Computing and Informatics, Vol. 40, 2021, 298–317, doi: 10.31577/cai_2021_2_298

LOGISTIC REGRESSION BASED ON STATISTICAL LEARNING MODEL WITH LINEARIZED KERNEL FOR CLASSIFICATION

Xiaochun Guan, Jianhua Zhang, Shengyong Chen

School of Computer Science and Technology
Zhejiang University of Technology
Liuhe Road No. 288
310 023 Hangzhou, China
e-mail: guanxc@wzu.edu.cn, zjh@zjut.edu.cn, csy@tjut.edu.cn

Abstract. In this paper, we propose a logistic regression classification method based on the integration of a statistical learning model with linearized kernel pre-processing. The single Gaussian kernel and fusion of Gaussian and cosine kernels are adopted for linearized kernel pre-processing respectively. The adopted statistical learning models are the generalized linear model and the generalized additive model. Using a generalized linear model, the elastic net regularization is adopted to explore the grouping effect of the linearized kernel feature space. Using a generalized additive model, an overlap group-lasso penalty is used to fit the sparse generalized additive functions within the linearized kernel feature space. Experiment results on the Extended Yale-B face database and AR face database demonstrate the effectiveness of the proposed method. The improved solution is also efficiently obtained using our method on the classification of spectra data.

Keywords: Elastic net, generalized additive model, kernel, lasso regression, spectra data

Mathematics Subject Classification 2010: 68U10

1 INTRODUCTION

In statistical data modeling, regression is a popular method to explore low-dimensional structures in Statistics and Computation. It uses the samples to estimate

the parameters in the formula [1, 2]. It is a widely used statistical analysis method for data modeling. There are two standards to evaluate the regression. One is the prediction accuracy, the other is a better interpretation. Prediction accuracy means the model's prediction accuracy on future testing data. The interpretation of the model refers to a more parsimonious model. Parsimony plays an important role in inference. The ordinary least squares (OLS) regression is obtained by minimizing the residual squared error, ridge regression plus the square sum of the regression coefficients as a penalty function on the residual squared error. The OLS regression tends to obtain a lower prediction accuracy compared with ridge regression. Ridge regression shrinks coefficients continually and hence is more practical and reliable. Its prediction accuracy is better than OLS regression, but it does not set any coefficients to 0 and hence it does not improve the model's interpretation. Subset selection can provide interpretable models because regressors are either retained or dropped from the model by subset selection, but its prediction accuracy tends to be very unstable because of its inherent discreteness. An influential regularization technique called least absolute shrinkage and selection operator (lasso) was proposed by Tibshirani [3]. Lasso is a penalized least squares method that imposes an L1 penalty on the regression coefficients. Owing to the sparse nature of the L1 penalty, the lasso can compress some coefficients and simultaneously set some coefficients to zero, thus it can produce a sparse representation of the model. It also has been proved that the L1 penalty can discover the "right" sparse representation of the model under certain conditions [4, 5, 6]. The success of the lasso is accomplished by the L1 penalty applied to the coefficients. This L1 penalty approach is also called basis pursuit in the field of signal processing [7]. In 2004, Efron et al. proposed the least angle regression algorithm (LARS) to solve the entire lasso solution path efficiently [8]. The LARS makes lasso widely used in feature selection and parameter estimation. Lasso also has been supported by much theoretical work in sparse representation and compressed sensing. Especially since 2006, Donoho and Tao et al. have put forward a theoretical basis for compressed sensing, which successfully constructed theory and practical methods in the field [4, 9, 10, 11, 12, 13].

However, for high dimension and small sample data, such as the gene selection problem in microarray data analysis, the lasso can not select the grouping information in situations consisting of grouped variables [14]. Zou et al. proposed a new feature selection algorithm called elastic net. The elastic net can not only simultaneously do automatic variable selection and continuous shrinkage, but also select groups of closely correlated variables, i.e. either group selection or omission of the correlated variables. Also in 2015, Chouldechova and Hastie introduced an extension of the lasso to the additive model setting, the method is called Generalized Additive Model Selection (GAMSEL), this method can select among zero, linear and non-linear fits as component functions in a generalized additive model framework by an overlap group-lasso penalty [15]. It also incorporates a penalized likelihood procedure for fitting sparse generalized additive models.

In this paper, we consider the classification as a multinomial/binomial logistic regression problem. There is research on the regression coefficients matrix of multinomial regression [16]. Due to the important and remarkable applications of face recognition technology, there have been many successful algorithms for face recognition, such as sparse representation classification, linear regression, elastic net et al. [11, 17, 18, 19, 20, 21]. In recent years, significant progress has been made on face recognition systems [22, 23]. Especially in [24], a semi-supervised sparse representation based classification method is proposed to address the problem of face recognition when the labeled samples are insufficient. Face recognition can be considered as a multi-class classification problem, it can also be regarded as a multinomial logistic regression problem. We use the elastic net's grouping effect to find the grouping features in the linearized kernel (LK) feature space of the samples based on the statistical learning model-the generalized linear model. In the model, an elastic net penalized negative log-likelihood function method was adopted to perform variables selection for classification. The elastic net regularization can also properly adapt to the situation where the number of samples is much smaller than the predicted variables. Thus the algorithm can perform well for some face databases which do not contain enough samples in each sample space. Simulation and experiments on publicly available face data and Raman spectra data are used to demonstrate the feasibility of our proposed method. The experiment results show that our method improved the classification accuracy by up to 0.83 % and 3.7 % on the Extended Yale-B face database and AR face database respectively compared with the best result in [25]. We also show the classification results of the GAMSEL model with or without LK pre-processing on spectra data, our method with LK pre-processing can improve the performance by 10 %. We apply the GAMSEL on a subset of Raman spectra data with or without LK pre-processing for the binomial logistic regression problem, the GAMSEL can fit the nonlinear functions within the linearized kernel feature space on the subset of Raman spectra data. It shows that the binomial classification accuracy of the subset of Raman spectra data can be improved with LK pre-processing based on GAMSEL.

