Robust Image Set Classification Using Partial Least Squares

Hui Jin[1] and Ruiping Wang[2]

[1] Peking University, Beijing, 100871, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
hjin@jdl.ac.cn, wangruiping@ict.ac.cn

Abstract. Image set classification has recently attracted increasing research interest in the field of visual information processing. Different from previous methods that usually characterize set data distribution explicitly using some parametric or non-parametric models, this paper proposes a simple yet effective Partial Least Squares (PLS) regression based method, which seeks to directly learn the underlying statistical relationship between the distributions of set data and their class memberships. With no assumption on the form of set data distribution, the learned model finally reduces to an efficient linear regression from the data space to the class label space, facilitating robust classification of novel test data. Experiments on face recognition and object categorization have shown that the proposed method is competitive to the state-of-the-arts and also quite robust to the noisy set data in practical applications.

Keywords: image set classification, PLS, regression.

1 Introduction

In recent years, with the increase of available video cameras and large capacity storage media, many new applications are emerging, such as visual surveillance, video retrieval, digital photo albums management, etc. In such applications, each object of interest can have a number of image sets for both training and testing, where each image set generally contains lots of images belonging to the same class and covering large appearance variations in pose, lighting, and non-rigid deformations. This is the so-called image set classification problem. By efficiently exploiting the rich set information, more robust object classification can be expected under more realistic conditions [1], [5], [14].

During the past decade, a number of methods have been proposed to solve the problem of image set classification [1], [2], [4], [5]. Generally, these methods make different prior assumptions on the form of set data distribution, and exploit either parametric or non-parametric mathematical models to explicitly characterize data variations in the image set. For parametric modeling, single Gaussian and Gaussian mixture models (GMM) have been explored in earlier works [1], [10] as the parametric distribution function of the image set. Their success highly rely on the assumption that the training and novel test data sets have strong statistical correlations.

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 200–207, 2013.

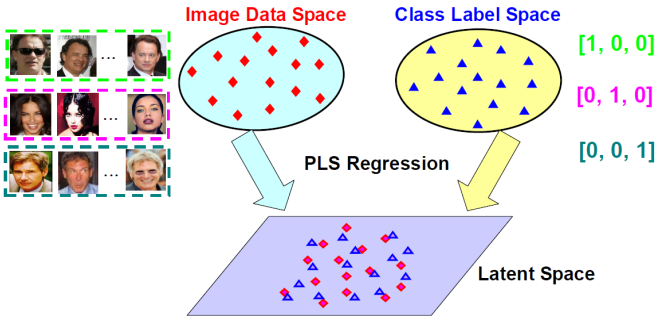

Fig. 1. The basic overview of the proposed SLR method. In the PLS learned latent space, the covariance of the projected image samples and their associated class labels are maximized.

For non-parametric modeling, one class of prevalent methods are based on the model of single linear subspace [5], [15] or more sophisticated nonlinear manifold [6], [12], [14]. Since they can flexibly characterize complex data variation, such methods have gained wide success in past several years. However, linear subspace is a relatively loose representation of the data distribution as noted in [2], while manifold typically needs a large data for reliable estimation, which are unavailable in some practical applications. More recently, a new type of non-parametric methods based on affine subspace model have been introduced [2], [4]. While data variations can be effectively handled, such methods are shown to be sensitive to outliers and have much higher computational cost, due to their inherent single sample-based matching mechanism [13], [14].

In this paper, we propose a simple yet effective Sample-Label Regression (SLR) approach to image set classification. By exploiting the Partial Least Squares (PLS) regression, SLR aims to directly learn the underlying statistical relationship between the set samples distribution and their class labels distribution. Different from previous methods, our approach makes no assumption on the form of set data distribution and the learned model finally reduces to an efficient linear regression from the image data space to the class label space. When applying the learned model to a novel test image set, it only involves computing the class membership score for each image in the set using linear regression, and then aggregating these scores to finally determine the label of the whole set. Fig.1 illustrates the basic idea of the proposed SLR method.

