Discriminative sparse representation learning using multiclass hinge loss

Ryota Kamiya and Yoshikazu Washizawa
The University of Electro-Communications, Tokyo, Japan.
E-mail: k1731051@edu.cc.uec.ac.jp

***Abstract*—Sparse representation methods have been researched widely in recent years. Sparse representation classification methods, such as sparse representation classifier (SRC) and label-consistent K-SVD (singular value decomposition) learn classification parameters, dictionary, and sparse representation simultaneously, so that they find an optimal sparse representation to discriminate categories. However, these classifiers use least square error (LSE) strategy to design the classifiers. LSE of the empirical risk is not optimal for classifiers because even if a training sample correctly classified, it may increase the empirical cost. We, therefore, introduce the hinge loss to design the classifier. The hinge loss is employed in support vector machines, and it shows better performance than LSE based methods. We provide an optimization algorithm to minimize the proposed criterion that is the linear combination of the hinge loss and sparse representation error. Experimental results show that the proposed method exhibited conventional sparse representation classification methods.**

I. Introduction

As the internet of things (IoT) grows, the redundancy of data increases. Exploration of latent information is important, and the sparse representation is one of the promising tools in signal processing and machine learning. Sparse representation compactly expresses data in a vector having a few non-zero elements using an overcomplete dictionary, and it has been widely applied in many fields such as image classification and recognition, feature learning, image noise removal, and sensing [1], [3]. Compressed sensing is a method to acquire and reconstruct signals in compressed form using sparse representation [3]. The dictionary learning methods such as K-SVD (singular value decomposition) have also been proposed to obtain the overcomplete dictionary using given dataset, and applied to image compression and denoising [6].

For classification of redundant input data, sparse representation classification methods have been proposed. The sparse representation classifier (SRC) classifies input vectors using reconstruction residual of sparse representation for each class [2]. The label consistent K-SVD (LC-KSVD) is a classification method using K-SVD. LC-KSVD simultaneously minimizes reconstruction error of the sparse representation, and empirical classification error using class labels of training data [11]. LC-KSVD has been extended to the joint embedding and dictionary learning (JEDL) and the locality preserving KSVD (LP-KSVD) [4], [5]. However, these classifiers use the squared error between class labels and the output of classifiers, that does not evaluate misidentification directly. The squared error may be increased even if a training data is correctly classified.

In this paper, we therefore, introduce the multiclass hinge loss for sparse representation classifier. The hinge loss is employed in support vector machines (SVMs) and shows better performance than least square error for classification problems since the hinge loss is always zero if a training data is correctly classified with appropriate margin. We define a new optimization criterion to minimize both reconstruction error of the sparse representation and misidentification error using the hinge loss. We provide an optimization algorithm to minimize our criterion.

II. Related works

We briefly introduce sparse expression methods of sparse representation classifier (SRC), K-SVD and LC-KSVD1 and LC-KSVD2.

A. Sparse representation classifier (SRC)

SRC is a classifier using the sparse representation [2]. First, SRC obtains dictionaries $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K] \in \mathbb{R}^{n \times K}$ for each class using labeled training data, where n is the number of input dimension, and K is the number of dictionary vectors (atoms). For unknown input $\boldsymbol{x}$, its sparse representation $\boldsymbol{s}$ is obtained by,

$$\min_{\boldsymbol{s}} \|\boldsymbol{D}\boldsymbol{s} - \boldsymbol{x}\|^2 + \lambda_1 \|\boldsymbol{s}\|_1, \tag{1}$$

where λ_1 is a parameter for trading-off the reconstruction error and the sparsity. Then, the residual $r_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{D}\delta_i(\boldsymbol{s})\|_2$ is obtained for each class $i = 1, 2, \ldots, C$, where $\delta_i(\boldsymbol{s})$ is set to zero except for non-zero elements of $\boldsymbol{s}$ associated with class $i \in 1, 2, ..., C$. $\boldsymbol{x}$ is assigned to the class having the minimum residual $r_i(\boldsymbol{x})$, $i = 1, \ldots, C$.

