as negative, is calculated using the equation:

$$FN = \frac{c}{c+d} = 0.24$$

6. Finally, *precision* (*P*) is the proportion of the predicted positive results that were correct, is calculated using the equation:

$$P = \frac{d}{b+d} = 0.86$$

The accuracy of the Classifier is 92 % and is proved efficiently.

VII. CONCLUSION AND FUTURE WORK

This paper described about an optimal approach of subset selection for classification of leaves based on GA and KPCA.

The deterministic shape, texture and colour features are extracted from the leaf images. The optimal deterministic subset is selected by separately applying GA and KPCA technique. Following which the intersection operation is performed to efficiently select the best common matching subset. These subsets are finally forwarded to train the SVM. Our experimental results indicate that the application of GA and KPCA for feature subset selection using SVM as a classifier proves computationally effective and improves the accuracy of the classifier.

REFERENCES

[1] A.Kadir,L.E. Nugroho, A. Susanto and P.I. Santosa, A Comparative Experiment of Several Shape Methods in Recognizing Plants, International Journal of Computer Science and Information Technology (IJCSIT), Vol 3, No 3, P.256-263, June 2011.
[2] Abdul Kadir, Lukito Edi Nugroho, Adhi Susanto, Paulus Insap Santosa, Leaf Classification Using Shape, Color, and Texture Features, International Journal of Computer Trends and Technology, P.224-230,2011.
[3] Jyotismita Chaki, Ranjan Parekh, Plant Leaf Recognition using Shape based Features and Neural Network classifiers, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 10, P. 41-47, 2011.
[4] Wahyu Wibowo, Hugh E. Williams, Simple and Accurate Feature Selection for Hierarchical Categorisation, ACM Digital library, 2002.
[5] P. Tzionas, S.E. Papadakis, and D. Manolakis, "Plant leaves classification based on morphological features and a fuzzy surface selection technique", in Proceeding of International Conference on Technology and Automation, Thessaloniki, Greece, P. 365-370, 2005.
[6] Xiaodong Zheng, Xiaojie Wang, Leaf Vein Extraction Based on Gray-scale Morphology, I.J. Image, Graphics and Signal Processing, Vol.2, 2P.25-31,2010.
[7] N. Kumar, S. Pandey, A. Bhattacharya, and P. S. Ahuja, "Do leaf surface characteristics affect agrobacterium infection in tea *J. Biosci.*, vol. 29, no. 3, P. 309–317, 2004.
[8] G. Guo, S. Li, and K. Chan, "Support vector machines for face recognition," Image and Vision Computing, vol. 19, no. 9, P. 631–638, 2001.
[9] S. Papadakis, P. Tzionas, V. Kaburlazos, and J. Theocharis, "A genetic based approach to the Type I structure identification problem," Informatica, vol. 5, no. 3, 2005.
[10] Yan Li, Zheru Chi, and David D. Feng, "Leaf Vein Extraction Using Independent Component Analysis," 2006 IEEE Conference on Systems, Man and Cybernetics, Vol. 5, Taipei, P. 3890-3894,2006.

[11] Chomtip Pornpanomchai, Chawin Kuakiatngam Pitchayuk Supapattranon, and Nititat Siriwisesokul,, Leaf and Flower Recognition System (e-Botanist), International Journal of Engineering and Technology (IACSIT), Vol.3, No.4, ,P.347-351, 2011.
[12] B.Sathya Bama et.al., Content Based Leaf Image Retrieval (CBLIR) Using Shape, Color and Texture Features**,** Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2 ,No. 2 ,P. 202-211,2011.
[13] Maliheh Shabanzade, Morteza Zahedi and Seyyed Amin Aghvami, Combination of Local Descriptors and Global Features for Leaf Recognition, Signal and Image Processing : An International Journal (SIPIJ) Vol.2, No.3, P. 23-31,2011.
[14] R. Sinan Tumen1, M. Emre Acer2 and T. Metin Sezgin1, Feature Extraction and Classifier Combination for Image-based Sketch Recognition, EUROGRAPHICS Symposium on Sketch-Based Interfaces and Modeling , P.1-8,2010.
[15] Chomtip Pornpanomchai, Supolgaj Rimdusit, Piyawan Tanasap and Chutpong Chaiyod, Thai Herb Leaf Image Recognition System (THLIRS), Kasetsart J. (Nat. Sci.) , Vol.45, P. 551 - 562 ,2011.
[16] Krzyszt Michalak, Halina Kwasnicka, Correlation–Based Feature Selection Strategy in Classification Problems, Int. J. Appl. Math. Comput. Sci., Vol. 16, No. 4, P.503–511, 2006.
[17] Qisong Chen, Xiaowei Chen and Yun Wu, Optimization Algorithm with Kernel PCA to Support Vector Machines for Time Series Prediction, Journal of Computers, Vol. 5, NO. 3, P.380-387, 2010.
[18] Shanwen Zhang and Kwok-Wing Chau, Dimension Reduction Using Semi-Supervised Locally Linear Embedding for Plant Leaf Classification, ICIC 2009, LNCS 5754, P. 948–955, 2009.
[19] Debdoot Sheet and Jyotirmoy Chatterjee, Hrushikesh Garud, Feature Usability Index and Optimal Feature Subset Selection, International Journal of Computer Applications, Vol.12, No.2, P.29-37, 2010.
[20] D S Guru, Y. H. Sharath, S. Manjunath, Texture Features and KNN in Classification of Flower Images, IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition", P.21-29, 2010.
[21] Minh Hoai Nguyen, Fernando De la Torre, Optimal Feature Selection for Support Vector Machines, Pattern Recoginition, P. 1-25, 2009.
[22] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, Qi Tian, Feature Selection Using Principal Feature Analysis, ACM Multimedia, September 23-29, 2007.
[23] Amaro Lima, Heiga Zen, Yoshihiko, Keiichi Tokuda,Tadashi Kitamura, Members, and Fernando G. Resende, Applying Sparse KPCA for Feature Extraction in Speech Recognition, IEICE TRANS. INF. & SYST., Vol.E88–D, No.3, P. 401-402, 2010.