LETTER

Locality-Constrained Multi-Task Joint Sparse Representation for Image Classification

Lihua GUO[†a)], ***Member***

SUMMARY In the image classification applications, the test sample with multiple man-handcrafted descriptions can be sparsely represented by a few training subjects. Our paper is motivated by the success of multi-task joint sparse representation (MTJSR), and considers that the different modalities of features not only have the constraint of joint sparsity across different tasks, but also have the constraint of local manifold structure across different features. We introduce the constraint of local manifold structure into the MTJSR framework, and propose the Locality-constrained multi-task joint sparse representation method (LC-MTJSR). During the optimization of the formulated objective, the stochastic gradient descent method is used to guarantee fast convergence rate, which is essential for large-scale image categorization. Experiments on several challenging object classification datasets show that our proposed algorithm is better than the MTJSR, and is competitive with the state-of-the-art multiple kernel learning methods.

key words: *image classification, multi-task learning, sparse representation, manifold learning*

1. Introduction

Thousands of images are generated every day, and it is necessary to classify, organize and access them by methods that are easier and faster. With the exponential growth in digital images number, the need for semantic image classification is becoming increasingly important to support effective image database indexing and retrieval. Semantic images categorization, especially large scale image categorization, is a challenging and important problem nowadays.

Many hand-crafted methods [1]–[6] have been proposed to measure object similarity for object classification. A recent trend is to combine these discriminative features for class-level object classification. One popular method in machine learning is Multiple Kernel Learning (MKL) [7]–[9], which can be seen to linearly combine similarity functions between images such that the combined similarity function yields improved classification performance. Another popular method is sparse coding, which has received wide interest in the field of visual recognition. Wright [10] used the Lasso regularization to select some representative training subjects from the entire training set, and proposed sparse representation classification (SRC) method to implement robust face recognition. Obozinski [11] regarded the sparse representation as a combinational model of group Lasso [12] and multi-task Lasso [13], and proposed Multi-task Joint Covariate Selection (MTJCS) by penalizing the sum of l_2 norms on blocks of coefficients. Yuan [14] further proposed the multi-task joint sparse representation (MTJSR), and introduced this powerful sparse learning model into computer vision as a joint sparse visual representation method. The MTJSR also provided two kernel extensions to fuse the discriminative power of different visual descriptor kernels in recognition problems. Recently, some researchers extended this framework with extra contexts, such as, Chen [15] proposed an approach, which incorporated a new type of context(label exclusive context) with linear representation and classification, to multi-label image classification. This model, a recent advance in sparse learning, formulated the problem of linear representation and classification as an exclusive lasso model. Liu [16] proposed a regularized multi-task learning approach to train multiple binary-class Semi-Supervised Support Vector Machines (S3VMs). This method was used to solve the problem of multi-class classification problem in semi-supervised setting. Moreover, the framework of sparse representation has achieved success in many fields, such as the human gait recognition [17], face recognition [18], visual tracking [19] and others [20], [21]. In this paper, motivated by the success of multi-task sparse joint sparse representation, we introduce the local manifold into the sparse representation, and propose a locality-constrain multi-task joint sparse representation.

2. Locality-Constrained Multi-Task Joint Sparse Representation

First, let's review the MTJSR [14]. In the MTJSR, the general problem of jointly estimating models from multiple related data sets was often referred to as multi-view learning or multi-task learning in the machine learning literature. The representation task was defined as a training set $X^k = \left[X_1^k, \ldots, X_J^k\right]$ with J classes and a query data y^k to be represented, where $k \in \{1, \ldots, K\}$ was tasks. MTJSR aimed to find out a very few common classes of training samples that were mostly useful for query data reconstruction in these K tasks. For object recognition, we may generate K representation tasks from K different modalities of features associated with the same visual input. The goal of joint sparsity can be achieved by imposing $\ell_{1,2}$ mixed-norm penalty on the reconstruction coefficients. The mathematic formulated objective is as follows:

Manuscript received April 19, 2013.
Manuscript revised May 27, 2013.