B. K-SVD

K-SVD is a method to learn an overcomplete dictionary for sparse representation by minimizing the reconstruction error [6]. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{n \times N}$ be a set of n-dimensional training vector, where N is the number of training data. Then the optimization problem of K-SVD is

$$\min_{\boldsymbol{D}, \boldsymbol{S}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{S}\|_F^2, \quad \text{subject to } \|\boldsymbol{s}_i\|_0 \leq T_1, \tag{2}$$

where $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{R}^{K \times N}$ is a matrix of sparse vectors, $\|\cdot\|_0$ is a l_0 norm which is the number of non-zero elements, T_1 is the sparsity constraint factor.

K-SVD alternately updates $\boldsymbol{S}$ and $\boldsymbol{D}$. The update of the sparse matrix $\boldsymbol{S}$ by fixing the dictionary $\boldsymbol{D}$ is done for

each column using some sparse coding algorithm such as the orthogonal matching pursuit (OMP) [7]. The update of $\boldsymbol{D}$ is done for each column of $\boldsymbol{D}$, i.e., $\arg\min_{\boldsymbol{d}_i} \|\boldsymbol{X} - \boldsymbol{d}_i\tilde{\boldsymbol{s}}_i^\top - \sum_{j\neq i}\boldsymbol{d}_j\tilde{\boldsymbol{s}}_j^\top\|_F^2$, where $\tilde{\boldsymbol{s}}_i$ is the ith row of $\boldsymbol{S}$. The singular value decomposition (SVD) is used to obtain $\boldsymbol{d}_i$.

C. LC-KSVD

The label consistent K-SVD (LC-KSVD) is a classification method using K-SVD [12]. LC-KSVD considers additional two terms to be minimized.

$$\min_{\boldsymbol{D},\boldsymbol{S},\boldsymbol{A},\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{S}\|_F^2 + \alpha\|\boldsymbol{Q} - \boldsymbol{A}\boldsymbol{S}\|_F^2 + \beta\|\boldsymbol{H} - \boldsymbol{W}\boldsymbol{S}\|_F^2$$
$$\text{subject to } \|\boldsymbol{s}_i\|_0 \leq T_1, \quad i = 1, \ldots, N. \tag{3}$$

$\boldsymbol{Q} \in \mathbb{R}^{K\times N}$ is the discriminative sparse code of the input signal for classification.

$$[\boldsymbol{Q}]_{i,j} = \begin{cases} 1 & \text{(the class label of } \boldsymbol{d}_i \text{ and } \boldsymbol{s}_j \text{ are the same)} \\ 0 & \text{(otherwise).} \end{cases} \tag{4}$$

$\boldsymbol{A} \in \mathbb{R}^{K\times K}$ is a linear transform, so that the sparse vectors having the same class label are linearly transformed to similar code vectors. $\alpha > 0$ is a parameter to define the strength of the second term.

The third term is to design a classifier. $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_K] \in \mathbb{R}^{C\times K}$ is the classifier to be trained, and $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_N] \in \mathbb{R}^{C\times N}$ is a class label matrix of input signals, i.e., the jth element of $\boldsymbol{h}_i$ is one if $\boldsymbol{h}_i$ belongs the jth class, and otherwise zero. β is a parameter to define the strength of the third term. LC-KSVD1 is the case of $\beta = 0$, otherwise, it is denoted by LC-KSVD2. Jiang et al. [12] provided a K-SVD based algorithm to minimize the optimization problem (3).

For unknown input signal $\boldsymbol{x} \in \mathbb{R}^n$, its sparse representation $\boldsymbol{s} \in \mathbb{R}^K$ is obtained. LC-KSVD2 estimates the class label as the largest element of $\boldsymbol{W}\boldsymbol{s}$. In the LC-KSVD1, the class label is estimated in the same manner as SRC.

The squared error used in (3) is not optimal for classification. Let us consider a position that $\boldsymbol{Q}$ or $\boldsymbol{H}$ is one. If $\boldsymbol{A}\boldsymbol{S}$ or $\boldsymbol{W}\boldsymbol{S}$ has a value that is greater than one in the position, the total cost increases although it is not disadvantageous for classification. In a similar manner, if $\boldsymbol{A}\boldsymbol{S}$ or $\boldsymbol{W}\boldsymbol{S}$ has a value that is smaller than zero in the position that $\boldsymbol{Q}$ or $\boldsymbol{H}$ is zero, the total cost increases although it is not disadvantageous for classification. Therefore, the squared error criteria do not evaluate misidentification.

