Visual Classification With Multitask Joint Sparse Representation

Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan, *Senior Member, IEEE*

***Abstract*— We address the problem of visual classification with multiple features and/or multiple instances. Motivated by the recent success of multitask joint covariate selection, we formulate this problem as a multitask joint sparse representation model to combine the strength of multiple features and/or instances for recognition. A joint sparsity-inducing norm is utilized to enforce class-level joint sparsity patterns among the multiple representation vectors. The proposed model can be efficiently optimized by a proximal gradient method. Furthermore, we extend our method to the setup where features are described in kernel matrices. We then investigate into two applications of our method to visual classification: 1) fusing multiple kernel features for object categorization and 2) robust face recognition in video with an ensemble of query images. Extensive experiments on challenging real-world data sets demonstrate that the proposed method is competitive to the state-of-the-art methods in respective applications.**

***Index Terms*— Feature fusion, multitask learning, sparse representation, visual classification.**

I. Introduction

THE PROBLEM of recovering sparse linear representation of a query datum with respect to a dictionary of reference data has recently received wide interests in signal processing [1], [2] and computer vision [3]–[7]. By taking the training datum as *observations* of covariate and the query datum as *response*, the sparse linear representation problem can be cast into a problem of sparse covariate selection via linear regression model. It has been discovered in neural science [8] that the human vision system seeks a sparse representation for the incoming image using a few words in a feature vocabulary. Olshausen *et al.* [9] introduced a Bayesian framework to simulate the sparse coding mechanism of human vision system. For face recognition, Wright *et al.* [5] demonstrated that the ℓ_1-norm regularized reconstruction error minimization is able to enhance system robustness against occlusion and illustration changes. For image super-resolution, Yang *et al.* [7] proposed to jointly learn two dictionaries for low resolution and high resolution images respectively, and the super-resolution is accomplished by transferring the representation coefficients from the former to the latter. For semi-supervised learning, Yan and Wang [6] proposed the ℓ_1-*graph* in which the graph adjacency structure and the graph weights are derived simultaneously via datum-wise linear sparse representation.

In machine learning and computer vision, when the tasks to be learned share some latent factors, it may be advantageous to take into account these cross-task relations. For instance, when the number of samples for learning is small, transferring some knowledge from one task to another can be advantageous in term of generalization performance. This new learning scheme has inspired the study of multi-task learning (MTL). A large body of works have provided evidence on the benefit of such a framework in theory [10]– [12] and in practice [13]– [15].

In this paper, motivated by the success of *multi-task joint sparse linear regression* [15]– [17], we investigate the problem of *multi-task joint sparse representation and classification* (MTJSRC) and its applications to visual recognition. By using the term "multi-task", we mean that there are more than one linear representation models which are simultaneously estimated with proper regularization on parameters across all the models. For example, in the setting of object recognition, it is natural to get K different linear representation models from K different visual features (e.g., color, shape and texture). For each representation model, as argued by Wright *et al.* [5] in face recognition that if sufficient training samples are available from each class, it will be possible to sparsely represent the test sample as a linear combination of just those training samples from the same class. We generalize this argument to the setup of multi-task learning. The basic idea is to find out across related representation models a very few common classes of training samples that are most correlated to the query sample. Here the constraint of joint sparsity across different tasks is able to provide additional useful information to the classification problem because different tasks may favor different sparse representation coefficients, yet the joint sparsity may enforce the robustness in coefficient estimation. The joint sparsity is enforced by imposing the $\ell_{1,2}$-norm penalty on the representation coefficients, as has been investigated by Argyriou *et al.* [13] and Obozinski *et al.* [16]. The $\ell_{1,2}$ regularization term can be taken as an extension of the sparsity-inducing ℓ_1-norm from single task learning to MTL. As an alternative joint sparsity-inducing norm, the $\ell_{1,\infty}$-norm has been widely applied as penalties in signal processing [18]

Manuscript received July 5, 2011; revised April 19, 2012; accepted May 16, 2012. Date of publication June 18, 2012; date of current version September 13, 2012. This work was supported in part by the NExT Research Center, funded by MDA, Singapore, under the Research Grant WBS:R-252-300-001-490. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kenneth K. M. Lam.

X.-T. Yuan is with the Department of Statistics, Rutgers University, Newark, NJ 08854 USA (e-mail: xyuan@stat.rutgers.edu).

X. Liu is with the Department of Statistics, University of California, Los Angeles, CA 90095 USA (e-mail: lxb@ucla.edu).

S. Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, 117583 Singapore (e-mail: eleyans@nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2205006

and machine learning [15], [17]. In this paper, we focus on the usage of $\ell_{1,2}$-norm penalty to enforce joint sparsity across representation tasks, while our method can be easily modified to handle the case with $\ell_{1,\infty}$-norm penalty. The formulated objective is composed of a squared reconstruction error term and a non-smooth $\ell_{1,2}$-norm regularization term. Conventionally, we resort to the accelerated proximal gradient (APG) method [19], [20] for optimization with strong convergence guarantee. After the representation coefficients are estimated, the classification is ruled in favor of the class that has the lowest total reconstruction error accumulated from all the tasks. Taking the classification of flower images as an example, Figure 1 depicts the working mechanism of MTJSRC.

In many visual applications, the hand-crafted descriptors of images are encoded in the form of kernel (or similarity) matrices [21]–[24]. In order to combine multiple complementary feature kernels for visual recognition, we present two kernel-view extensions of MTJSRC. The first extension addresses the situation where the feature space is a Reproducing Kernel Hilbert Space (RKHS). The second extension takes the columns of a kernel matrix as extracted feature vectors and directly applies the linear version of our approach in this new feature space. The APG method is utilized to optimize both kernel extensions.

We have applied MTJSRC and its kernelised variants to two concrete visual classification problems.

1) *Multiple-kernel based object categorization*. We apply the kernelised MTJSRC to fuse the discriminative power of different complementary descriptors for object categorization. Our experimental results on some benchmark data sets show that MTJSRC is competitive to several state-of-the-art multiple kernel learning methods for object recognition [25]–[27].
2) *Face recognition for dynamic videos*. In this application, we assume that each video sequence contains only one single subject. Given an input video, we first employ a face detector [28] to localize faces within each frame. Then, while recognizing the face detected in current frame, the faces detected within the previous L frames are also collected to produce the ensemble of samples belonging to the same subject, and the subject identify could be determined through the MTJSRC. We move the window of size L along with the input sequence to process each input frame. The main purpose of this application is to show the natural capability of MTJSRC for combining multiple test instances for face recognition. Experimental results on real-world videos validates the advantages of MTJSRC over existing methods.