2 PLS-Based Image Set Classification

Formally, given m training image sets as: $S_1, S_2, ..., S_m$, we denote $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,Ni}]$ $(i = 1, 2, ..., m)$ as the data matrix of the i-th image set S_i with Ni samples, where $\boldsymbol{x}_{i,j} \in \Re^d$ is the j-th image sample with d-dimensional feature description. Each set belongs to one of c object classes denoted by $\{L_i \mid L_i \in \{1, 2, ..., c\}\}_{i=1}^{m}$. To facilitate following description, we group all training image samples in a single data matrix: $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_m]^T$ of size $N \times d$, where each row of the matrix is an image sample and $N = \sum_{i=1}^{m} Ni$ is the total number of image samples from all m training image sets.

In the next, we first introduce the basic mathematical model of the Partial Least Squares (PLS) method including both its linear and kernel formulations. Then we elaborate the training and testing framework of exploiting PLS for the specific task of image set classification.

2.1 Background of Partial Least Squares (PLS)

Partial Least Squares (PLS) is a wide class of methods for modeling relations between two sets of observed variables by means of latent variables. In its general form, PLS creates score/latent vectors by using existing correlations between different sets of variables while also keeping most of the variance of both sets. Please refer to [8] for more details.

Let $\boldsymbol{x} \in \mathcal{X} \subset \mathfrak{R}^d$ denote a d-dimensional vector of predictor variables in the first set of data and similarly $\boldsymbol{y} \in \mathcal{Y} \subset \mathfrak{R}^c$ denote a c-dimensional vector of response variables from the second set. Observing N data samples from each set of variables, PLS decomposes matrix $\boldsymbol{X}_{N\times d}$ (it has the same meaning as the above total training data matrix $\boldsymbol{X}$) and $\boldsymbol{Y}_{N\times c}$ into the form

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \\ \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} \end{aligned} \tag{1}$$

where $\boldsymbol{T}$ and $\boldsymbol{U}$ are $N \times p$ matrices containing the extracted p latent vectors, the $d \times p$ matrix $\boldsymbol{P}$ and the $c \times p$ matrix $\boldsymbol{Q}$ represent loadings, and the $N \times d$ matrix $\boldsymbol{E}$ and the $N \times c$ matrix $\boldsymbol{F}$ are the residuals. Basically, PLS proceeds to find weight vectors $\boldsymbol{w}$, $\boldsymbol{v}$ such that

$$\max_{|\boldsymbol{w}|=|\boldsymbol{v}|=1} [cov(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{Y}\boldsymbol{v})]^2 = [cov(\boldsymbol{t}, \boldsymbol{u})]^2, \tag{2}$$

where $\boldsymbol{t}$ and $\boldsymbol{u}$ are the column vectors of $\boldsymbol{T}$ and $\boldsymbol{U}$ respectively, $cov(\boldsymbol{t}, \boldsymbol{u})$ is the sample covariance. Grouping the sequentially obtained weight vectors $\boldsymbol{w}_i$ in a matrix $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_p]$, the regression coefficients between the two sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be estimated by:

$$\boldsymbol{B} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}\boldsymbol{T}^T\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{U}(\boldsymbol{T}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{U})^{-1}\boldsymbol{T}^T\boldsymbol{Y}, \tag{3}$$

which results in the linear PLS regression $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B}$ [8].

Since it can often bring desirable performance gain by extending linear methods in a so-called RKHS (reproducing kernel Hilbert space) feature space via the kernel trick, in [9] the kernel formulation of PLS (KPLS) has been presented. The basic idea is to map the original $\mathcal{X}$-space data into a RKHS feature space $\mathcal{F}$ with $\phi: \mathfrak{R}^d \mapsto \mathcal{F}$, where an inner product can be defined using the kernel function as: $\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and perform the kernel form of the optimization in Eq. (2). Let $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_N)]^T$ be the feature matrix of the training points, the kernel

Gram matrix can thus be written as $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ with $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then the regression coefficients $\boldsymbol{B}_\phi$ in the feature space will have the form:

$$\boldsymbol{B}_\phi = \boldsymbol{\Phi}^T \boldsymbol{U}(\boldsymbol{T}^T \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{T}^T\boldsymbol{Y}, \tag{4}$$

Given a testing data example $\boldsymbol{x}_t \in \Re^d$ in the $\mathcal{X}$-space, its KPLS prediction $\boldsymbol{y}_t$ in the $\mathcal{Y}$-space can be obtained by

$$\boldsymbol{y}_t^T = [\phi(\boldsymbol{x}_t)]^T \boldsymbol{B}_\phi = \boldsymbol{K}_t^T \boldsymbol{U}(\boldsymbol{T}^T \boldsymbol{K}\boldsymbol{U})^{-1}\boldsymbol{T}^T\boldsymbol{Y}, \tag{5}$$

where $\boldsymbol{K}_t = [k(\boldsymbol{x}_1, \boldsymbol{x}_t), \ldots, k(\boldsymbol{x}_N, \boldsymbol{x}_t)]^T$.