III. Sparse representation classifier using hinge loss

We introduce the hinge loss to the sparse representation classifier. The hinge loss directly evaluates the misidentification, and is used in SVMs. Let $y \in \{-1, +1\}$ be a binary class label, $\boldsymbol{x}$ be corresponding input vector, and $f(\boldsymbol{x})$ be a decision function of a binary classifier, i.e. if $f(\boldsymbol{x}) > 0$, $\boldsymbol{x}$ is assigned to the positive class, and otherwise, $\boldsymbol{x}$ is assigned to the negative class. The hinge loss is defined by

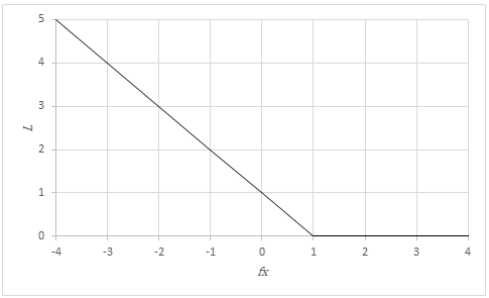

Fig. 1. hinge loss of binary classification

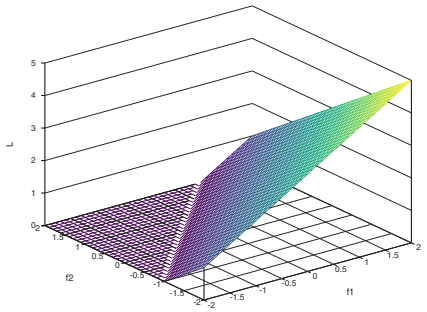

Fig. 2. multiclass hinge loss

$$L(f) = \begin{cases} 1 - yf(\boldsymbol{x}) & (yf(\boldsymbol{x}) \leq 1) \\ 0 & (yf(\boldsymbol{x}) \geq 1). \end{cases} \tag{5}$$

If $yf(\boldsymbol{x}) > 0$, the sample $\boldsymbol{x}$ is correctly classified by f. The hinge loss has the margin one, and linearly give the penalty of $yf(\boldsymbol{x}) \leq 1$ (Fig. 1).

The hinge loss is extended for multi-class problems. Let $f_i(\boldsymbol{x})$ is the decision function for the ith class, i.e., $\boldsymbol{x}$ is assigned to $\arg\max_{i=1,\ldots,C} f_i(\boldsymbol{x})$. The hinge loss for multi-class is

$$L(f) = \max\{0, 1 - (f_c(\boldsymbol{x}) - v)\}, \tag{6}$$
$$v = \max_{i\in\{1,\ldots,C\}\setminus\{c\}} f_i(\boldsymbol{x}) \tag{7}$$

where c is the class label of $\boldsymbol{x}$. v is the maximum value of decision functions of the other classes. If the difference between the decision function value of the true class label c, $f_c(\boldsymbol{x})$ and v is greater than one, $L(f)$ linearly gives the penalty (Fig. 2).

We define the linear multiclass decision function for sparse vector $\boldsymbol{s}$, $f_i(\boldsymbol{s}) = \boldsymbol{w}_i^\top\boldsymbol{s} + b_i$, and let $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C] \in \mathbb{R}^{K\times C}$, $\boldsymbol{b} = [b_1, \ldots, b_C] \in \mathbb{R}^C$. Then the optimization problem of the proposed method is

$$\min_{\boldsymbol{D},\boldsymbol{S},\boldsymbol{W},\boldsymbol{b}} J = \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{s}_i\|^2 + \lambda\|\boldsymbol{s}_i\|_1 + \mu L(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{s}_i), \tag{8}$$

where $L(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{s}_i)$ is the multiclass hinge loss function.