A. Related Work

Sparse linear models seek to predict an output by linearly combining a small subset of the features describing the data. To simultaneously address variable selection and model estimation, ℓ_1-norm regularization has become a popular tool, which benefits both from efficient algorithms [20], [29], [30] and well-developed theory for generalization properties and variable selection consistency [31], [32]. When the variables

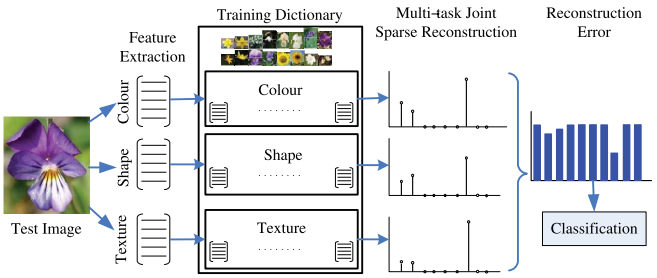

Fig. 1. Scheme illustration of visual classification with multitask joint sparse representation and classification. Given a query image, multiple modalities of features are extracted. Each feature is then represented as a linear combination of the corresponding training features in a joint sparse way across all of the features. Finally, the classification decision is achieved according to the overall reconstruction error of the individual class.

are regularized by the ℓ_1-norm, each variable is estimated independently, regardless of its position in the input feature vector. In many practical situations, however, the estimation can benefit from some type of prior knowledge on the relationships and structures between the variables. The group sparsity is a suitable method in these situations and has been an active topic in both machine learning [33]–[36] and signal processing [37]–[39]. Recently, sparsity and group sparsity have been applied to the task of recovering the sparse representation of a datum with respect to a set of bases or a dictionary in image processing and computer vision [40]–[42], [5]. Wright *et al.* [5] exploited the sparse representation-based classification (SRC) method for robust face recognition. They assumed that the training samples of a particular class approximately form a linear basis set for any test sample belonging to this class. The ℓ_1-norm minimization was utilized to select the representative training samples from the entire training set. To promote the representation of the test sample in terms of all the atoms from the correct group, Majumdar and Ward [42] proposed two alternative regularization methods, Elastic Net [43] and group Lasso [34], to improve SRC. Both these regularization methods favor the selection of multiple correlated training samples to represent the test sample.

Multi-task learning has received a lot of research interests in machine learning. The idea behind this paradigm is that, when the tasks to be learned are similar enough or are related in some sense, it may be advantageous to consider these relations between tasks in the model. Several works have experimentally demonstrated the benefit of such a framework [10], [44]. In general, MTL can be addressed through a regularization framework [44]. For example, the joint sparsity regularization favors learning a common subset of features for all tasks [13], [16], while the exclusive sparsity regularization [45] encourages exclusive feature selection among tasks. The proposed MTJSRC also falls in the regularized MTL framework. Methodologically, MTJSRC is motivated by the recent advance in sparse learning called *multi-task joint covariate selection* (MTJCS) [16]. In this approach, by penalizing the sum of ℓ_2 norms of the blocks of coefficients associated with each covariate group across different classification problems, similar sparsity patterns in all models are encouraged. From the standpoint of linear regression, MTJCS can be

regarded as a combinational model of group Lasso [34] and multi-task Lasso [15], [17]. We introduce this appealing sparse learning model to computer vision as a joint sparse visual representation method.

Combining multiple discriminative features for object recognition is a recent trend in class-level object recognition and image classification. It is clear that among several types of feature descriptors, not every one has the same discriminative power for all classes. Therefore it is widely accepted that, instead of using a single modality of feature, it is better to adaptively combine a set of diverse and complementary modalities of features in order to discriminate each class best from all other classes. One popular method for feature combination in computer vision is Multiple Kernel Learning (MKL) which linearly combines multiple similarity functions such that the combined one yields improved classification performance [46], [27]. Recently, several support vector machines (SVMs) ensemble methods inspired by linear programming Boosting have also been proposed for multi-kernel object classification [25]. In contrast to this family of work, our method combines multiple features for visual classification in a regularized MTL framework.

B. Paper Organization

We outline the remainder of this paper as follows. In Section II we briefly review the sparse representation and classification methods in object recognition. In Section III we present the proposed multi-task joint sparse representation and classification method. The kernel-view extensions of our method are described in Section IV. The applications of our method to object categorization and face recognition are studied in Section V. We conclude this paper in Section VI.

II. Sparse Representation

Sparsity has been a key factor of intensive research over the last decade. This line of research witnessed the development of nice theoretical frameworks [1], [47] and the emergence of many efficient algorithms [29], [48]. In computer vision, it has been demonstrated in [41], [5] that the sparse representation model is discriminative and particularly useful for building robust multi-class visual classification systems. Let us consider a J class visual classification problem. Let matrix $X \in \mathbb{R}^{d\times n}$ be a stack of n columns of training image feature vectors of dimension d. Denote $X_j \in \mathbb{R}^{d\times n_j}$ as the n_j ($\sum_{j=1}^{J} n_j = n$) columns of X labeled as class $j \in \{1,\ldots,J\}$. Given a testing image feature $y \in \mathbb{R}^d$, the sparse linear representation model seeks to solve the following optimization problem:

$$\hat{w} = \arg\min_{w} \|w\|_0, \quad s.t. \quad \|y - Xw\| \le \epsilon \tag{1}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm which counts the number of non-zero entries in a vector and ϵ is the noise level parameter. Unfortunately, it is known that problem (1) is NP-hard in general case, and thus is intractable. Recent results [1], [49] show that under mild assumptions on X and given that the solution is sparse enough, the sparse representation can be recovered by solving the following ℓ_1-norm minimization:

$$\hat{w} = \arg\min_{w} \|w\|_1, \quad s.t. \quad \|y - Xw\| \le \epsilon. \tag{2}$$

This optimization problem is convex and can be efficiently optimized via several well implemented toolboxes, e.g., NESTA [50]. Given the optimal solution $\hat{w}$, the class label of y is decided based on the following criterion of minimum reconstruction error:

$$\hat{j} = \arg\min_{j\in\{1,\ldots,J\}} \|y - X_j \hat{w}_j\| \tag{3}$$

where $\hat{w}_j$ is the components of $\hat{w}$ restricted on class j. The model (2) together with the decision rule (3) is known as sparse representation-based classification (SRC) in the study of face recognition [5].

III. Multitask Joint Sparse Representation and Classification

The SRC model described in the previous section was originally developed for single feature based visual recognition. In this section, we will generalize SRC to the setup of multiple features and multiple instances based recognition. Suppose we have a training set with J classes in which each sample has K different modalities of features (e.g., color, shape and texture). For each modality index $k = 1,\ldots,K$, denote $X^k \in \mathbb{R}^{d_k\times n}$ the training feature matrix. Let $X_j^k \in \mathbb{R}^{d_k\times n_j}$ be the n_j columns of X^k associated with the jth class. Also we suppose that a testing sample is given by $y = \{y^{kl} \in \mathbb{R}^{d_k}, k = 1,\ldots,K, l = 1,\ldots,L\}$ as an ensemble of L different instances (e.g., multiple views of a human face), each of which is represented by the same K modalities as training images. Let us consider the following $K \times L$ linear representation models:

$$y^{kl} = X^k W^{kl} + \varepsilon^{kl}, \qquad k = 1,\ldots,K, \qquad l = 1,\ldots,L, \tag{4}$$

where $W^{kl} \in \mathbb{R}^n$ is the coefficient vector for y^{kl} and $\varepsilon^{kl} \in \mathbb{R}^{d_k}$ is the residual term which is assume to be i.i.d. Gaussian noise. $\{W^{kl}\}$ can be estimated by fitting the following least squared regression (LSR) model:

$$\min_{W} \left\{ f(W) := \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{L} \left\| y^{kl} - X^k W^{kl} \right\|^2 \right\}, \tag{5}$$

where $W \in \mathbb{R}^{n\times KL}$ is the matrix stacked by $K \times L$ columns of coefficient vectors $\{W^{kl}\}$. To avoid singularity of linear systems, an additional regularization term $\lambda\|W\|_F^2$ is typically imposed in (5). From the viewpoint of multi-task learning, problem (5) is a multi-task regression model with $K \times L$ independent LSR.