The main contributions of this paper can be summarized in two aspects. First, it proposes a novel method that integrates linearized kernel pre-processing into a statistical learning model for multiclass classification. It provides us a perspective to explore the low-dimensional space embedded in the high dimension data. Second, it adopts the fusion of Gaussian and cosine kernels for linearized kernel pre-processing with improved accuracy compared with a single Gaussian kernel. The rest of this paper is organized as follows: Section 2 briefly introduces sparsity and statistical learning. Section 3 depicts information on the kernel and linearized kernel pre-processing. Section 4 describes the logistic regression classification method combining the statistical learning model with linearized kernel pre-processing. Section 5 elaborates extensive experiments. Section 6 includes the analysis and conclusion remarks.

2 SPARSITY AND STATISTICAL LEARNING

Research of sparse representation had started in the 1990s [26], it has been flourishing since the beginning of this century. Sparse representation of signal has attracted many concerns from researchers, for example, the typical image compression algorithm JPEG utilizes image's sparsity in the DCT domain to achieve image compression. The core model in the sparse domain is the linear equations to describe an underdetermined system that has infinitely many solutions. The sparsest solution, which has the least nonzero terms, is the most interesting. The L0 norm can find a sparse solution, and the L0 norm represents the total number of non-zero elements in a vector $\mathbf{x}$. The L0 norm can be defined as in Equation (1), x_i is the elements of the vector $\mathbf{x}$.

$$\|\mathbf{x}\|_0 = \#\left(i \mid x_i \neq 0\right). \tag{1}$$

However, the optimization problem of the L0 norm is an NP-hard problem. It is proved theoretically that the L1 norm is the optimal convex approximation of the L0 norm, so the L1 norm is usually used instead of the L0 norm. The L1 norm represents the sum of the absolute values of each element in a vector. The L1 norm can be defined as in Equation (2).

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|. \tag{2}$$

The solution of the L1 norm is usually sparse and tends to select a very small number of very large values or a small number of insignificant values. L1 norm regularization adds the L1 norm to the cost function, which makes the learning result satisfy the sparsity, so the main features can be extracted. L1 norm has become a popular tool in many research fields [27]. Lasso is one typical example of L1 norm regularization. Given the predictors, $x_1, \ldots, x_p$, the usual linear regression model with response y can be predicted by Equation (3).

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \ldots + x_p\hat{\beta}_p. \tag{3}$$

The vector of coefficients $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ should be fitted by the model. We can assume without loss of generality that the mean of y is 0 and hence omit β_0. Lasso regularization can be defined as in Equation (4).

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} x_j\beta_j \right\|^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq t \tag{4}$$

where t is a nonnegative tuning parameter. It can control the amount of shrinkage that is applied to the estimates. If t is sufficiently small, the lasso can cause continuous shrinkage of the coefficients to 0 as t decreases, and some coefficients can be exactly shrunk to zero. The bias and variance trade-off introduced by the L1 norm

penalty can lead to the coefficients continuous shrinkage and variable selection and thus improve the prediction accuracy [28]. Lasso can produce a parsimonious model. However, lasso can not also reveal the grouping information in situations consisting of grouped variables. To overcome its limitations, Zou et al. proposed a new feature selection algorithm called elastic net [14]. Elastic net combines the L1 norm and L2 norm together as the penalty function on the regression coefficients. The elastic net estimates are defined as in Equation (5).

$$\hat{\boldsymbol{\beta}}_{\text{Enet}} = \arg\min_{\boldsymbol{\beta}} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 \quad \text{s.t. } (1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2 \leq t. \tag{5}$$

$\|\boldsymbol{\beta}\|_1$and $\|\boldsymbol{\beta}\|_2$ represent the vector β's L1 norm and L2 norm, respectively. L1 norm refers to the sum of absolute values of each element in a vector. L2 norm is the square root of the sum of squares of each element in a vector. Similar to lasso, the algorithm LARS-EN is proposed to compute the entire elastic net regularization paths efficiently [14], just like algorithm LARS for the lasso. However, both for the lasso and elastic net, it retains the linear fit for all the predictors, Chouldechova and Hastie introduced an extension of the lasso to the additive model setting in 2015. The method is called Generalized Additive Model Selection (GAMSEL), this method selects between zero, linear and non-linear fits for predictor functions in a generalized additive model framework by an overlap group-lasso penalty. It incorporates a penalized likelihood procedure for fitting sparse generalized additive models [15]. For data $(\mathbf{x}, y)$, a simple linear fit is of the form in Equation (6).

$$\eta(\mathbf{x}) = \sum_{j=1}^{p} \beta_j x_j. \tag{6}$$

For more generative, the generalized additive model was defined as in Equation (7).

$$\eta(\mathbf{x}) = \sum_{j=1}^{p} f_j\left(x_j\right) \tag{7}$$

where the f_j are unknown functions that should be estimated, which can be zero, linear or nonlinear. The GAMSEL is to optimize a penalized negative log-likelihood criterion of the form defined in Equation (8).

$$\hat{f}_1, \ldots, \hat{f}_p = \arg\min_{f_1, \ldots f_p \in \mathcal{F}} l\left(y; f_1, \ldots, f_p\right) + \sum_{j=1}^{p} J\left(f_j\right). \tag{8}$$

It can fit each f_j as zero, linear or nonlinear, as determined by the data. It can capture non-linear relationships among the data. In this paper, we also employ it to explore the nonlinearity of the linearized kernel feature space. The last term in Equation (8) represents the sum of the penalty term of each component f_j. For more detail, please refer to [15].

3 KERNEL AND LINEARIZED KERNEL PRE-PROCESSING

In recent years, with the development of machine learning, a series of kernel function learning methods have been developed. The kernel method is a powerful nonparametric modeling tool. In some cases, it can make problems such as classification and regression easier to solve. It is the Reproducing Kernel Hilbert Space (RKHS) underlying the kernel method that provides linearity, convexity, and general approximation capability, the research of RKHS technology began in the 1940s. The theory of kernel function can be traced back to 1909. The main idea of the kernel method is to transform low-dimensional linear inseparable data into high-dimensional linear separable data. It transforms low-dimensional data into a high-dimensional feature space by a kernel function, which can be equivalent to the inner product of corresponding high-dimensional feature vectors. Then the high-dimensional feature data can be processed by the appropriate linear method as long as the algorithm of that linear method can be expressed by the inner product of the high-dimensional feature vectors of the samples. It is not necessary to know what is the specific high-dimensional features. This is called the "kernel trick". In 1992, Vapnik et al. successfully used this technology to extend linear support vector machine (SVM) to nonlinear SVM [29], its potential was fully realized by researchers. The popular kernel functions mainly include the Gaussian kernel, the polynomial kernel, and the sigmoid kernel. In [30], the author proposed a fusion kernel that fuses the Euclidean and cosine distance measures. The fusion kernel can also be applied to our problem achieving better performance.