2.2 Exploiting PLS for Image Set Classification

As illustrated in Fig.1, we exploit PLS regression to directly learn the underlying statistical relationship between the set samples distribution and their class labels distribution. Specifically, we use the training image sets S_i and their associated class labels L_i ($i = 1, 2, ..., m$) to learn the PLS or KPLS latent model. As described in Sec.2.1, the total training data matrix $\boldsymbol{X}$ of size $N \times d$ acts as the *predictor* matrix $\boldsymbol{X}_{N\times d}$. For each training image sample $\boldsymbol{x}_{i,j}$ ($i = 1, 2, ..., m$, $j = 1, 2, ..., Ni$) with its corresponding set class label L_i, we define its class membership indicator vector: $\boldsymbol{y}_{i,j} = [0, ..., 1, ..., 0]^T \in \Re^c$, where the k-th entry being 1 and all other entries being 0 indicates that $\boldsymbol{x}_{i,j}$ belongs to the *k*-th class. The *response* matrix $\boldsymbol{Y}_{N\times c}$ can then be easily constructed with $\boldsymbol{y}_{i,j}^T$ as its row vector. Taking the two matrices $\boldsymbol{X}_{N\times d}$ and $\boldsymbol{Y}_{N\times c}$ as input, either linear PLS or KPLS can then be used to learn the regression model in Eq. (3) or (4) respectively. In the KPLS formulation, we choose the widely used Gaussian RBF kernel function:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / \sigma^2). \tag{6}$$

In the testing phase, suppose we are given a test image set S_t with $\boldsymbol{X}_t = [\boldsymbol{x}_{t,1}, \boldsymbol{x}_{t,2}, ..., \boldsymbol{x}_{t,M}]$ as its data matrix containing M image samples, the classification task is to determine the class label of the image set. To this end, for each individual sample $\boldsymbol{x}_{t,j}$ ($j = 1, 2, ..., M$) we first compute its class membership indicator vector $\boldsymbol{y}_{t,j}$ (which is c-dimensional) using the PLS/KPLS regression model in Eq. (3) or (4). By aggregating these individual vectors, we then obtain the indicator vector of the whole image set as:

$$\boldsymbol{y}_t = \frac{1}{M}\sum\nolimits_{j=1}^{M} \boldsymbol{y}_{t,j}. \tag{7}$$

Intuitively, the weight score in the k-th entry of $\boldsymbol{y}_t$ indicates the probability that the set belongs to the *k*-th class. Thus, the entry index with the largest score in $\boldsymbol{y}_t$ finally determines the class label of the test image set S_t.

From above analysis, it can be seen that our method has the following advantages: (1) it makes no assumption of data distribution, thus can be stably applied in different scenarios; (2) it effectively integrates set information from individual samples, resulting in quite robust and efficient classification. Such properties will be verified in the following experiments.

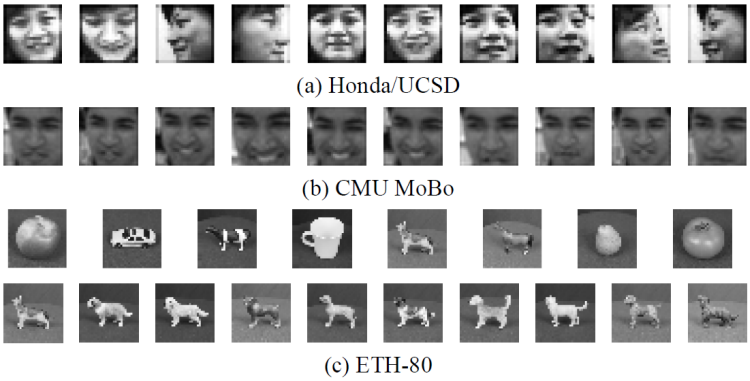

Fig. 2. Example images of the three benchmark databases. In (a) and (b), each row shows representative facial images from one video sequence of an individual. In (c), the first row shows images of the 8 different categories, and the second row shows example images of the 10 different objects for one category.