A. Imposing Class-Level Joint Sparsity Regularization

This formulation of problem (5), however, does not take into account the relationship among the features and instances since the minimization is independently performed for individual features and instances. To combine the strength of multiple features/instances for recognition, we improve the independent learning model (5) to a joint learning model by imposing a class-level sparsity-inducing term, which is derived in the sequel. The intuition of designing such a term is: for the purpose of multi-class classification, it can be useful to

jointly selected a few common classes of training samples to represent a testing image over each feature and each instance. Therefore, the desired sparse representation vectors for the multiple features/instances should share certain class-level sparsity patterns.

For each test image feature y^{kl}, let us rewrite the representation vector as $W^{kl} = [(W_1^{kl})^T, \ldots, (W_J^{kl})^T]^T$ in which $W_j^{kl} \in \mathbb{R}^{n_j}$ consists the components of W^{kl} restricted on class j. Denote $W_j = [W_j^{11}, \ldots, W_j^{KL}] \in \mathbb{R}^{n_j \times KL}$ the representation coefficients associated with class j across different features and instances. To combine the strength of all the atoms within class j, we apply ℓ_2-norm over W_j. To promote sparsity to allow a small number of classes to be involved during the joint sparsity representation, we apply ℓ_0-norm across the ℓ_2-norm of the W_j. Thus, we arrive at the following class-level joint sparsity-inducing term:

$$\Omega(W) := \|[\|W_1\|_F, \ldots, \|W_J\|_F]\|_0 . \tag{6}$$

To recover the sparse representation coefficient matrix W with the joint sparsity regularization for the multiple observations, we propose to solve the following multi-task joint sparse representation model:

$$\hat{W} = \arg\min_W \Omega(W), \quad s.t. \quad f(W) \le \epsilon,$$

or equivalently

$$\hat{W} = \arg\min_W f(W) + \lambda\Omega(W), \tag{7}$$

where the expressions of $f(W)$ and $\Omega(W)$ are given in (5) and (6), respectively, and λ is a trade-off parameter balancing the effects of these two terms. The problem (7), however, is NP-hard due to the ℓ_0-norm used in $\Omega(W)$. To make the problem tractable, we apply the following convex relaxation to problem (7):

$$\hat{W} = \arg\min_W f(W) + \lambda P(W), \tag{8}$$

where $P(W)$ is the following $\ell_{1,2}$-norm

$$P(W) := \sum_{i=1}^{J} \|W_j\|_F .$$

As is well known [34] that when used as a regularization term, $P(W)$ will encourage $\|W_j\|_F = 0$, and thus $W_j = 0$. Therefore the convex formulation (8) still encourages the test features $\{y^{kl}\}$ to be sparsely reconstructed by the most parsimonious and representative classes in the training set. The problem (8) is known as the multi-task joint covariate selection model in sparse learning [16].

B. Classification Rule

Given the optimal coefficient matrix $\hat{W}$, one can approximate feature y^{kl} as $\hat{y}^{kl} = X_j^k \hat{W}_j^{kl}$. The decision is ruled in favor of the class with the lowest reconstruction error accumulated over all the $K \times L$ tasks:

$$\hat{j} = \arg\min_{j \in \{1,\ldots,J\}} \sum_{k=1}^{K} \sum_{l=1}^{L} \|y^{kl} - X_j^k \hat{W}_j^{kl}\|^2. \tag{9}$$

We call the model (8) together with the decision rule (9) as MTJSRC, namely multi-task joint sparse representation and classification. Particularly, when $L = K = 1$, MTJSRC reduces to a regularization form of SRC [5].

C. Optimization

We resort to the widely applied Accelerated Proximal Gradient (APG) method [19], [20], [51] to optimize problem (8). In the seminal work [19], Nesterov considered the general minimization problem of which the objective composes a smooth convex term and a non-smooth convex term, and proposed the APG algorithm which achieves $O(1/t^2)$ rate of convergence by simply using first order information. The success of APG method largely depends on the structure of the non-smooth part, e.g., there exists a closed-form minimizer of the sum of the non-smooth part with a quadratic auxiliary function. Tseng [20] proposed a nearly unified treatment of the existing APG methods. The application of APG algorithm to group/multi-task joint sparse learning has been addressed in [51]–[53]. In this paper, we have implemented an APG optimization procedure similar to that of [51]. Algorithm 1 summaries the details of the optimization and classification. The scheme alternately updates a weight matrix sequence $\{W^{(t)}\}_{t\ge1}$ and an aggregation matrix sequence $\{V^{(t)}\}_{t\ge1}$. Each iteration consists of the following two steps.

1) **The generalized gradient mapping step to update matrix** $W^{(t+1)}$. Given the current matrix $V^{(t)}$, we calculate

$$W^{(t+1)} = \arg\min_W f(V^{(t)}) + \langle \nabla f(V^{(t)}), W - V^{(t)} \rangle + \frac{1}{2\eta}\|W - V^{(t)}\|^2 + \lambda P(W), \tag{10}$$

where η is a proper step-size parameter which should be no less than the inverse number of the Lipschitz constant of ∇f. For the $\ell_{1,2}$-norm regularization $P(W)$, it is shown in [54] that the solution of problem (10) is given in closed-form by:

$$W^{(t+1/2)} = V^{(t)} - \eta \nabla f(V^{(t)}),$$
$$W_j^{(t+1)} = \left[1 - \frac{\lambda\eta}{\|[W^{(t+1/2)}]_j\|}\right]_+ W_j^{(t+1/2)}, \tag{11}$$

where we define $[\cdot]_+ = \max\{0, \cdot\}$.

2) **The aggregation forward step to update** $V^{(t+1)}$. We then construct a linear combination of $W^{(t)}$, $W^{(t+1)}$ to update $V^{(t+1)}$ as follows:

$$V^{(t+1)} = W^{(t+1)} + \frac{\alpha_{t+1}(1 - \alpha_t)}{\alpha_t}(W^{(t+1)} - W^{(t)}). \tag{12}$$

Here the sequence $\{\alpha_t\}_{t\ge1}$ satisfies

$$\frac{1 - \alpha_t}{\alpha_t^2} \le \frac{1}{\alpha_{t-1}^2}. \tag{13}$$

We conventionally set $\alpha_t = 2/(t + 2)$ [20] in our implementation.

Although proved to be convergent to global minimum with the optimal rate $O(1/t^2)$, APG does not guarantee to monotonously decrease the objective value [29]. In our implementation, we found that running Algorithm 1 until convergent is not necessary for the best recognition performance. Indeed, satisfying recognition accuracy can be obtained within a few hundred times of iteration. This can be partially explained by the fact that the objective of MTJSRC is to minimize reconstruction error of a testing image. This is in contrast to those classifier training based methods such as SVMs and logistic regression which directly address empirical losses on training data.

D. Computational Complexity

On computational complexity of MTJSRC, we point out that at each iterate of Algorithm 1, the dominant computational cost comes from the calculation of gradient (15) in step 3. The costs of the two terms in (15) are typically $O(KLnd_k)$ and $O(2KLnd_k)$ floating-point operations (*flops*), respectively. Since X^k and y^{kl} are pre-fixed, the first term $-(X^k)^T y^{kl}$ in (15) can be pre-computed. Let T be the average number of iterations for the running of Algorithm 1. The total flops for gradient estimation in (15) is typically $O(KLnd_k + 2TKLnd_k)$. The computational overload in the steps $4 \sim 6$ and the step 8 are negligible comparing to that of step 3.