We alternately update $(\boldsymbol{W}, \boldsymbol{b})$, $\boldsymbol{s}_i$, $i = 1, \ldots, N$, and $\boldsymbol{D}$.

a) Update of $\boldsymbol{W}$ and $\boldsymbol{b}$: Let us optimize J for $\boldsymbol{W}$ and $\boldsymbol{b}$ while $\boldsymbol{D}$ and $\boldsymbol{s}_n$ $n = 1, \ldots, N$ are fixed. Then the first and second terms are constant for $\boldsymbol{W}$ and $\boldsymbol{b}$. Then we consider the subproblem,

$$\min_{\boldsymbol{W}, \boldsymbol{b}} \sum_{i=1}^{N} L(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{s}_i). \tag{9}$$

The multiclass hinge loss (6) can be rewritten by using slack variable ξ,

$$\begin{aligned} L(f) = &\min \xi \\ &\text{subject to } \xi \geq 0 \\ &\quad \xi \geq 1 - (f_c(\boldsymbol{x}) - f_i(\boldsymbol{x})),\ i \in \{1, \ldots, C\} \setminus \{c\}, \end{aligned} \tag{10}$$

Then the subproblem (13) is reduced to the following linear programming,

$$\begin{aligned} &\min_{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\xi}} \sum_{i=1}^{N} \xi_i \\ &\text{subject to } \xi_i \geq 0,\ i = 1, \ldots, N \\ &\qquad \xi_i \geq 1 - (\boldsymbol{w}_c - \boldsymbol{w}_k)^\top \boldsymbol{s}_i + (b_c - b_k), \\ &\qquad\quad i = 1, \ldots, N,\ k \in \{1, \ldots, C\} \setminus \{c\}. \end{aligned} \tag{11}$$

The solution can be obtained by a solver such as GLPK. Alternatively, this is a special case of multiclass SVM [13], and its implementations such as liblinear [14] obtain the solution.

b) Update of $\boldsymbol{s}_n$: The sparse vector $\boldsymbol{s}_i$ can be optimized for each i. We, here omit the subscript i, and let $\boldsymbol{s} = \boldsymbol{s}_+ - \boldsymbol{s}_-$, $(\boldsymbol{s}_+ \geq 0,\ \boldsymbol{s}_- \geq 0)$ and $\boldsymbol{z} = [\boldsymbol{s}_+^\top | \boldsymbol{s}_-^\top]^\top \in \mathbb{R}^{2K}$ $(\boldsymbol{z} \geq 0)$, where "$\geq$" is for each element of vectors. Then the second term, the l_1 regularization, of J can be rewritten to

$$\|\boldsymbol{s}\|_1 = \min_{\boldsymbol{z}} \mathbf{1}_{2K}^\top \boldsymbol{z}, \quad \text{subject to } \boldsymbol{z} \geq 0,\ \ [\boldsymbol{I}_K | -\boldsymbol{I}_K]\boldsymbol{z} = \boldsymbol{s},$$

where $\mathbf{1}_{2K}$ is a $(2K)$-dimensional vector whose elements are one, and $\boldsymbol{I}_K$ is the identity matrix of size K.

Let $\tilde{\boldsymbol{w}}_i = [\boldsymbol{w}_i^\top | -\boldsymbol{w}_i^\top]$. Then the third term of J is transformed by using Eq. (10),

$$\begin{aligned} L(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{s}) = &\min \xi \\ &\text{subject to } \xi \geq 0, \\ &\quad \xi \geq 1 - (\tilde{\boldsymbol{w}}_c - \tilde{\boldsymbol{w}}_i)^\top \boldsymbol{z} + (b_c - b_i), \\ &\quad i \in \{1, \ldots, C\} \setminus \{c\}, \end{aligned} \tag{12}$$

where c is the class label of $\boldsymbol{x}$.