Recently kernel method has been widely used in the machine learning field. The kernel method has played an important role in system identification, machine learning, and function estimation [31]. Kernel method has been integrated with sparse representation-based classifier (KSRC) for face recognition with good representation and classification performance [32]. Recently, a new video semantic analysis method with kernel discriminative sparse representation was adopted to efficiently detect the event and concept in video surveillance [33]. Kernel-based machine learning method has been also applied for Chinese license plate recognition [34]. Kernel method is a common way of extending a specific algorithm to deal with a higher dimension "feature space", it has been also incorporated into dictionary learning (DL) [25, 35, 36, 37, 38, 39, 40, 41, 42]. One typical application is its incorporation with dictionary learning in sparse land. There are many successful image processing applications based on DL [43, 44, 45, 46, 47]. Popular algorithms for DL are Method of Optimal Directions (MOD) [48] and other algorithms based on K-means clustering via singular value decomposition (K-SVD), such as label consistent K-SVD1 (LC-KSVD1), label consistent K-SVD2 (LC-KSVD2) and the kernel K-SVD algorithm (KKSVD) [36, 49, 50]. And in [25], Golts et al. give out a new method of incorporating linearized kernel pre-processing into these dictionary learning algorithms, termed "Linearized Kernel Dictionary Learning" (LKDL), which typically gets a relatively good experiment result compared with LC-KSVD1 and LC-KSVD2.

In this paper, we explore the linearized kernel feature space using a statistical learning model. Our method is inspired by the kernel DL method termed "Linearized Kernel Dictionary Learning" (LKDL) by Golts and Elad [25]. They introduce a pre-processing stage based on the kernel method for the dictionary learning algorithm. The idea of Linearized Kernel (LK) pre-processing using the Nyström method to obtain a good approximation of the regular kernel matrix. Please refer to [25] for the details of LKDL. The LK pre-processing method can address the problems of high computational cost and the large storage space for a very large kernel matrix when the kernel trick is used. Without too much effort the LK pre-processing can be incorporated into the algorithms as a kernel layer application. This paper proposed a logistic regression classification method based on the fusion of statistical learning models and linearized kernel pre-processing. The adopted two statistical learning models are the generalized linear model and the generalized additive model. With the generalized linear model, the elastic net regularization is adopted to explore the grouping effect of the linearized kernel feature space. With the generalized additive model, an overlap group-lasso penalty is used to fit the sparse generalized additive functions within the linearized kernel feature space. It can explore the nonlinearity of the linearized kernel feature space.

4 LOGISTIC REGRESSION AND STATISTICAL LEARNING MODEL

Logistic regression is a widely-used method for classification. The logistic regression method is mainly applied to the study of the occurrence probability of certain events. When there are more than two possible outcomes in a problem, multinomial logistic regression can be adopted. For logistic regression, when facing a regression or classification problem, firstly it establishes a cost function, and then iteratively solves the optimal model parameters through a specific optimization method on a training set, and then to verify the quality of the logistic regression model on the testing set. In this study, we adopt two statistical learning models to do logistic regression. They are the generalized linear model and the generalized additive model. Supposing the response variable has K classes $G = (1, 2, \ldots, K)$, for the multinomial logistic regression model, the model can be defined as follows:

$$\Pr(G = k \mid X = \mathbf{x}) = \frac{e^{\beta_{0k} + \beta_k^T \mathbf{x}}}{\sum_{l=1}^{K} e^{\beta_{0l} + \beta_l^T \mathbf{x}}}. \tag{9}$$

For the multinomial logistic regression model, we adopt the Glmnet R package. The Glmnet can fit the generalized linear model via penalized maximum likelihood. Its regularization path can be computed for the elastic net penalty at different regularization parameter lambda. The Glmnet's elastic-net penalized negative log-likelihood function is defined as Equation (10), which can realize the grouping effect

of variables [51].

$$l\left(\{\beta_{0k},\beta_k\}_1^K\right) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}\left(\beta_{0k}+\mathbf{x}_i^T\beta_k\right)-\log\left(\sum_{k=1}^{K} e^{\beta_{0k}+x_i^T\beta_k}\right)\right)\right] + \lambda\left[(1-\alpha)\|\beta\|_F^2/2+\alpha\sum_{j=1}^{p}\|\beta_j\|_q\right]. \tag{10}$$

Here Y to be the $N\times K$ indicator response matrix, with elements $y_{il}=\mathrm{I}(\mathrm{g_i}=1)$, I() is the indication function. β is a $p\times K$ matrix of coefficients. β_k refers to the k^{th} column of outcome class k, and β_j refers to the j^{th} row vector of K coefficients for variable j. For the last penalty term $\|\beta_j\|_q$, if $q=2$, it is a grouped-lasso penalty on all the K coefficients for the particular variables. The tuning parameter λ controls the overall strength of the penalty.

The algorithm flow of multinomial logistic regression with elastic net (MLR-elastic net) is shown in Algorithm 1. It is based on the generalized linear statistical learning model with LK pre-processing.

Algorithm 1. Multinomial logistic regression with elastic net based on linearized kernel pre-processing

1: Input: $\mathrm{X_{train}}=[\mathrm{X}_l,\ldots,\mathrm{X}_L]$, $\mathrm{X_{test}}$, the kernel κ, sampling-method, c, k
2: $\mathrm{X}_R=\mathrm{sub\_sample}\ (\mathrm{X_{train}}\ ,\ \text{sampling-method},c)$
3: Compute $\mathrm{C_{train}}=\mathrm{K}\left(\mathrm{X_{train}},\mathrm{X}_R\right)$
4: Compute $\mathrm{W}=\mathrm{K}\left(\mathrm{X}_R,\mathrm{X}_R\right)$
5: Approximate W_k using k largest eigenvalues and eigenvectors $\mathrm{W}_k=\mathrm{V}_k\Sigma_k\mathrm{V}_k^T$
6: Compute virtual train set $\mathrm{F_{train}}=\left(\Sigma_k^{\dagger}\right)^{1/2}\mathbf{V}_k^T\mathbf{C}_{\text{train}}^T$
7: Compute $\mathrm{C_{test}}=\mathrm{K}\left(\mathrm{X_{test}},\mathrm{X}_R\right)$
8: Compute virtual test set $\mathrm{F_{test}}=\left(\Sigma_k^{\dagger}\right)^{1/2}\mathbf{V}_k^T\mathbf{C}_{\text{test}}^T$
9: Using $\mathrm{F_{train}}$ to obtain the model parameters by multinomial logistic regression with elastic net based on generalized linear model
10: Carry out classification of $\mathrm{F_{test}}$ using the model obtained above
11: Output: classification result of $\mathrm{F_{test}}$