IV. Kernel-View Extensions

So far, MTJSRC is developed for sparse representation with feature vectors. In many visual recognition problems, however, the descriptors are encoded as similarity or kernel matrices without raw features available. For the purpose of combining multiple feature kernels, we further present two kernel extensions of MTJSRC. The first extension aims to encode the features in a Reproducing Kernel Hilbert Space (RKHS), while the second extension is simply derived based on column generation strategy.

A. Joint Sparse Representation in RKHS

Let us consider the setup where the linear representation model is sparse in a RKHS. The intuition of such a kernel trick is to use a non-linear function ϕ^k for each feature k to map the training and test samples from the original space to another higher dimensional RKHS in which we have $\phi^k(x_i)^T\phi^k(x_j) = g^k(x_i, x_j)$ for some given kernel function g^k. In this new space, we can write the problem (8) as:

$$\hat{W} = \arg\min_W \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\phi^k(y^{kl}) - \sum_{j=1}^{J}\phi^k(X_j^k)W_j^{kl}\right\|^2 + \lambda P(W), \tag{14}$$

where $\phi^k(X_j^k) = [\phi^k(X_{j,1}^k), \ldots, \phi^k(X_{j,n_j}^k)]$. Notice that the gradient mapping step (10) only involves inner product of features, and thus can be straightforwardly extended to solve problem (14). Let $G^k = \phi^k(X^k)^T\phi(X^k)$ with $\phi^k(X^k) = [\phi^k(X_1^k), \ldots, \phi^k(X_J^k)]$ be the training kernel matrix associated with feature k, and $h^{kl} = \phi^k(X^k)^T\phi^k(y^{kl})$ be the test kernel vector associated with feature k and instance l, we have

Algorithm 1: MTJSRC Algorithm

Inputs : Reference image feature matrices $\{X^k \mid k = 1, \ldots, K\}$, an ensemble of query image features $\{y^{kl} \mid k = 1, \ldots, K, l = 1, \ldots, L\}$, the regularization parameter $\lambda > 0$, and the step-size parameter $\eta > 0$.
Output: $W^{(t)}$, $\hat{j}$.
Initialization: Set $V^{(0)} = 0$, $\alpha_0 = 1$ and $t = 0$.
repeat
  // The generalized gradient mapping step.
  Calculate $W^{(t+1/2)} = V^{(t)} - \eta\nabla f(V^{(t)})$, in which $\nabla f(V^{(t)})$ is given by

$$[\nabla f(V^{(t)})]^{kl} = -(X^k)' y^{kl} + (X^k)' X^k [V^{(t)}]^{kl}, \tag{15}$$

  $l = 1, \ldots, L, k = 1, \ldots, K$.
  Calculate $W^{(t+1)}$ as

$$W_j^{(t+1)} = \left[1 - \frac{\lambda\eta}{\|[W^{(t+1/2)}]_j\|}\right]_+ W_j^{(t+1/2)}, \quad j = 1, \ldots, J.$$

  // The aggregation step.
  Set $\alpha_{t+1} = \frac{2}{t+3}$. Update $V^{(t+1)} = W^{(t+1)} + \frac{(1-\alpha_t)\alpha_{t+1}}{\alpha_t}(W^{(t+1)} - W^{(t)})$.
  Set $t \leftarrow t + 1$.
until *Converges*;
// Make classification decision.
Calculate $\hat{j} = \arg\min_{j\in\{1,\ldots,J\}} \sum_{k=1}^{K}\sum_{l=1}^{L} \|y^{kl} - X_j^k[W^{(t)}]_j^{kl}\|^2$.

that $[\nabla f(V^{(t)})]^{kl} = -h^{kl} + G^k[V^{(t)}]^{kl}$ in (11). When the optimal reconstruction coefficient matrix $\hat{W}$ is estimated, the classification decision is made by

$$\begin{aligned}\hat{j} &= \arg\min_{j\in\{1,\ldots,J\}} \sum_{k=1}^{K}\sum_{l=1}^{L}\left\|\phi^k(y^{kl}) - \phi^k(X_j^k)\hat{W}_j^{kl}\right\|^2 \\ &= \arg\min_{j\in\{1,\ldots,J\}} \sum_{k=1}^{K}\sum_{l=1}^{L} -2h_j^{kl}\hat{W}_j^{kl} + (\hat{W}_j^{kl})^T G_j^k \hat{W}_j^{kl},\end{aligned}$$

where $h_j^{kl} = \phi^k(y^{kl})^T\phi^k(X_j^k)$ indicates the elements of h^{kl} associated with class j, and $G_j^k = \phi^k(X_j^k)^T\phi^k(X_j^k)$ is the block diagonal of G^k associated with class j. The details of such a kernel-view extension of MTJSRC, called as KMTJSRC-RKHS, are described in Algorithm 2. The complexity analysis of Algorithm 2 is analog to that of Algorithm 1 as stated in Section III-D.

B. Column Generation

One simple way to make use of the available kernel matrices is to directly take vector h^{kl} and the columns of each kernel matrix G^k as features. In essence, this idea is similar to a simplified column generation strategy for CG-Boost in multiple kernel learning [25]. In this way, problem (8) can be written as:

$$\hat{W} = \arg\min_W \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left\|h^{kl} - \sum_{j=1}^{J} G_j^k W_j^{kl}\right\|^2 + \lambda P(W), \tag{16}$$

to which Algorithm 1 can be directly applied. This kernel extension is referred to as KMTJSRC-CG throughout the rest of this paper.

Algorithm 2: KMTJSRC-RKHS Algorithm

Inputs : Reference kernel image feature matrices $\{G^k \mid k = 1, \ldots, K\}$, an ensemble of query kernel image features $\{h^{kl} \mid k = 1, \ldots, K, l = 1, \ldots, L\}$, the regularization parameter $\lambda > 0$, and the step-size parameter $\eta > 0$.

Output: $W^{(t)}$, $\hat{j}$.

Initialization: *Set* $V^{(0)} = 0$, $\alpha_0 = 1$ *and* $t = 0$.

repeat

// The generalized gradient mapping step.

Calculate $W^{(t+1/2)}$ as

$$[W^{(t+1/2)}]^{kl} = [V^{(t)}]^{kl} - \eta\left(-h^{kl} + G^k[V^{(t)}]^{kl}\right),$$

$l = 1, \ldots, L, k = 1, \ldots, K$.

Calculate $W^{(t+1)}$ as

$$W_j^{(t+1)} = \left[1 - \frac{\lambda\eta}{\|[W^{(t+1/2)}]_j\|_2}\right]_+ W_j^{(t+1/2)}, \quad j = 1, \ldots, J.$$

// The aggregation step.

Set $\alpha_{t+1} = \frac{2}{t+3}$;

Update $V^{(t+1)} = W^{(t+1)} + \frac{(1-\alpha_t)\alpha_{t+1}}{\alpha_t}(W^{(t+1)} - W^{(t)})$;

Set $t \leftarrow t + 1$;

until *Converges*;

// Make classification decision.

Calculate

$\hat{j} = \arg\min_j \sum_{k=1}^{K}\sum_{l=1}^{L}\left(-2h_j^{kl}[W^{(t)}]_j^{kl} + ([W^{(t)}]_j^{kl})^T G_j^k \hat{W}_j^{kl}\right)$.

V. Experiments

We have conducted several groups of experiments to evaluate the effectiveness of MTJSRC and its kernel-view extensions for visual classification. As an example, we demonstrate the feature combination capability of MTJSRC in a face recognition experiment. We then apply MTJSRC to multi-class object categorization and video based face recognition, and systematically evaluate the performances on several benchmark data sets. Our algorithms were implemented in Matlab 7.6/Windows Vista 64, and run on a personal desktop equipped with 2 Intel quadcore 3.0 GHz CPU and 32GB memory.