[†]The author is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China.
a) E-mail: guolihua@scut.edu.cn

DOI: 10.1587/transinf.E96.D.2177

Copyright © 2013 The Institute of Electronics, Information and Communication Engineers

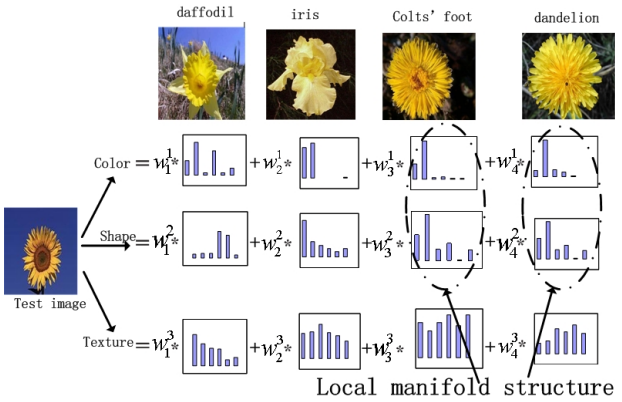

Fig. 1 The working mechanism of our LC-MTJSR algorithm.

$$\min_{W} \frac{1}{2}\sum_{k=1}^{K}\left\|y^k - \sum_{j=1}^{J} X_j^k w_j^k\right\|_2^2 + \lambda \sum_{j=1}^{J}\left\|w_j\right\|_2^1 \tag{1}$$

But, in the real situation, not only the different modalities of features have the constraint of joint sparsity across different tasks, but also these features have local manifold structure. Let's take flower image classification as an example, which is shown in Fig. 1. The color and shape descriptions are the image features, and among them, the Colts' foot class and dandelion class have similar feature values; thus, the features of these two images have local manifold structure, which can be added into the formulated objective to explicitly encourage system to choose the similar training image for data reconstruction. We present this new coding algorithm called Locality-constrained multi-task joint sparse representation (LC-MTJSR). As suggested by LCC [22], locality is more essential than sparsity, because locality must lead to sparsity but not necessary vice versa. The LC-MTJSR method incorporates locality constraint instead of the sparsity constraint in Eq. (1), Specifically, our LC-MTJSR method uses the following criteria:

$$\min_{W} \frac{1}{2}\sum_{k=1}^{K}\left\|y^k - \sum_{j=1}^{J} X_j^k w_j^k\right\|_2^2 + \lambda \sum_{j=1}^{J}\left\|\sum_{k=1}^{K} D_j^k \Theta w_j^k\right\|_2 \tag{2}$$

whereΘ denotes the element-wise multiplication, λ is the regularization parameter and D_j^k is the locality adaptor that gives different freedom to different modalities of features X_j^k , which are proportional to its similarity to the input descriptor y^k. Specifically,

$$D_j^k = \left\|y^k - X_j^k\right\|_2^2 \tag{3}$$

D_j^k is the Euclidean distance between the training image X_j^k and the query data y^k. The principle of the locality constraint in regularization term may be explained by comparing Eq. (1) with Eq. (2). The MTJSR process may select quite different features for data reconstruction to favour sparsity, thus loses correlations between different feature space of training examples, but the explicit locality using in our LC-MTJSR ensures that similar test feature will have similar training feature for sparsity presentation. In practice, the LC-MTJSR in Eq. (2) is not sparse in the sense of l_0 norm. We just set those small coefficients to zero.

The recent research in image classification is focused on the large-scale image categorization. To efficiently apply our method into large-scale image categorization, we conventionally adopt the Stochastic gradient descent method(SGD) [23] during optimization to guarantee the formulated object(Eq. (2)) with fast convergence rate. SGD is a simple approach to find the local minima of a cost function whose evaluations are corrupted by noise, and is perhaps the most commonly used optimization procedure. To economize the computational cost at every iterations, SGD samples a subset training sample to optimize at every step. This is very effective in the case of large-scale machine learning problems. We update w_j^k in Eq. (2) according to the SGD method as follows:

$$\begin{cases} w_j^{k,t+1} = w_j^{k,t} - \eta\nabla_w \\ \nabla_w = -X^k y^k + X^k X^k w_j^{k,t} + \lambda\left\|y^k - X_j^k\right\|_2^2 \cdot w_j^{k,t} \end{cases} \tag{4}$$

Where η is the step size value.

Like the MTJSR method, we also extend our method using the kernel trick. The intuition of such a kernel trick is to use a non-linear function φ^k for each task k to map the training and test samples from the original space to another higher dimensional RKHS, where we have $\varphi^k(x_i)^T\varphi^k\left(x_j\right) = g^k\left(x_i, x_j\right)$ for some given kernel function g^k. In this new space, we can rewrite the function (4) as:

$$\begin{cases} w_j^{k,t+1} = w_j^{k,t} - \eta\nabla_w \\ \nabla_w = -h^k + G^k w_j^{k,t} + \lambda\left(P^k - 2h^k + G^k\right)\cdot w_{\mathrm{j}}^{\mathrm{k,t}} \end{cases} \tag{5}$$

Where $h^k = \varphi(X^k)\varphi(y^k)$, $G^k = \varphi(X^k)\varphi(X^k)$ and $P^k = \varphi(y^k)\varphi(y^k)$ are the kernel matrix associated with kth modality of feature.