By letting $\tilde{\boldsymbol{z}} = [\boldsymbol{z}^\top | \xi]^\top \in \mathbb{R}^{2K+1}$, and $\tilde{\boldsymbol{D}} = [\boldsymbol{D}| - \boldsymbol{D}]$, the optimization problem (8) is reduced to the following constrained quadratic programming,

$$\begin{aligned} &\min_{\tilde{\boldsymbol{z}}} \tilde{\boldsymbol{z}}^\top \begin{bmatrix} \tilde{\boldsymbol{D}}^\top \tilde{\boldsymbol{D}} & \mathbf{0}_{2K} \\ \mathbf{0}_{2K}^\top & 0 \end{bmatrix} \tilde{\boldsymbol{z}} + [(\lambda \mathbf{1}_{2K}^\top) - 2\boldsymbol{x}^\top \tilde{\boldsymbol{D}} | \mu] \tilde{\boldsymbol{z}} \\ &\text{subject to } \tilde{\boldsymbol{z}} \geq 0 \\ &\quad [(\tilde{\boldsymbol{w}}_c^\top - \tilde{\boldsymbol{w}}_i^\top) | 1] \tilde{\boldsymbol{z}} \geq 1 - (b_c - b_i) i \in \{1, \ldots, C\} \setminus \{c\}. \end{aligned} \tag{13}$$

c) Update of $\boldsymbol{D}$: Since the second and third terms of J are constant for $\boldsymbol{D}$, $\boldsymbol{D}$ is updated by the least square,

$$\boldsymbol{D}^{\text{new}} = \arg\min_{\boldsymbol{D}} \|\boldsymbol{X} - \boldsymbol{D}\boldsymbol{S}\|_F^2 \tag{14}$$

$$= \boldsymbol{X}\boldsymbol{S}^\top (\boldsymbol{S}\boldsymbol{S}^\top)^{-1}. \tag{15}$$

$\boldsymbol{D}$ is initiated by randomly selecting vectors from $\boldsymbol{X}$, then $\boldsymbol{D}$ and $\boldsymbol{S}$ are initiated by the standard K-SVD [6]. We summarize our algorithm in Algorithm 1.

Algorithm 1 Optimization algorithm of the proposed method

Require: $\boldsymbol{X}$
Ensure: $\boldsymbol{D}, \boldsymbol{S}, \boldsymbol{W}, \boldsymbol{b}$
1: Initialize $\boldsymbol{D}^{(0)}$ by selecting randomly from $\boldsymbol{X}$
2: Compute $\boldsymbol{D}^{(0)}$ and $\boldsymbol{S}^{(0)}$ by using K-SVD
3: **repeat**
4: Update $\boldsymbol{W}^{(t)}$, $\boldsymbol{b}^{(t)}$ by (11) ;
5: **for** $i = 1$ **to** N **do**
6: Obtain $\tilde{\boldsymbol{z}}$ by Eq. (13)
7: Update $\boldsymbol{s}_i = [\boldsymbol{I}_K | -\boldsymbol{I}_K]\boldsymbol{z}$
8: **end for**
9: Compute $\boldsymbol{D}^{(t)}$ by (15) ;
10: **until** J converges

IV. Experiment

We used the Isolet spoken letter recognition database. It has 150 subjects speaking the name of each letter of the alphabet twice. The described features contain spectral coefficients, contour features, sonorant features, pre-sonorant features, and postsonorant features. The number of input dimension n is 617, and the number of classes C is 26. The speakers are grouped into sets of 30 speakers each and are named Isolet 1 to Isolet 5. In the experiment, Isolet 1 to Isolet 5 are used to evaluate the performance of sparse codes expression classifiers. We conducted five fold cross validation for Isolet 1 to Isolet 5, i.e., each dataset is divided into five subsets, and four subsets are used for training, and remaining one subset is used for testing. We evaluated averaged classification accuracy for five folds. α and β refer to the parameters the reference paper. The sparsity T_1 and dictionary vector K used grid search. For the experiment the parameter setting was as follows; $\alpha = 3.0$, $\beta = 4.0$, the sparsity $T_1 = 10$, and the number of dictionary vectors $K = 25$.

We show the experimental results in Table. I. The classification accuracy is almost the same for SRC, LCKSVD1 and LCKSVD2.

Table. II represents the number of nonzero elements in a sparse vector. We used OMP for SRC, LCKSVD1, and LCKSVD2, that is, the number of nonzero elements is directly determined by the hyperparameter. For this reason the number of nonzero elements in sparse vectors are the same for SRC and LC-KSVDs. The our method improves the identification rate more than the related method. Moreover the number of nonzero elements of the proposed method is large.