For the binomial logistic regression problem, the generalized additive model is adopted. The GAMSEL R package is used. The generalized additive model uses overlap grouped-lasso penalties, it can select whether a term in a general additive model is zero, linear, or a non-linear spline for Gaussian or binomial applications [15]. We adopt LK pre-processing for binomial logistic regression based on the GAMSEL. The algorithm flow is the same as in Algorithm 1. The difference only lies in feeding the virtual samples to the GAMSEL model. Figure 1 shows the block diagram of our proposed method.

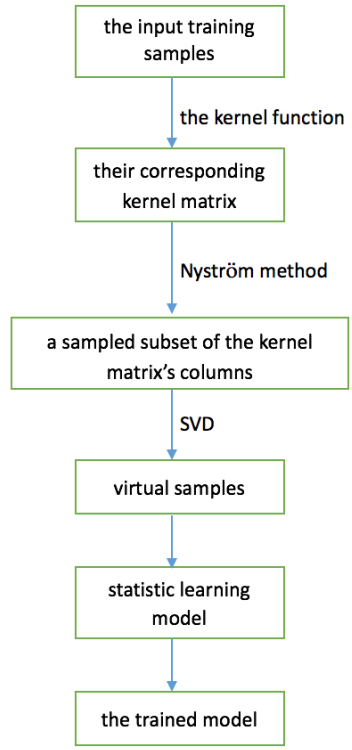


Figure 1. Block diagram of the proposed method

5 EXPERIMENT AND SIMULATION

In this section, we evaluate the performance of multinomial/binomial logistic regression based on a statistical learning model with LK pre-processing on face recognition databases and spectra dataset.

5.1 Evaluation on Face Recognition Database

The adopted face recognition databases are the Extended YaleB face database and AR-face database [52, 53]. The "Extended YaleB" face database has 38 classes with 2 414 frontal face images taken under varying lighting conditions. Each class nearly has 64 images. The AR Face database possesses 126 classes with 4 000 color images, which are with different lighting conditions, facial variations, and facial disguises for each class. For the sake of fairness and convenience for contrast experiments, our experimental details are configured with the same settings as [25]. For LK pre-

processing, the Gaussian kernel and the fusion kernel [30] are used. The specific parameters of the Gaussian kernel are configured with the same configuration as in [25]. For a fair comparison, in the case of LC-KSVD1 and LC-KSVD2, the parameters are chosen as identical to the best classification result in [25]. The classification results were shown in Table 1 with our method comparing with the best results obtained in [25]. The LK pre-processing with the fusion kernel is already stated in Table 1. Other LK pre-processing is with the Gaussian kernel. It can be seen that the addition of LK pre-processing and the elastic net's grouping effect can increase the prediction accuracy. The experiment result also shows that the LK pre-processing with the manual fusion kernel outperforms that with the Gaussian kernel. The fusion kernel can exploit the reciprocating properties of the Euclidean and cosine distance measures. We also used the support vector machine(SVM) toolbox LIBSVM [54] in our experiments. We use a coarse grid search for the SVM parameters, we use the Gaussian kernel. We use the grid search strategy to look for suitable parameters. We chose $g \in [0.001\ 0.001\ 0.01\ 0.1\ 0.5\ 1\ 10\ 100\ 500\ 1\,000]$ and $c \in [0.000006\ 0.000008\ 0.000009\ 0.0000092\ 0.0000093\ 0.0000095\ 0.0000096\ 0.0000097\ 0.0000098\ 0.00001]$ and run the search for 100 times. And choosing the parameters for the best accuracy. Experiment result with SVM using LIBSVM toolbox shows inferior accuracy compared to that of the proposed method. For method comparisons, we use the same random training and testing samples for each algorithm. Our method improves the classification results by up to 2.42 % and 4.8 %, when compared to LC-KSVD2 results [25] on the Extended YaleB face dataset and AR-face dataset, respectively. Due to the grouping effect of the elastic net, a group of related or correlated variables can be detected, when the LK pre-processing is adopted, more grouped features from the high dimension kernel space can be incorporated, so it can obtain the grouped features in the high dimension, which can lead to increased performance. Figure 2 shows the solution path of multinomial regression with the elastic net for the 38^{th} subject on Yale-B face database based on the Glmnet. Each curve corresponds to a variable. Each curve shows the solution path of its coefficient against the L2 norm of the whole coefficient vector as $\log(\lambda)$ varies. The axis above indicates the number of nonzero coefficients at each corresponding $\log(\lambda)$. There are about 332 variables selected for this algorithm when $\log(\lambda)$ is equal to -8. It shows that more variables will be shrunk to be zero eventually as $\log(\lambda)$ increases. The λ represents the tuning parameter in Equation (10) for controlling the overall strength of the penalty.

To further justify the performance of our method, on the AR face database, we perform simulations by randomly selecting training samples, and the remaining samples are used as testing samples. Figure 3 shows the classification accuracy of a total of 20 experiments by randomly selecting 20 training samples from the input samples. The Y-axis is the testing accuracy and the horizontal axis is the experiment number(ID). We adopted different kernel functions for LK preprocessing, the solid line in Figure 3 uses the Gaussian kernel, the dashed line in Figure 3 uses the fusion kernel. The result shows that the proposed method is effective.