When classifying test image, the test image can be reconstructed by using only the optimal coefficients $\widehat{w}_j^k$ associated with the jth class, and the kth modality y^k of a test image can be approximated as $\widehat{y}^k = X_j^k\widehat{w}_j^k$. The final decision is ruled in favor of the class with the lowest total reconstruction error accumulated over all the K tasks:

$$j^* = \arg\min_{j}\sum_{k=1}^{K}\left\|y^k - X_j^k\widehat{w}_j^k\right\|_2^2 \tag{6}$$

The details of this kernel-view extension of LC-MTJSR are given in Table1.

3. Experimental Results

To evaluate the effectiveness of our proposed method for object classification by feature combination, we apply it to several multi-class object categorization datasets, and compare the overall recognition performance of our proposed algorithms with the following methods: a) SRC [10],

MTJSR [14] and Some others multiple kernel learning methods from literatures [7], [9], [24].

Our datasets are chosen as follow:

Oxford flower17 dataset [25]: This dataset totally has 1360 images from 17 species of flowers, and 80 images in each class. The flowers are chosen from some common flowers in the UK. The images have large scale, pose and light variations, and also there are classes with large variations of images within the class and close similarity to other classes. The dataset has been randomly split the dataset into 3 different training, validation and test sets. The handcrafted descriptions include the HSV, HOG and SIFT, which are three matrices derived from color, shape and texture. The χ^2 distance matrices of these features along with a predefined training /validation/test split are publicly available on the dataset website.

Caltech-101 dataset [26] and Caletech-256 dataset [27]: The Caltech-101 dataset contains images of 101 categories of objects as well as a background class, and the Caltech-256 dataset holds 30,607 images in 256 categories, and presents much higher variability in object size, location, pose than the Caltech-101 dataset. In the Caltech-101, Most categories have about 50 images, but in the Caltech-256, each class contains at least 80 images. The experiment on the Caltech-256 dataset can evaluate the performance of large scale image categorization more efficiently than that on the Caltech-101 dataset. The classification is carried out on the basis of χ^2 distance matrices of Geometric blur [6] PHOW gray [4], PHOW color [4] and SSIM [5]. These features are extracted using the MKL code package from [28]. The data set is divided into a training set, a validation set and a test set. Two different sizes of training set are used to evaluate performance, which include 15 training images and 30 training images per class, and each validation set consists of 15 images per class, and the test set consists of the remaining images. We calculate the χ^2 distance matrices of these four features along with a training/validation/test split. On both datasets, Kernel matrices are computed as $\exp\left(-\chi^2(x,x)/\mu\right)$, where μ is set to be the mean value of the pairwise χ^2 distance matrices on the training set. The classification average accuracy over all classes is chosen as the final performance during the testing.

Before the experiment, the regularization parameter λ and step size value η should been set. The step size value η is the learning rate of the SGD method. Considering the convergence of the SGD method, we make the step size value η decreasing with the iteration number t, and set η as $\frac{1}{1+100t}$. The regularization parameter λ is used to balance the reconstruction error and locality constraint, and we optimize the best parameter value using cross-validation. The regularization parameter λ is chosen from the set $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, and the performance of the oxford flower17, the Caltech-101 and the Caltech-256 validation set is the best when the parameter λ is 0.01, 0.001 and 0.001, respectively.

During experiment of Oxford flower17 dataset, the accuracies of single feature kernel matrices of our method and the several other methods are shown in Fig. 2. The results show that the performance using the shape and SIFT features is better than that using the other features because the shapes of flower are the easiest to distinguished between the different categories. The performance of LC-MTJSR is better than that of SVM and the SRC only using the single features, and in the feature of color, shape, SIFTint and SIFTbdy, the performance of LC-MTJSR has more recognition accuracies than that of MTJSR. The performance of the combined feature are shown in Fig. 3, and the results

Table 1 The proposed LC-MTJSR algorithm.