3 Experiments

We evaluate the proposed method on three widely used datasets: Honda/UCSD [6], CMU MoBo [3] for image sets based face recognition, and ETH-80 [7] for object categorization.

3.1 Databases and Settings

The **Honda/UCSD** consists of 59 video sequences of 20 persons and each video contains about 300~500 frames covering large variations in head pose and facial expression. The **CMU MoBo** contains 96 sequences of 24 subjects and each subject has 4 sequences captured in different walking situations. Each sequence has about 300 frames. We used a cascaded face detector [11] to collect faces in each video, and then resized each face to a 20×20 intensity image. Histogram equalization was used to eliminate lighting effects. Each video generated an image set of faces. The **ETH-80** contains images of 8 categories with each category including 10 objects. Each object has 41 images of different views which form an image set. 20×20 intensity images were also used. Fig. 2 shows some example images from each of the three databases.

For comparison with the literature, we adopted the same protocol as [2],[13]. On all three datasets, we conducted ten-fold cross validation experiments, i.e., 10 randomly selected training/testing combinations, to report average recognition rates of

different methods. Specifically, for both Honda and MoBo, each person had one image set for training and the rest sets for testing. For ETH-80, each category had 5 objects for training and the other 5 objects for testing.

3.2 Comparative Methods and Settings

We compared our approach with several representative non-parametric methods for image set classification, including (i) Mutual Subspace Method (MSM) [15] as the baseline linear subspace based method, and (ii) Affine Hull based Image Set Distance (AHISD) [2], (iii) Convex Hull based Image Set Distance(CHISD) [2], (iv) Sparse Approximated Nearest Point (SANP) [4], which are all affine subspace based methods recently proposed in the literature.

For fair comparison, the key parameters of each method were empirically tuned according to the recommendations in the original references as well as the source codes provided by the original authors. In MSM, PCA was performed to learn the linear subspaces by preserving 95% of data energy. For both AHISD and CHISD, we used their linear version and retained 95% energy by PCA. The error penalty in CHISD was set to $C = 100$ as [2]. For SANP, we adopted the same weight parameters as [4] for the convex optimization.

In our proposed SLR method, we tested both the linear and kernel PLS regression models, referred to as "SLR_L" and "SLR_K" respectively. From Sec.2.1 it can be seen the PLS model has only one parameter, i.e., the number of latent vectors p, which was fixed to 100 in all three databases. We found the classification accuracy was quite stable while varying this number in our experiments. In the KPLS model, there is another important parameter, i.e., the window width σ in the RBF kernel Eq. (6). It was adaptively set for each dataset as the mean of all sample pairs' Euclidean distances.

3.3 Results and Analysis

We tabulate the classification results of all methods on the three datasets in Tab. 1. Each reported rate is an average over the ten-fold trials. From the comparison results, we have the following observations: (1) Our method is very competitive to the state-of-the-art ones in all three datasets. In both Honda and ETH-80 datasets, our method delivered the highest rate, and in MoBo dataset, our kernel regression ranked the second highest among all methods. (2) Compared with the linear PLS model, our kernel PLS model can further boost the performance with a modest margin, indicating that the difficult linearly inseparable problem can be effectively alleviated by the nonlinear kernel mapping. (3) In the ETH-80 object dataset, it is interesting to find that the affine subspace methods [2], [4] exhibit much lower accuracy due to that the common intra-class object variations cannot be handled adequately by the single points based matching.