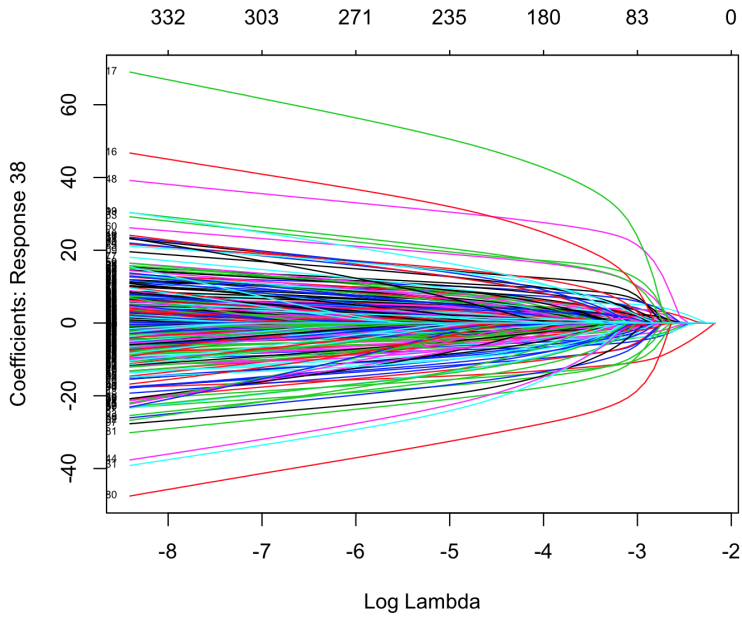


Figure 2. Solution path of multinomial logistic regression with elastic net for the 38^{th} subject on Extended Yale-B face database

We also use 10 times 5-fold cross validation to further show the effectiveness of our method on the Extended Yale-B face database and AR face database. The experiment was shown in Table 2.

For the LK pre-processing, the time complexity is $\mathrm{O}\left(Nck + c^2k\right)$, $\mathrm{O}(Nck)$ represents the complexity of getting the virtual samples, $\mathrm{O}\left(c^2k\right)$ stands for the eigenvalue decomposition of the sampled kernel matrix. Although the process of computing the

Algorithm	Yale-B	AR-Face
LC-KSVD1	94.49	92.5
LC-KSVD1 + LK	96.08	94.8
LC-KSVD2	94.99	93.7
LC-KSVD2 + LK	96.58	94.8
SVM	91.32	95.5
SVM + LK	95.41	94.7
MLR-elastic net	93.90	97.3
MLR-elastic net+ LK	97.16	98.3
MLR-elastic net+ LK (fusion kernel)	**97.41**	**98.5**

Table 1. Classification accuracy of LC-KSVD1, LC-KSVD2, SVM and our method on Extended Yale-B and AR face database, with or without LK pre-processing

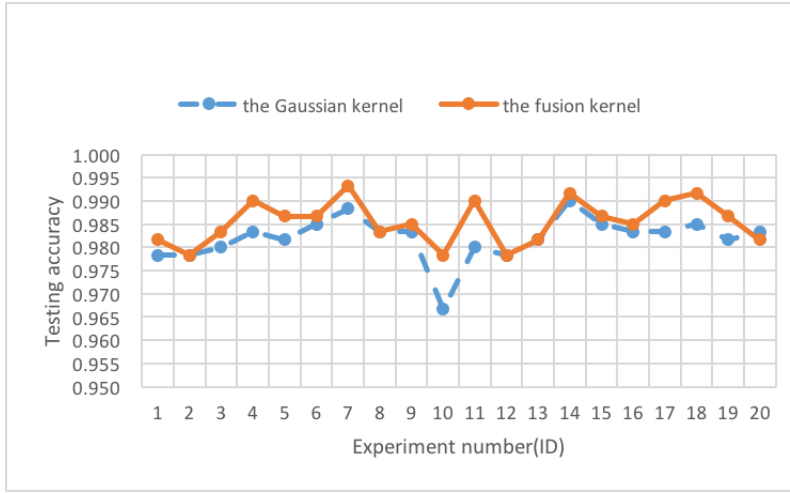


Figure 3. Classification accuracy by randomly selecting training samples with different kernel function on AR face database

Algorithm	Yale-B	AR-Face
SVM + LK	95.41	94.7
MLR-elastic net+ LK (fusion kernel)	**98.73**	**98.62**

Table 2. Average classification accuracy of our method on Extended Yale-B and AR face database with LK pre-processing by 10 rounds 5-fold cross validation

virtual samples may seem inefficient, it is only performed once, after which the complexity is dictated by the chosen model. The total training time and test time required to classify one sample with or without LK pre-processing on the AR face are shown in Table 3. The experiment is carried out on MacBook Pro with a 2.6 GHz Intel Core i5 processor and 8 GB 1 600 MHz DDR3 memory. Its operation system is OS X Yosemite 10.10.2.

Algorithm	Total training time	Test time for one sample
MLR-elastic net	579.503 s	0.006 s
MLR-elastic net + LK (fusion kernel)	**323.125 s**	**0.004 s**

Table 3. Total training time and testing time required to classify one sample with or without LK pre-processing on AR face database

It shows that the training time and test time on the statistical model is decreased greatly with LK pre-processing.

5.2 Evaluation on Spectra Data

We also evaluated our method on the spectral dataset. A publicly available near infrared (NIR) transmittance dataset and a Raman transmittance spectroscopy dataset are adopted. for evaluating the method. This dataset is about Escitalopram tablets from the pharmaceutical company H. Lundbeck A/S. The tablets are qualitatively divided into four categories according to dosage values of this pharmaceutical drug, which are 5, 10, 15, and 20 mg tablets, respectively. Classification is carried out on the four types of tablets. The instrument for collecting the Raman spectra data of each tablet is a Perkin–Elmer System 2000 NIR FT-Raman spectrometer equipped with a diode pumped Nd:YAG laser emitting 400 mW at $v_0 = 9\,394.69\,\mathrm{cm}^{-1}$ and an InGaAs detector. Its Raman wavenumber shifts range is 200–3 600 cm^{-1} with the interval of 1 cm^{-1} and the resolution of 8 cm^{-1} [55]. Please refer to the following website for further detail about the dataset: `http://www.models.life.ku.dk/Tablets`. The NIR transmittance spectra data include 310 samples with 404 variables. Approximately 80 samples belong to each class. In the experiment, nine–tenths samples of each class were randomly selected for training, and the remaining samples were used for testing. We repeated the experiment 10 times and compared the classification results with and without LK pre-processing, as shown in Table 4. Our method improves the testing accuracy by 6.78 % on this NIR transmittance spectra data.