LC-MTJSR Pseudo Code
Input: some subset training samples $X_i^k, i = 1, \ldots, n$, the regularization parameter λ and step size value η
Step1: Properly initialize $\widehat{w}^{k,0}$, set $t \leftarrow 0$ **Step2**: Randomly shuffle examples in the training set X_i^k, and calculate the kernel matrix h_i^k ,G_i^k and P_i^k **Step3**: For $i = 1, \ldots, n$, update the construct coefficients using: $\widehat{w}_i^{k,t+1} = \widehat{w}_i^{k,t} - \eta\nabla_{i,w}$, where $\nabla_{i,w} = -h_i^k + G_i^k\widehat{w}_i^{k,t} + \lambda\left(P_i^k - 2h_i^k + G_i^k\right)\cdot\widehat{w}_i^{k,t}$, $k = 1, \ldots, K$ **Step4**: $t = t + 1$ **Step5**: If the loss function in Equations 2 does not decrease, or the iteration number t is larger than a predefined threshold, then exit. Otherwise go to step 2
Output: $j^* = \arg\min_j \sum_{k=1}^{K}\left(-2h_j^k\widehat{w}_j^k + \left(\widehat{w}_j^k\right)^T G_j^k\left(\widehat{w}_j^k\right)\right)$

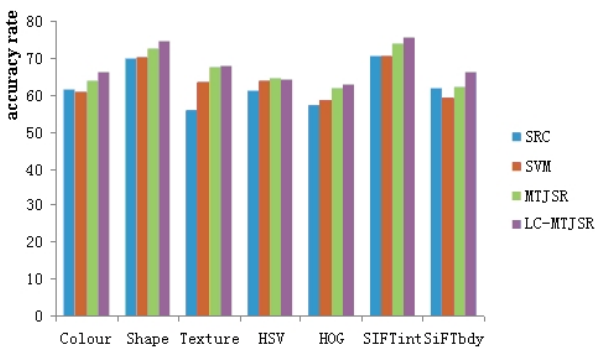

Fig. 2 The accuracies of single feature kernel matrices between our method and the several other methods.

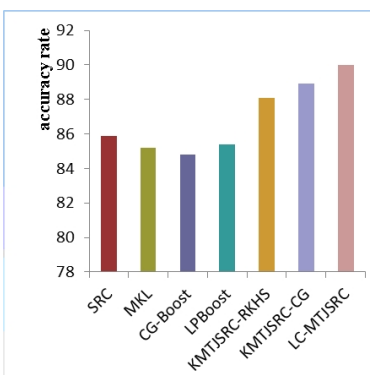

Fig. 3 The accuracies of combined feature kernel matrices between our method and the several other methods.

show that the classification performance using all feature has been dramatically improved than that using single feature, and our method is better than MKL [7], MTJSR [14], CG-Boost [9] and LPBoost [9] methods.

During the experiment of Caltech-101 category dataset, Table 2 lists the average accuracies of our LC-MTJSR methods along with the results from [14]. We observe that the average accuracies of LC-MTJSR are higher than those of MTJSR when testing GB and PHOW features, and LC-MTJSR can achieve the highest performance among all testing methods using all features. Table 3 shows the average accuracies between our LC-MTJSR methods and the other methods on the 256 category dataset using single feature kernel matrices. The experimental results show that the average accuracy of SVM method is the highest when testing GB feature, and the performance of LC-MTJSR is almost similar to that of SVM. When testing the PHOW gray and PHOW color features, The average accuracy of LC-MTJSR are 30.4% and 30.2%, respectively, which are better than that of MTJSR and SVM. When testing the SSIM feature, the performance of LC-MTJSR is similar to that of MTJSR. Table 4 shows the average accuracies of the feature combination using different training images. The average accuracy of LC-MTJSR is higher 0.8% than that of MTJSR when we use only 15 training images, and LC-MTJSR achieve the best performance. Moreover, the performance of LC-MTJSR is higher 2.7% than that of MTJSR, and the performance of LC-MTJSR is competitive with that of MKL using all features.

Table 2 The accuracy(%) performance on the Caltech-101 dataset using the single features and combined features when the number of training image is 15.

features	MKL	MTJSR	LC-MTJSR
GB	**62.6±1.2**	58.3±0.4	59.4±0.7
Phow-gray	63.9±0.8	65.0±0.7	**65.3±0.9**
Phow-color	54.5±0.6	56.1±0.5	**57.0±1.0**
SSIM	54.3±0.6	**61.8±0.6**	61.5±0.7
All feature	70.0±1.0	71.0±0.3	**71.7±0.9**

Table 3 The accuracy(%) performance on the Caltech-256 dataset using the single features when the number of training image is 15.

features	SRC	SVM	MTJSR	LC-MTJSR
GB	21.4±0.4	**29.4±0.5**	26.8±0.6	28.5±0.5
Phow-gray	20.4±0.5	26.3±0.7	29.4±0.5	**30.4±0.6**
Phow-color	19.8±0.4	27.7±0.4	28.9±0.4	**30.2±0.3**
SSIM	18.4±0.6	19.2±0.5	23.7±0.6	**23.8±0.5**

Table 4 The accuracy(%) performance on the Caltech-256 dataset using the combined features.