A. Illustrating Example on Feature Combination

We utilize the Extended Yale Face Database B[1] to demonstrate the feature combination capability of MTJSRC. It contains 16 128 images of 38 human subjects under 9 poses and 64 illumination conditions. For each individual, we use 64 near frontal cropped face images and resize them to be 32×32 pixels with 256 gray-levels per pixels. The features (pixel values) are then stacked to form a 1024-dimension descriptor. In this experiment, we utilize the gray-level and the Local Binary Patterns (LBP) [55] features which are widely used in face recognition. Therefore $K = 2$ for this feature combination test. Intuitively, these two features are complementary in terms of appearance and robustness to illumination variations. We use $L = 1$ instance for each probe face. Therefore we have $K \times L = 2 \times 1$ sparse representation tasks. For each subject j, we randomly select $n_j \in \{5, 10, 20, 30, 40, 50\}$ images for training (as reference images), and the remaining for testing. The reported mean and standard variance of recognition accuracy are estimated over 50 random splits.

[1]Available at http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html.

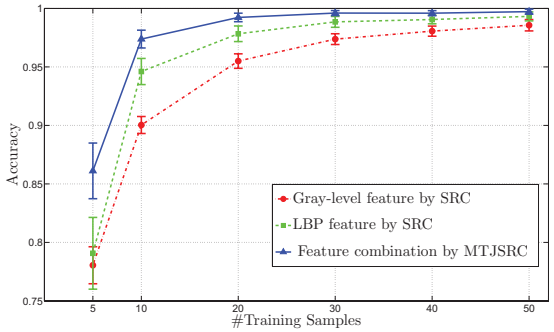

Fig. 2. Accuracy curves on the Extended Yale B set. The MTJSRC model consistently outperforms each single task SRC model.

Figure 2 shows the accuracy curves of individual features and their combination. From this figure we can see that MTJSRC well combines features to improve the performance. Several example classifications (with $n_j = 40$) are provided in Figure 3(a). As expected, the failing cases from the gray-level feature are mainly due to poor illumination condition (see the first five columns in Figure 3(a). In contrast, the LBP feature is robust to illumination variations and thus can classify these five samples correctly. The LBP still fails, however, in some other samples (see the last two columns in Figure 3(a) which the gray-level feature recognizes correctly. Combination of the two complementary features can fuse their benefits and achieve improved classification performance. In detail, Figure 3(b) and 3(c) show the sparse reconstruction coefficients and residuals for the test images of the 1st and 7th column in Figure 3(a).

B. Multiclass Object Categorization

We have applied the kernel extensions of MTJSRC to multi-class object classification on two Oxford flower data sets and the Caltech101 data set. In this group of experiments, we use $L = 1$ instance for each query image and simplify the notation W^{k1} as W^k.

1) Baselines: We compare our algorithms with the following methods.

1) Feature combination based on nearest subspace (NS). Here the column generation strategy is applied to handle kernel matrices, and the coefficients $\hat{W}_j^k$ are independently learnt through least squared regression for each feature k and each class j.
2) Feature combination based on independent SRC. This method can be taken as a simplification of our method without enforcing the joint sparsity across tasks. The column generation strategy is applied, and the coefficients $\hat{W}^k$ are independently learned by SRC.
3) Some representative multiple kernel learning methods in the study of object recognition [25], [26], [27].

Meanwhile, for single features, the kernel-views of our method reduce to the kernelizations of SRC, of which the performances shall be compared with NS and SVM.

2) Oxford Flowers Data Sets: In this subsection, we present results on two Oxford flower data sets with 17 classes [56] and 102 classes [26] respectively. For both data sets, the background of each image is removed using segmentation so

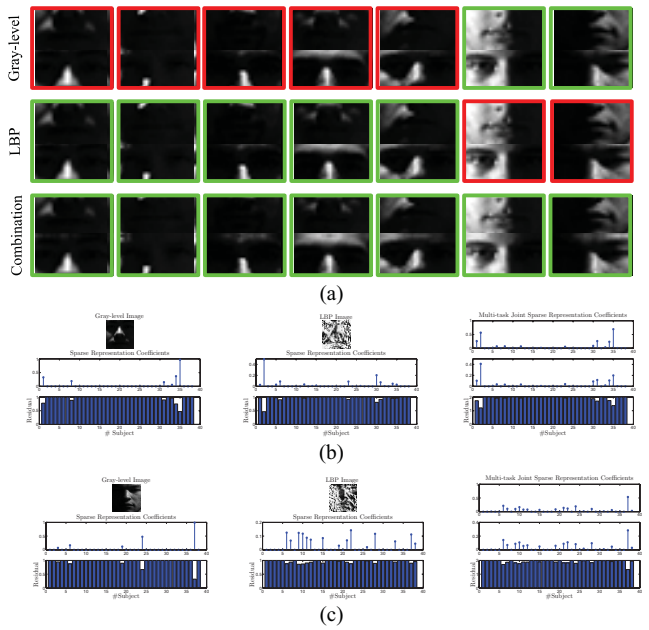

Fig. 3. (a) Example classifications. Each column shows the classifications of a testing image by single features and their combination. (b) and (c) Detailed results of sparse representation for two test samples [columns 1 and 7 in (a)]. For both cases, our MTJSRC method makes the right decision. For each column in (a), a green border indicates correct classification and a red border indicates incorrect classification. In all of these cases, one of the single features classifies incorrectly but the combination makes the right decision. The testing image in (b) belongs to subject 2. Due to illumination variation, the gray-level feature fails while the LBP feature succeeds. The testing image in (c) belongs to subject 37. In this case, the LBP feature fails but the gray-level feature classifies correctly. For better viewing, please see original color pdf file.

as to extract features from the flowers alone and not from the surrounding vegetation. Some details on data set description and experiment setup are given below:

1) **The 17 category data set:** This data set consists of 17 species of flowers with 80 images per class, totalling 1360 images. The classification is carried out on the basis of χ^2 distance matrices of clustered HSV, HOG, SIFT on the foreground internal region (SIFTint), SIFT on the foreground boundary (SIFTbdy) and three matrices derived from color, shape and texture vocabularies. The authors of [26] provide the χ^2 distance matrices of these seven features along with the three predefined training (17×40 images), validation (17×20 images) and test (17×20 images) splits on the database website[2].
2) **The 102 category data set:** This larger data set consists of 8,189 images divided into 102 flower classes. Each class consists of 40-250 images. The classification is carried out on the basis of χ^2 distance matrices of clustered HSV, HOG, SIFTint and SIFTbdy. The whole data set is divided into a training set, a validation set and a testing set. The training set and validation set each consist of 10 images per class. The test set consists of the remaining 6149 images (minimum 20 per class). The χ^2 distance matrices of these four features along with a predefined training/validation/test split are publicly available on the database website[3].

On both data sets, we use the predefined splits as aforementioned for training and parameter selection. Kernel matrices are computed as $\exp(-\chi^2(x, x')/\mu)$ where μ is set to be the mean value of the pairwise χ^2 distance on the training set. The parameter selection is conducted on the validation set. For the test set, accuracy performance is measured per class, i.e., the final performance is the classification accuracy averaged over all classes.