Algorithm	Testing accuracy on NIR transmittance spectra data
MLR-elastic net	89.67
MLR-elastic net + LK (fusion kernel)	**96.45**

Table 4. Classification accuracy of the method on near infrared transmittance dataset with and without LK pre-processing

The tablets' Raman spectra data include 120 samples. Approximately 30 samples belong to each class. We conducted a binomial logistic regression experiment on a subset of the Raman spectra data. The first class and the fourth class were selected as the subset of the Raman spectra data, corresponding to the tablets with a dosage of 5 and 20 mg, respectively. Each of these two classes has 30 samples. We used 10 rounds of 5-fold cross-validation to further illustrate the effectiveness of our method. In the experiment, for each class, 25 samples were chosen for training, and the remaining 5 samples were selected for evaluation. Binomial logistic regression based on the generalized additive model was adopted for fitting the regularization path of the data. We compared the classification results based on the statistical learning model GAMSEL and SVM with or without LK preprocessing, as shown in Table 5. The classification accuracy was improved by GAMSEL; therefore, exploring the LK feature space with the generalized additive model is more effective than using SVM. The experiment shows that exploration of the LK feature space based on a statistical learning model is effective.

Algorithm	Testing accuracy on one subset of Raman spectra data
SVM	96
SVM + LK	94.83
GAMSEL	**100**
GAMSEL + LK	**100**

Table 5. 10 rounds of 5-fold cross-validation classification accuracy of SVM and GAMSEL on one subset of Raman spectra data with LK pre-processing

6 ANALYSIS AND CONCLUSION

This paper proposes a novel method that integrates LK pre-processing into a statistical learning model for classification. The Gaussian kernel and the fusion of Gaussian and cosine kernels are adopted for linearized kernel pre-processing. For multinomial logistic regression, we use the elastic net's grouping effect to find the grouping features in the high dimension features space. Experimental results on the Extended Yale-B database and AR-Face database demonstrated the good performance of the multinomial regression with elastic net methods based on the statistical learning model. This method can overcome the restriction of a small number of samples. The elastic net penalty can guarantee the robustness of the least square solution and strengthen the sparseness of the solution vector so that the model is more parsimonious and its accuracy is greatly improved. The elastic net can also deal with high dimensional and small sample data effectively, and the model can obtain a good trade-off between sparsity and prediction accuracy. A relatively high accuracy model can be established with the combination of LK pre-processing and elastic net based on statistical learning.

We also examined the classification of the different dosages of the active substance in Escitalopramtablets using Raman transmittance spectroscopy. This method also makes progress in classification accuracy with LK pre-processing on spectral data. In this study, the experiment was further carried out on a subset of the Raman transmittance spectroscopy dataset as a binomial logistic regression problem. We adopt the GAMSEL R package to capture the generalized additive model within the LK pre-processing feature spaces. The GAMSEL still can be used to fit the nonlinearity on the linearized kernel feature space. But the relatively small number of samples leads to kind of over fitting in it. The experiment shows that it is effective to explore the linearized kernel feature space based on the statistical learning model. We found that linearized kernel pre-processing is an effective descending dimension method and the fusion of Gaussian and cosine kernels for linearized kernel pre-processing with improved accuracy compared with a single Gaussian kernel. It provides us a new perspective to explore the low-dimensional space embedded with LK pre-processing in the high dimension data. Nowadays deep convolution neural network has achieved satisfactory performance in many fields. One of our future directions is to use the proposed model to explore the deep CNN feature space.

We will also focus on combining our method with Chemometric for spectral data analysis.

Acknowledgements

This work was supported by the Youth Science Foundation Project of Zhejiang Natural Science Foundation: Study on Grouping Characteristics of High Dimensional Data in Spectral Data Analysis (LQ19F020006).

REFERENCES

[1] Hastie, T.—Tibshirani, R.—Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, USA, 2001.

[2] James, G.—Witten, D.—Hastie, T.—Tibshirani, R.: An Introduction to Statistical Learning: With Applications in R. Springer, New York, USA, 2013.

[3] Tibshirani, R.: Regression Shrinkage Selection via the LASSO. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 58, 1996, No. 1, pp. 267–288.

[4] Donoho, D. L.: For Most Large Underdetermined Systems of Equations, the Minimal $l(1)$-Norm Near-Solution Approximates the Sparsest Near-Solution. Communications on Pure and Applied Mathematics, Vol. 59, 2006, No. 7, pp. 907–934, doi: 10.1002/cpa.20131.

[5] Donoho, D. L.—Elad, M.: Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via $\ell 1$ Minimization. Proceedings of the National Academy of Sciences of the United States of America (PNAS), Vol. 100, 2003, No. 5, pp. 2197–2202, doi: 10.1073/pnas.0437847100.

[6] Donoho, D. L.—Huo, X.: Uncertainty Principles and Ideal Atomic Decomposition. IEEE Transactions on Information Theory, Vol. 47, 2001, No. 7, pp. 2845–2862, doi: 10.1109/18.959265.

[7] Chen, S. S.—Donoho, D. L.—Saunders, M. A.: Atomic Decomposition by Basis Pursuit. SIAM Review, Vol. 43, 2001, No. 1, pp. 129–159, doi: 10.1137/S003614450037906X.

[8] Efron, B.—Hastie, T.—Johnstone, I.—Tibshirani, R.: Least Angle Regression. The Annals of Statistics, Vol. 32, 2004, No. 2, pp. 407–499, doi: 10.1214/009053604000000067.

[9] Candes, E. J.—Romberg, J.—Tao, T.: Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 2, pp. 489–509, doi: 10.1109/TIT.2005.862083.

[10] Candes, E. J.—Tao, T.: Near-Optimal Signal Recovery from Random Projections: Universal Encoding Strategies? IEEE Transactions on Information Theory, Vol. 52, 2006, No. 12, pp. 5406–5425, doi: 10.1109/TIT.2006.885507.

[11] Donoho, D. L.: Compressed Sensing. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 4, pp. 1289–1306, doi: 10.1109/TIT.2006.871582.

[12] Donoho, D. L.—Elad, M.—Temlyakov, V. N.: Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. IEEE Transactions on Information Theory, Vol. 52, 2006, No. 1, pp. 6–18, doi: 10.1109/TIT.2005.860430.

[13] Tsaig, Y.—Donoho, D. L.: Extensions of Compressed Sensing. Signal Processing, Vol. 86, 2006, No. 3, pp. 549–571, doi: 10.1016/j.sigpro.2005.05.029.