Methods	15 training images	30 training images
SRC	30.1±0.7	36.7±0.6
MKL	37.5±0.6	**43.9±0.7**
MTJSR	37.7±1.0	41.1±0.9
LC-MTJSR	**38.5±0.9**	43.8±1.0

4. Conclusion

In the large scale image categorization, multiple manually crafted features are extracted to represent the images. Our proposal, the locality-constrained multi-task joint sparse representation (LC-MTJSR) method, considers these descriptions as different tasks, and further introduces a constraint of local manifold stucture into the formulated objective. During the optimization of formulated objective, the stochastic gradient descent method (SGD) is used to guarantee fast convergence rate of optimization. Experiments were performed on several category datasets: Oxford flower17 dataset, Clatech-101 dataset and Caltech-256 dataset. When challenged with Oxford flower17 dataset, our LC-MTJSR achieved better performance than the MKL, MTJSR, CG-Boost and LPBoost methods; when challenged with Caltech-101 datset and Caltech-256 dataset, the performance of our methods is better than that of the MTJSR method, and is comparable with that of MKL method. Our LC-MTJSR method, like MTJSR method, imposed $\ell_{1,2}$ mixed-norm penalty on the reconstruction coefficients for its relative simplicity for optimization, but in some applications, some $\ell_{p,q}(p > 1, q > 1)$ norm maybe outperforms $\ell_{1,2}$ norm due to non-sparsity. The selection of best norm regularization is a very interesting research topic in the multi-task joint sparse representation framework that merits our future study. Moreover, how to efficiently apply some semantic context, such as the hierarchical structure of training sample, to the multi-task joint sparse representation framework is another interesting research topic.

Acknowledgments

This work is supported in part by the Guangzhou Science and Technology Plan Project(2012J2200010), the Fundamental Research Funds for the Central Universities of SCUT, the National Science Foundation of China (NSFC)(grant no.61201348,61202292) and the National Science and Technology Support Plan (2013BAH65F01-2013BAH65F04).

References

[1] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.
[2] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," Proc. IEEE Comput. Vis. Pattern Recognit., pp.887–893, 2005.
[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories," Proc. IEEE Comput. Vis. Pattern Recognit., pp.2169–2178, 2006.
[4] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," Proc. CIVR, 2007.
[5] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," Proc. CVPR, 2007.
[6] A.C. Berg, T.L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," Proc. CVPR, 2005.

[7] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple Kernels for object detection," Proc. International Conference on Computer Vision, 2009.
[8] G. Lihua and J. Lianwen, "Laplacian Support vector machines with multi-kernel learning," IEICE Trans. Inf. & Syst., vol.E94-D, no.2, pp.379–383, Feb. 2011.
[9] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," International Conference on Computer Vision, 2009.
[10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.2, pp.210–226, 2009.
[11] G. Obozinski, B. Taskar, and M. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," J. Statistics and Computing, pp.1–22, 2009.
[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," J. Royal Statistical Society, Series B, vol.68, no.1, pp.49–67, 2006.
[13] J. Zhang, "A probabilistic framework for multi-task learning," Technical report, CMU-LTI-06-006, 2006.
[14] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," CVPR 2010.
[15] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua, "Multi-label visual classification with label exclusive context," ICCV 2011, pp.834–841, 2011.
[16] X. Liu, X.-T. Yuan, S. Yan, and H. Jin, "Multi-class semi-supervised SVMs with positiveness exclusive regularization," ICCV 2011, pp.1435–1442, 2011.
[17] D. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation," IEEE Trans. Image Process., vol.21, no.1 pp.316–326, 2012.
[18] Y.-W. Chao, "Locality-constrained group sparse representation for robust face recognition," 2011 18th IEEE International Conference on Image Processing (ICIP), pp.761–764, 2011.
[19] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," Int. J. Comput. Vis., vol.101, no.2, pp.367–383, Jan. 2013.
[20] H. Zhang, N.M. Nasrabadi, Y. Zhang, and T.S. Huang, "Multi-observation visual recognition via joint dynamic sparse representation Computer Vision (ICCV)," 2011 IEEE International Conference on, pp.595–602, Nov. 2011.
[21] S. Gao, "Sparse representation with kernels," IEEE Trans. Image Process., vol.22, no.2, pp.423–434, 2013.
[22] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," NIPS'09, 2009.
[23] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," Advances in Neural Information Processing Systems, vol.20, pp.161–168, 2008.
[24] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," IEEE International Conference on Computer Vision, 2007.
[25] http://www.robots.ox.ac.uk/ṽgg/data/flowers/17/index.html
[26] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[27] http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[28] http://www.robots.ox.ac.uk/ṽgg/software/MKL/