The accuracies by our proposed algorithms along with baselines and several state-of-the-art results directly from literatures on the 17 category data set are tabulated in Table I. The results on single feature kernel matrices are tabulated in Table I(a). One can observe that KMTJSRC based methods perform better than SVM on single features. The baseline NS method is also comparable with SVM on these kernels. The results by feature combination methods are listed in Table I(b), from which we can see that all feature combination methods dramatically improve the classification performance while our two algorithms are slightly better than the MKL, CG-Boost and LPBoost methods presented in [25]. This is not surprising because all the kernels used here are clean and thus kernel selection (which is a common key merit of MKL, CG-Boost and LPBoost) is not an issue. On the other hand, the kernel features are divergent enough to be complementary to each other in discriminability. Therefore, the multi-task learning mechanism of our method works reasonably well on this data set. As a simplification of our methods, the independent SRC combination is also competitive to the MKL, but slightly inferior to our methods which take into account the joint sparsity across different tasks. Figure 4 shows several exemplar classifications based on the seven individual features and their combinations. Note that our methods do not require any classifier training procedures and thus is more flexible in practice. The per query time of our method is around 0.1 second, while the values are 0.02s for the NS Combination and 0.13s for the SRC Combination. Here we do not report the running time for the other baselines since we did not implement those methods but directly take the accuracy results from the original papers with exactly the same experimental protocol.

Table II lists the accuracies of our methods along with the results from [26] on the 102 category data set. Table II(a) shows the results on single feature kernel matrices, from which we can see that KMTJSRC methods are competitive to SVM for single features on this data set. Table I(b) lists the results by feature combination methods, from which we again observe that our algorithms perform comparably to the state-of-the-art MKL method in [26]. The per query time of our method is around 0.07s, while the values are 0.005s for the NS Combination and 0.05s for the SRC Combination.

3) Caltech101 Data Set: We report in this subsection the evaluation results on the Caltech101 [57] which contains images of 101 categories of objects and a background class. Following the experimental protocol stated by the designers

[2]Available at http://www.robots.ox.ac.uk/~vgg/data/flowers/17/index.html.

[3]Available at http://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html.

TABLE I

ACCURACY (MEAN ± STD%) PERFORMANCE ON THE 17 CATEGORY OXFORD FLOWERS DATA SET

(a) SINGLE FEATURES

Features	NS	SVM [25]	KMTJSRC-RKHS	KMTJSRC-CG
Color	61.7 ± 3.3	60.9 ± 2.1	64.0 ± 2.1	64.0 ± 3.3
Shape	69.9 ± 3.2	70.2 ± 1.3	72.7 ± 0.3	71.5 ± 0.8
Texture	55.8 ± 1.4	63.7 ± 2.7	67.6 ± 2.4	67.6 ± 2.2
HSV	61.3 ± 0.7	62.9 ± 2.3	64.7 ± 4.1	65.0 ± 3.9
HOG	57.4 ± 3.0	58.5 ± 4.5	61.9 ± 3.6	62.6 ± 2.7
SIFTint	70.7 ± 0.7	70.6 ± 1.6	74.0 ± 2.2	74.0 ± 2.0
SIFTbdy	61.9 ± 4.2	59.4 ± 3.3	62.4 ± 3.2	63.2 ± 3.3

(b) FEATURE COMBINATION METHODS

Methods	Accuracy
NS Combination	83.2 ± 2.1
SRC Combination	85.9 ± 2.2
MKL [25]	85.2 ± 1.5
CG-Boost [25]	84.8 ± 2.2
LPBoost [25]	85.4 ± 2.4
KMTJSRC-RKHS	86.8 ± 1.8
KMTJSRC-CG	88.2 ± 2.3

TABLE II

ACCURACY (%) PERFORMANCE ON THE 102 CATEGORY OXFORD FLOWERS DATA SET

(a) SINGLE FEATURES

Features	NS	SVM [26]	KMTJSRC-RKHS	KMTJSRC-CG
HSV	39.8	43.0	43.6	42.5
HOG	34.9	49.6	46.7	48.1
SIFTint	46.6	55.1	54.7	55.2
SIFTbdy	34.1	32.0	33.0	31.6

(b) FEATURE COMBINATION METHODS

Methods	Accuracy
NS Combination	59.2
SRC Combination	70.0
MKL [26]	72.8
KMTJSRC-RKHS	71.5
KMTJSRC-CG	71.2

TABLE III

ACCURACY (MEAN ± STD%) PERFORMANCE ON THE CALTECH101 DATA SET (15 TRAINING/15 TEST)

(a) SINGLE FEATURES

Features	NS	MKL [27]	KMTJSRC-RKHS	KMTJSRC-CG
GB	40.8 ± 0.6	62.6 ± 1.2	58.3 ± 0.4	58.5 ± 0.3
PHOW-g	45.4 ± 0.9	63.9 ± 0.8	65.0 ± 0.7	64.5 ± 0.5
PHOW-c	37.3 ± 0.5	54.5 ± 0.6	56.1 ± 0.5	54.4 ± 0.7
SSIM	39.8 ± 0.8	54.3 ± 0.6	61.8 ± 0.6	59.7 ± 0.4

(b) FEATURE COMBINATION METHODS

Methods	Accuracy
NS Combination	51.7 ± 0.8
SRC Combination	69.2 ± 0.7
MKL [27]	70.0 ± 1.0
LPBoost [25]	70.7 ± 0.4
KMTJSRC-RKHS	69.5 ± 0.6
KMTJSRC-CG	70.2 ± 0.7

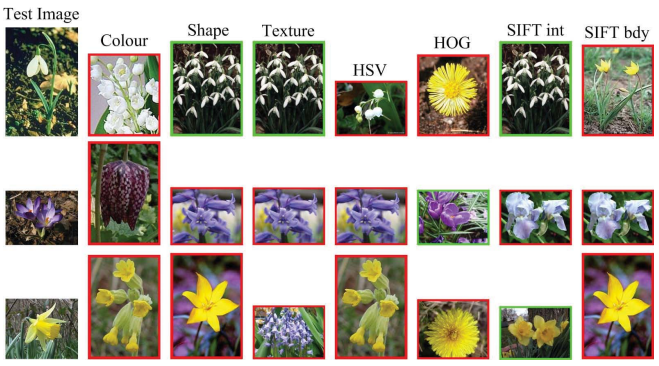

Fig. 4. Each row shows the classification results for each individual feature given a testing image (first column). The classification is indicated by the starting image from that class (i.e., not the closest image to the test image). A green border indicates correct classification and a red border incorrect classification. The last two rows, for example, show images where the SIFTint is the only single feature that classifies them correctly, but the combination of features still gets it right.

of this data set, we select 15 training images per category and test on 15 images per category. Evaluation includes all 102 classes averaged over three random training/test splits, and the performance is measured as the mean accuracy per class. In our experiment, we use four image features including Geometric Blur (GB) [21], Phow-gray/color [23] and SSIM [24]. The latter three are represented in spatial pyramid with two levels. These features are extracted using the MKL code package from [27]. Table III lists the accuracies of our methods along with the results from [25], [27]. Once again we observe that our methods are quite competitive to MKL and LPBoost for visual feature combination. The SRC combination method, which as aforementioned is a simplification of our method, can also achieve state-of-the-art results. The per query time of our method is around 0.16s, while the values are 0.01s for the NS Combination and 0.15s for the SRC Combination.