[14] Zou, H.—Hastie, T.: Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 67, 2005, No. 2, pp. 301–320, doi: 10.1111/j.1467-9868.2005.00503.x.

[15] Chouldechova, A.—Hastie, T.: Generalized Additive Model Selection. Statistics, 2015, arXiv: 1506.03850v2.

[16] Powers, S.—Hastie, T.—Tibshirani, R.: Nuclear Penalized Multinomial Regression with an Application to Predicting at Bat Outcomes in Baseball. Statistical Modelling, Vol. 18, 2018, No. 5-6, pp. 388–410, doi: 10.1177/1471082X18777669.

[17] Zhang, Z.—Lai, Z.—Xu, Y.—Shao, L.—Wu, J.—Xie, G. S.: Discriminative Elastic-Net Regularized Linear Regression. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 3, pp. 1466–1481, doi: 10.1109/TIP.2017.2651396.

[18] Li, G. Z.—Wang, S. T.: Face Recognition Based on Sparse Representation and Elastic Network. Journal of Computer Applications, Vol. 37, 2017, No. 3, pp. 901–905, doi: 10.11772/j.issn.1001-9081.2017.03.901.

[19] Jiang, Z. L.—Zhe, L.—Davis, L. S.: Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, 2013, No. 11, pp. 2651–2664, doi: 10.1109/TPAMI.2013.88.

[20] Naseem, I.—Togneri, R.—Bennamoun, M.: Linear Regression for Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, 2010, No. 11, pp. 2106–2112, doi: 10.1109/TPAMI.2010.128.

[21] Wright, J.—Yang, A. Y.—Ganesh, A.—Sastry, S. S.—Ma, Y.: Robust Face Recognition via Sparse Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, 2009, No. 2, pp. 210–227, doi: 10.1109/TPAMI.2008.79.

[22] Hassaballah, M.—Aly, S.: Face Recognition: Challenges, Achievements, and Future Directions. IET Computer Vision, Vol. 9, 2015, No. 4, pp. 614–626, doi: 10.1049/iet-cvi.2014.0084.

[23] Parkhi, O. M.—Vedaldi, A.—Zisserman, A.: Deep Face Recognition. Proceedings of the British Machine Vision Conference (BMVC), 2015, doi: 10.5244/c.29.41.

[24] Gao, Y.—Ma, J.—Yuille. A. L.: Semi-Supervised Sparse Representation Based Classification for Face Recognition with Insufficient Labeled Samples. IEEE Transactions on Image Processing, Vol. 26, 2017, No. 5, pp. 2545–2560, doi: 10.1109/TIP.2017.2675341.

[25] Golts, A.—Elad, M.: Linearized Kernel Dictionary Learning. IEEE Journal of Selected Topics in Signal Processing, Vol. 10, 2016, No. 4, pp. 726–739, doi: 10.1109/JSTSP.2016.2555241.

[26] Levy, S.—Fullagar, P. K.: Reconstruction of a Sparse Spike Train from a Portion of Its Spectrum and Application to High-Resolution Deconvolution. Geophysics, Vol. 46, 1981, No. 9, pp. 1235–1243, doi: 10.1190/1.1441261.

[27] Kim, S. J.—Koh, K.—Lustig, M.—Boyd, S.—Gorinevsky, D.: An Interior-Point Method for Large-Scale $l(1)$-Regularized Least Squares. IEEE Journal of Selected Topics in Signal Processing, Vol. 1, 2007, No. 4, pp. 606–617, doi: 10.1109/JSTSP.2007.910971.

[28] Zou, H.: The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, Vol. 101, 2006, No. 476, pp. 1418–1429, doi: 10.1198/016214506000000735.

[29] Boser, B. E.—Guyon, I. M.—Vapnik, V. N.: A Training Algorithm for Optimal Margin Classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92), 1992, pp. 144–152, doi: 10.1145/130385.130401.

[30] Khan, S.—Naseem, I.—Togneri, R.—Bennamoun, M.: A Novel Adaptive Kernel for the RBF Neural Networks. Circuits, Systems, and Signal Processing, Vol. 36, 2017, No. 4, pp. 1639–1653, doi: 10.1007/s00034-016-0375-7.

[31] Pillonetto, G.—Dinuzzo, F.—Chen, T.—De Nicolao, G.—Ljung, L.: Kernel Methods in System Identification, Machine Learning and Function Estimation: A Survey. Automatica, Vol. 50, 2014, No. 3, pp. 657–682, doi: 10.1016/j.automatica.2014.01.001.

[32] Zhang, L.—Zhou, W. D.—Chang, P. C.—Liu, J.—Yan, Z.—Wang, T.—Li, F. Z.: Kernel Sparse Representation-Based Classifier. IEEE Transactions on Signal Processing, Vol. 60, 2012, No. 4, pp. 1684–1695, doi: 10.1109/TSP.2011.2179539.

[33] Zhan, Y.—Dai, S.—Mao, Q.—Liu, L.—Sheng, W.: A Video Semantic Analysis Method Based on Kernel Discriminative Sparse Representation and Weighted KNN. The Computer Journal, Vol. 58, 2015, No. 6, pp. 1360–1372, doi: 10.1093/comjnl/bxu121.

[34] Yang, Y.—Li, D.—Duan, Z.: Chinese Vehicle License Plate Recognition Using Kernel-Based Extreme Learning Machine with Deep Convolutional Features. IET Intelligent Transport Systems, Vol. 12, 2018, No. 3, pp. 213–219, doi: 10.1049/iet-its.2017.0136.

[35] Van Nguyen, H.—Patel, V. M.—Nasrabadi, N. M.—Chellappa, R.: Design of Non-Linear Kernel Dictionaries for Object Recognition. IEEE Transactions on Image Processing, Vol. 22, 2013, No. 12, pp. 5123–5135, doi: 10.1109/TIP.2013.2282078.

[36] Van Nguyen, H.—Patel, V. M.—Nasrabadi, N. M.—Chellappa, R.: Kernel Dictionary Learning. Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 2021–2024, doi: 10.1109/ICASSP.2012.6288305.

[37] Yin, J.—Liu, Z.—Jin, Z.—Yang, W.: Kernel Sparse Representation Based Classification. Neurocomputing, Vol. 77, 2012, No. 1, pp. 120–128, doi: 10.1016/j.neucom.2011.08.018.