TABLE I
CLASSIFICATION ACCURACY

	Method	
	SRC Mean ±STD (%)	LCKSVD1 Mean ±STD (%)
Isolet1	91.73±1.44	89.16±2.91
Isolet2	89.87±1.19	86.85±1.46
Isolet3	85.32±2.72	79.61±2.82
Isolet4	80.10±3.57	76.63±3.16
Isolet5	88.77 ±1.65	83.96±2.16
Mean	87.16±2.11	86.60±2.50
	LC-KSVD2 Mean ±STD (%)	Our Model Mean ±STD (%)
Isolet1	90.57±2.10	95.83±0.32
Isolet2	87.56±1.45	91.45±1.17
Isolet3	81.73±1.63	87.65±2.40
Isolet4	79.59 ±1.14	88.38±0.55
Isolet5	86.01 ±0.85	86.71±2.36
Mean	84.85±1.43	90.04 ±1.36

TABLE II
NUMBER OF NON-ZERO IN SPARSE REPRESENTATION

SRC	10
LC-KSVD1	10
LC-KSVD2	10
Our method	11

V. CONCLUSION AND FUTURE WORK

We have proposed a sparse representation classification method using multiclass hinge loss. The hinge loss directly evaluate the empirical misidentification risk, and optimization is reduced to a linear programming. We provided an algorithm to minimize the criterion which is the linear combination of sparse representation error and the hinge loss. The algorithm alternately updates weights for classifier $\boldsymbol{W}$ and $\boldsymbol{b}$, the sparse representation $\boldsymbol{S}$, and the dictionary $\boldsymbol{D}$, thus the sparse representation of input vector $\boldsymbol{x}$ is not only approximation of $\boldsymbol{x}$, but also discriminative representation.

Future works include the investigation of the relation between sparsity level, dictionary size, and classification accuracy, and a method using the l_0 norm constraint.

REFERENCES

[1] M. Elad, *Sparse and Redundant Representations,* Springer, 2010.

[2] J. Wang, C. Y. Lu, M. Wang, P.P. Li, S.C. Yan, and X.G. Hu, "Robustface recognition via adaptive sparse representation," *IEEE Trans. Cybern.,* vol. 44, pp. 2368–2378, 2014.

[3] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory* vol. 52, pp. 1289-1306, 2006.

[4] Z. Zhang, F. Li, T.W.S. Chow, L. Zhang, and S. Yan, "Sparse codes auto-extractor for classification: a joint embedding and dictionary learning framework for representation," *IEEE Trans. Signal Process.* vol. 64, pp. 3790-3805, 2016.

[5] Y.-S. Lee, C.-Y. Wang, S. Mathulaprangsan, J.-H. Zhao, and J.-H. Zhao, "Locality-preserving K-SVD Based Joint Dictionary and Classifier Learning for Object Recognition," *Proceedings of the 2016 ACM on Multimedia Conference,* pp. 481-485, 2016.

[6] M. Aharon, M. Elad, and A.M. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.* vol. 54, no. 11, pp. 4311-4322, 2006.

[7] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing,* vol. 20, no. 1, pp. 33-61, 1998.

[8] Y. C. Pati, R. Rezaiifar and S. K. Perinkulam, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition." *IEEE Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on.*, 1993, pp.40-44.

[9] D. R. Musicant, "MATLAB/CPLEX MEX-Files", *Computer Sciences Department, University of Wisconsin, Madison, www.cs.wisc.edu/∼musicant/data/cplex/.,* 2000

[10] Z. Qiang, and B. Li, "Discriminative K-SVD for dictionary learning in face recognition." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.*, IEEE, 2010, pp.2691-2698.

[11] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," *Computer Vision and Pattern Recognition, 2011 IEEE Conference on.*, IEEE, 2011, pp.1697-1704.

[12] Z. L. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol.35, 2013, pp.2651-2664.

[13] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research,* vol 2, pp. 265–292, 2001.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research,* vol 9, pp. 1871–1874, 2008.