C. Face Recognition in Video Sequences

In this subsection, we apply MTJSRC to robust face recognition in videos. The problem of video-based face recognition is to identify the subjects of faces within each video sequence. To address this challenging task, we represent each probe face as an ensemble of samples detected from consecutive video frames, and further seek for the linear sparse representation of

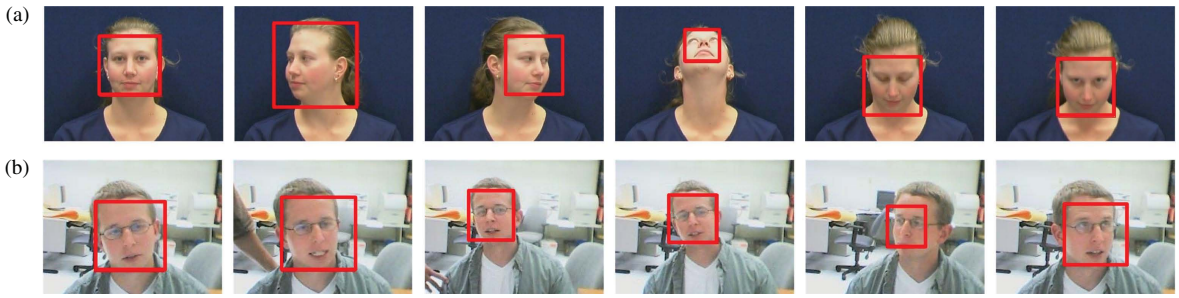

Fig. 5. Two probe video sequences from (a) VidTIMIT database [58] and (b) NRC-IIT database [59]. Each row plots six video frames of one person, overlaid with the bounding box of faces region detected by the algorithm [28].

Fig. 6. Example recognition results of various recognition algorithms for for subject 7 from the NRC-IIT database. For each frame, a blue border indicates incorrect recognition. The baselines algorithms, SRC, LDA+NN, and PCA+NN, achieve 4, 4, 2 failure recognitions respectively. In contrast, MTJSRC has made the right decisions for all these frames.

these probe images following the group constraint. Different image samples of the same subject are capable of characterizing the slightly different properties (e.g., poses, expressions, or other variations), and are expected to be complementary to each other while recognizing the face. In reconstruction of the probe image, the coefficients on these faces are enforced (by MTJSRC) to be joint sparse, which leads to a robust estimation against appearance and/or expressions variants.

1) Experimental Setting: In this experiment, we use the face detector proposed by Viola *et al.* [28] to automatically localize

TABLE IV
ACCURACY COMPARISONS ON VIDTIMIT DATABASE FOR VIDEO-BASED FACE RECOGNITION [MEAN(±STD)%]

Algorithms	Accuracy
SRC [5]	92.8(±0.4)
MTJSRC	**93.4**(±0.4)

TABLE V
ACCURACY COMPARISONS OF VARIOUS ALGORITHMS ON NRC-IIT DATABASE FOR VIDEO-BASED FACE RECOGNITION (%)1267

Algorithms	Accuracy
SRC [5]	95.6
MTJSRC	**97.8**

and crop the most confident face image from each video frame. The cropped face images are then resized to 32×32 gray images with 256 gray levels. We collect the faces detected in the most recent L frames and try to classify them in a joint manner to decide the face identity in the current frame. By taking each detected face as one instance of the same probe subject, we could apply MTJSRC for recognition. Here we describe each cropped and resized face image with the LBP feature, i.e., $K = 1$. Note that our method is immediately applicable to the setup of multiple features.

2) Data Sets: We use two data sets for evaluation: the VidTIMIT database [58] and the NRC-IIT Facial Video Database [59]. The VidTIMIT is a multi-modal database consisting of video sequences from 43 distinct volunteers (19 female and 24 male), reciting short sentences. All the videos are recorded in different sessions and the delay between sessions allows for changes in the hair style, make-up, clothing and expressions. For each person, there are 10 videos recording the sentence speaking and 3 videos recording the head moving and rotating. The video of each person is stored as a sequence of JPEG images with a resolution of 384×512 pixels. We randomly select three of the 10 videos of reciting sentences as the gallery set and use the 3 videos of moving head as the probe set. The NRC-IIT database [59] contains pairs of short video clips each showing a face of a computer user sitting in front of the monitor. There are 11 subjects and each has one pair of videos, which exhibit a wide range of facial expressions and orientations as captured by a web camera mounted on the computer monitor. The video resolution is kept to 160×120 pixels. In our experiment, one sequence of each subject is used as gallery set while the other is used as probe set. Figure 5 shows several probe video frames overlaid with the detected faces from the VidTIMIT and the NRC-IIT databases.

3) Baseline: We evaluate MTJSRC by comparing it with the SRC [5] algorithm as aforementioned in the previous experiment. For SRC, we slightly modify the implementations of SRC to elicit the representations of ensemble of face instances. First, each face instance is independently reconstructed by the gallery face images. The reconstruction errors related to the specific subject are then accumulated cross all the probe face instances. Finally, the probe image is assigned to the subject that achieves the least reconstruction error. It is worth noting that we do not try to compare our method with other face recognition algorithms (see, e.g., [60] for a survey) since SRC is reported [3], [5] to be the most promising face recognition algorithm in terms of robustness against various variations commonly occurred in real-world images.

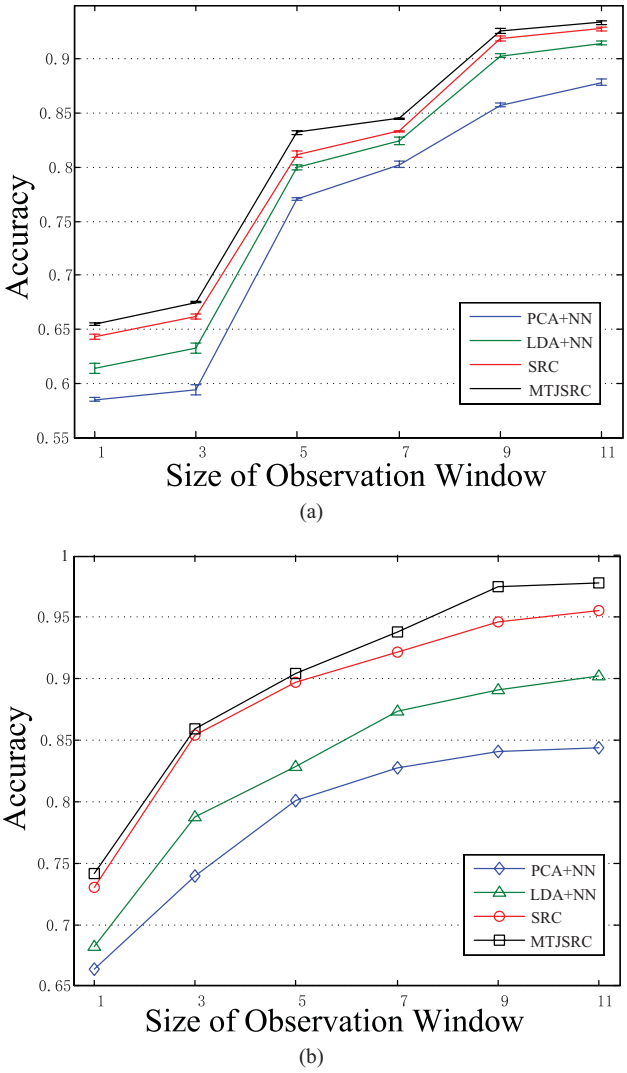

Fig. 7. Recognition accuracies of various algorithms increase along with the increases of the sliding window size. (a) VidTIMIT database. (b) NRC-IIT database.