[38] Gao, S. H.—Tsang, I. W. H.—Chia, L. T.: Kernel Sparse Representation for Image Classification and Face Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (Eds.): Computer Vision – ECCV 2010. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 6314, 2010, pp. 1–14, doi: 10.1007/978-3-642-15561-1_1.

[39] Chen, Z. H.—Zuo, W. M.—Hu, Q. H.—Lin, L.: Kernel Sparse Representation for Time Series Classification. Information Sciences, Vol. 292, 2015, pp. 15–26, doi: 10.1016/j.ins.2014.08.066.

[40] Gangeh, M. J.—Ghodsi, A.—Kamel, M. S.: Kernelized Supervised Dictionary Learning. IEEE Transactions on Signal Processing, Vol. 61, 2013, No. 19, pp. 4753–4767, doi: 10.1109/TSP.2013.2274276.

[41] Harandi, M. T.—Sanderson, C.—Hartley, R.—Lovell, B. C.: Sparse Coding and Dictionary Learning for Symmetric Positive Definite Matrices: A Kernel Approach. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.): Computer Vision – ECCV 2012. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7573, 2012, pp. 216–229, doi: 10.1007/978-3-642-33709-3_16.

[42] Shrivastava, A.—Nguyen, H. V.—Patel, V. M.—Chellappa, R.: Design of Non-Linear Discriminative Dictionaries for Image Classification. In: Lee, K. M., Matsushita, Y., Rehg, J. M., Hu, Z. (Eds.): Computer Vision – ACCV 2012. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7724, 2012, pp. 660–674, doi: 10.1007/978-3-642-37331-2_50.

[43] Bryt, O.—Elad, M.: Compression of Facial Images Using the K-SVD Algorithm. Journal of Visual Communication and Image Representation, Vol. 19, 2008, No. 4, pp. 270–282, doi: 10.1016/j.jvcir.2008.03.001.

[44] Zepeda, J.—Guillemot, C.—Kijak, E.: Image Compression Using Sparse Representations and the Iteration-Tuned and Aligned Dictionary. IEEE Journal of Selected Topics in Signal Processing, Vol. 5, 2011, No. 5, pp. 1061–1073, doi: 10.1109/JSTSP.2011.2135332.

[45] Elad, M.—Aharon, M.: Image Denoising via Sparse and Redundant Representations over Learned Dictionaries. IEEE Transactions on Image Processing, Vol. 15, 2006, No. 12, pp. 3736–3745, doi: 10.1109/TIP.2006.881969.

[46] Fadili, M. J.—Starck, J.-L.—Murtagh, F.: Inpainting and Zooming Using Sparse Representations. The Computer Journal, Vol. 52, 2009, No. 1, pp. 64–79, doi: 10.1093/comjnl/bxm055.

[47] Mairal, J.—Elad, M.—Sapiro, G.: Sparse Representation for Color Image Restoration. IEEE Transactions on Image Processing, Vol. 17, 2008, No. 1, pp. 53–69, doi: 10.1109/TIP.2007.911828.

[48] Engan, K.—Aase, S. O.—Hakon Husoy, J.: Method of Optimal Directions for Frame Design. Proceedings of 1999 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99), Vol. 5, 1999, pp. 2443–2446, doi: 10.1109/ICASSP.1999.760624.

[49] Aharon, M.—Elad, M.—Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. IEEE Transactions on Signal Processing, Vol. 54, 2006, No. 11, pp. 4311–4322, doi: 10.1109/TSP.2006.881199.

[50] Jiang, Z.—Lin, Z.—Davis, L. S.: Learning a Discriminative Dictionary for Sparse Coding via Label Consistent K-SVD. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1697–1704, doi: 10.1109/CVPR.2011.5995354.

[51] Friedman, J.—Hastie, T.—Tibshirani, R. et al.: Lasso and Elastic-Net Regularized Generalized Linear Models. 2017, available at: http://web.stanford.edu/~hastie/glmnet/glmnet_beta.html.

[52] Martinez, A.—Benavente, R.: The AR Face Database. CVC Technical Report No. 24, 1998, available at: http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html.

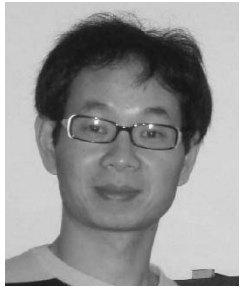
[53] Georghiades, A. S.—Belhumeur, P. N.—Kriegman, D. J.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, 2001, No. 6, pp. 643–660, doi: 10.1109/34.927464.

[54] Chang, C. C.—Lin, C. J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, 2011, No. 3, Art. No. 27, pp. 1–27, doi: 10.1145/1961189.1961199.

[55] Dyrby, M.—Engelsen, S. B.—Nørgaard, L.—Bruhn, M.—Lundsberg-Nielsen, L.: Chemometric Quantitation of the Active Substance (Containing $C \equiv N$) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. Applied Spectroscopy, Vol. 56, 2002, No. 5, pp. 579–585, doi: 10.1366/0003702021955358.

Xiaochun Guan received her B.Sc. and M.Sc. degrees in measurement technology and automation apparatus from the University of Shanghai for Science and Technology, Shanghai, in 2002 and 2005, respectively. She joined Wenzhou University in 2005 where she is currently Associate Professor in the College of Electrical and Electronic Engineering. She is now pursuing the Ph.D. degree in the School of Computer Science and Technology, Zhejiang University of Technology. Her research interests include end-side deep neural network deployment, machine learning, sparse representation and statistical learning.

Jianhua Zhang received his Ph.D. degree from the University of Hamburg, Hamburg, Germany in 2012. He is currently Professor with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China. His current research interests include SLAM, 3D vision, reinforcement learning, and machine vision.



Shengyong Chen received his Ph.D. degree in robot vision from the City University of Hong Kong, Hong Kong, in 2003. He is currently Professor with Tianjin University of Technology, China. He received a fellowship from the Alexander von Humboldt Foundation of Germany and worked with the University of Hamburg, Hamburg, Germany, from 2006 to 2007. He has authored over 100 scientific papers in international journals and is an inventor of over 100 patents. His research interests include computer vision, robotics, and image analysis. He is Fellow of IET and Senior Member of CCF. He was the recipient of the National Outstanding Youth Foundation Award of China in 2013.