Table 1. Average classification rate of different methods on three datasets by ten-fold trials

Datasets	MSM [15]	AHISD [2]	CHISD [2]	SANP [4]	SLR_L	SLR_K
Honda/ UCSD	0.925	0.885	0.905	0.936	0.931	0.956
CMU MoBo	0.852	0.951	0.940	0.963	0.930	0.954
ETH-80	0.878	0.773	0.735	0.755	0.895	0.903

We further conducted experiment to test the robustness of different methods to a practical challenge where the image sets have noisy data from other classes. We followed the same setting as [2] to study this problem on the dataset Honda/UCSD. We tested three cases in which the training gallery and/or the testing probe sets were corrupted by adding one image from each of the other classes. The original clean data and the three noisy cases are referred to as "Clean", "N_G" (only gallery has noise), "N_P" (only probe has noise), and "N_G+P" (both) respectively. Fig. 3 demonstrates the comparison result. It can be seen that our proposed SLR method shows high robustness against the noisy data challenge, with quite slight accuracy drop. This is mainly attributed to the advantage of our method by effectively integrating the set information from individual samples, without any assumption on the data distribution. Another finding is that the linear subspace based method MSM is more stable than the affine subspace based ones AHISD/CHISD and SANP, since the former treats the set samples as a whole and can suppress the noisy data effectively.

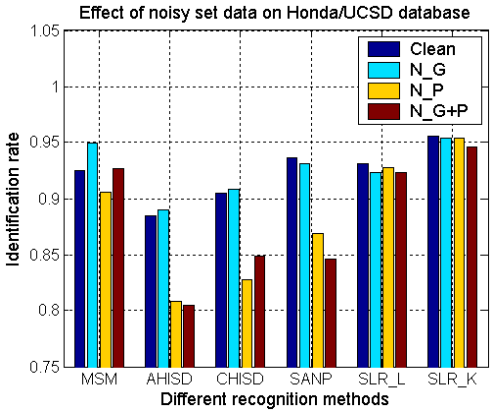

Fig. 3. Comparison of different methods on the practical problem of noisy set data

4 Conclusions

In this paper, we have proposed a simple yet effective Sample-Label Regression (SLR) approach to image set classification. With no assumption on the form of set data distribution, our approach exploits the Partial Least Squares (PLS) to directly

learn an efficient linear regression model from the image data space to the class label space. When applying the learned model to classify novel test image set, it involves only simple linear operations and can effectively integrate the set information from individual samples. The extensive experimental results have shown the effectiveness of our method and its favorable robustness to noisy set data in practical applications.

Acknowledgments. This paper is partially supported by Natural Science Foundation of China under contracts No. 61001193 and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract No.4111003.

References

1. Arandjelović, O., Shakhnarovich, G., Fisher, J., Cipolla, R., Darrell, T.: Face Recognition with Image Sets Using Manifold Density Divergence. In: CVPR, pp. 581–588 (2005)
2. Cevikalp, H., Triggs, B.: Face Recognition Based on Image Sets. In: CVPR, pp. 2567–2573 (2010)
3. Gross, R., Shi, J.: The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University (2001)
4. Hu, Y., Mian, A.S., Owens, R.: Sparse Approximated Nearest Points for Image Set Classification. In: CVPR (2011)
5. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. PAMI 29(6), 1005–1018 (2007)
6. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In: CVPR, pp. 313–320 (2003)
7. Leibe, B., Schiele, B.: Analyzing Appearance and Contour Based Methods for Object Categorization. In: CVPR, vol. 2, pp. 409–415 (2003)
8. Rosipal, R., Krämer, N.C.: Overview and Recent Advances in Partial Least Squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) SLSFS 2005. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006)
9. Rosipal, R., Trejo, L.J.: Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. J. Machine Learning Research 2(2), 97–123 (2001)
10. Shakhnarovich, G., Fisher III, J.W., Darrell, T.: Face Recognition from Long-term Observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 851–865. Springer, Heidelberg (2002)
11. Viola, P., Jones, M.: Robust Real-Time Face Detection. Int'l J. Computer Vision 57(2), 137–154 (2004)
12. Wang, R., Chen, X.: Manifold Discriminant Analysis. In: CVPR, pp. 429–436 (2009)
13. Wang, R., Guo, H., Davis, L., Dai, Q.: Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification. In: CVPR, pp. 2496–2503 (2012)
14. Wang, R., Shan, S., Chen, X., Dai, Q., Gao, W.: Manifold-Manifold Distance and Its Application to Face Recognition with Image Sets. IEEE Transactions on Image Processing 21(10), 4466–4479 (2012)
15. Yamaguchi, O., Fukui, K., Maeda, K.: Face Recognition Using Temporal Image Sequence. In: FG, pp. 318–323 (1998)