4) Results and Analysis: We tabulate the recognition accuracies of MTJSRC and SRC in Tables IV and V, for the VidTIMIT and NRC-IIT respectively. The results show an improved performance of MTJSRC over SRC. Some exemplar results of classification are shown in Figure 6. As expected, the failure cases by SRC are mainly due to the variations of poses and scales. The proposed multi-task joint sparse representation framework, on the other hand, is robust to these variations. Here we use $L = 12$ most recent frames for joint face recognition. Moreover, we plot in Figure 7 the recognition accuracy as a function of sliding window size L. From these curves a consequence of observations are made in order. First, under different L, MTJSRC consistently outperforms SRC. Second, all the four algorithms achieve encouraging

improvements when enlarging the size of sliding window. This makes sense because more face instances will cover a wider range of poses, illuminations, expressions and scales. In addition, it is interesting to see that although MTJSRC degenerates to the single-task sparse representation when the size of sliding window is one, it still achieves higher performance as compared to SRC on both the VidTIMIT and NRC-IIT databases. This is partially due to the effectiveness of the proposed proximal gradient descent procedure for optimization.

VI. Conclusion

In this paper we develop the MTJSRC algorithm and its kernel extensions for visual classification applications. We observe that the multi-task joint sparse representation is a simple yet effective way to fuse multiple complementary visual features and instances to improve the classification accuracy. Experiments on challenging multi-class object recognition and face recognition data sets show that our method performs quite competitive to several representative state-of-the-art approaches. It is interesting to note that when $K = L = 1$, the proposed KMTJSRC-RKHS and KMTJSRC-CG methods reduce to two kernel-views of SRC, both of which perform favorably to SVM. Similar to SRC, one appealing aspect of our methods lies in that it is free of classifier training, and thus novel reference samples can be introduced without additional efforts for classifier update. In summary, we can conclude with observations that multi-task joint sparse representation is an effective method for visual classification with multiple features and/or ensemble of testing image instances.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments on this paper.

References

[1] E. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[2] J. Wright and Y. Ma, "Dense error correction via ℓ^1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.

[3] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Toward a practical face recognition system: Robust registration and illumination by sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 597–604.

[4] J. Wright, Y. Ma, J. Mairal, G. Spairo, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[5] 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, Feb. 2009.

[6] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 792–801.

[7] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[8] R. Rao, B. Olshausen, and M. Lewicki, *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press, 2002.

[9] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[10] R. Caruana, "Multi-task learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[11] T. Kato, H. Kashima, M. Sugiyama, and K. Asai, "Multi-task learning via conic programming," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2008.

[12] S. Ozawa, A. Roy, and D. Roussinov, "A multitask learning model for online pattern recognition," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 430–445, Mar. 2009.

[13] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[14] S. Bickel, J. Bogojeska, T. Lengauers, and T. Scheffer, "Multi-task learning for HIV therapy screening," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 56–63.

[15] H. Liu, M. Palatucci, and J. Zhang, "Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 649–656.

[16] G. Obozinski, B. Taskar, and M. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *J. Stat. Comput.*, vol. 20, no. 2, pp. 231–252, 2009.

[17] J. Zhang, "A probabilistic framework for multi-task learning," School Comput. Sci., Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-LTI-06-006, 2006.

[18] M. Fornasier and H. Rauhut, "Recovery algorithm for vector-valued data with joint sparsity constraints," *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[19] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE, Catholic Univ. Louvain, Louvain-la-Neuve, Belgium, Tech. Rep. 2007/076, 2007.

[20] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *SIAM J. Optim.*, submitted for publication.

[21] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 26–33.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 887–893.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Oct. 2006, pp. 2169–2178.

[24] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[25] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2009, pp. 221–228.

[26] M. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[27] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[28] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[29] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–451, 2004.

[31] T. Zhang, "Some sharp performance bounds for least squares regression with L1 regularization," *Ann. Stat.*, vol. 37, no. 5A, pp. 2109–2144, 2009.

[32] P. Zhao and B. Yu, "On model selection consistancy of lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2567, Dec. 2006.

[33] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," Dept. Stat. Sci., Univ. California, Berkeley, Tech. Rep. HAL 00621245-v2, 2012.

[34] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[35] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Ann. Stat.*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[36] D. Singaraju, R. Tron, E. Elhamifar, A. Yang, and S. S. Sastry, "On the lagrangian biduality of sparsity minimization problems," Dept. Electr. Eng. Comput. Sci., Univ. California, Berkeley, Tech. Rep. UCB/EECS-2011-70, 2011.

[37] B. D. Rao, "Analysis and extensions of the focuss algorithm," in *Proc. Conf. Record 30th Asilomar Conf. Signals, Syst. Comput.*, 1996, pp. 1218–1223.

[38] J. A.Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *EURASIP J. Appl. Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.
[39] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.
[40] M. Elad, M. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.
[41] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1873–1879.
[42] A. Majumdar and R. Ward, "Classification via group sparsity promoting regularization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1. Apr. 2009, pp. 861–864.
[43] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.
[44] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.
[45] Y. Zhou, R. Jin, and S.-C. Hoi, "Exclusive lasso for multi-task feature selection," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2010, pp. 988–995.
[46] Y. Lin, T. Liu, and C. Fuh, "Local ensemble kernel learning for object category recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2007, pp. 1–8.
[47] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[48] J. Friedman, T. Hastie, H. Hudotofling, and R. Tibshirani. "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.
[49] D. Donoho, "For most large undertetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–826, 2006.
[50] S. Becker, J. Bobin, and E. J. Candès, "NESTA: A fast and accurate first-order method for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 1–39, 2009.
[51] X. Chen, W. Pan, J. Kwok, and J. Garbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2009, pp. 746–751.
[52] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmonic Anal.*, vol. 27, no. 3, pp. 303–324, 2009.
[53] M. Kowalski and B. Torreesani, "Sparsity and persistence: Mixed norms provide simple signals models with dependent coefficient," *Signal, Image Video Process.*, vol. 3, pp. 251–264, 2009.
[54] M. Schmidt, E. Berg, M. Friedlander, and K. Murphy, "Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm," in *Proc. Int. Conf. Artif. Intell. Stat.*, vol. 5. 2009, pp. 456–463.
[55] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[56] M. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Oct. 2006, pp. 1447–1454.
[57] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. CVPR Workshop Generat.-Model Based Vis.*, 2004, pp. 59–70.
[58] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. Saarbrücken, Germany: VDM-Verlag, 2008.
[59] D. Gorodnichy, "Video-based framework for face recognition in video," in *Proc. 2nd Canadian Conf. Comput. Robot Vis.*, 2005, pp. 330–338.
[60] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

Xiao-Tong Yuan received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009.

He is currently a Post-Doctoral Research Fellow with the Department of Statistics and Biostatistics, Rutgers University, Newark, NJ. He has authored or co-authored over 30 technical papers over a wide range of research topics. His current research interests include machine learning, data mining, and computer visions.

Dr. Yuan was the recipient of the Classification Task Prize in PASCAL VOC in 2010.

Xiaobai Liu is pursuing the Ph.D. degree with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China.

He has been a Research Scholar with the Department of Statistics, University of California, Los Angles, since July 2011. From September 2007 to December 2008, he was a Research Associate with the Lotus Hill Research Institute, Ezhou, China, under the supervision of Prof. S. C. Zhu. From December 2008 to July 2011, he was a Research Associate with the Learning and Vision Group, National University of Singapore, Singapore, under the supervision of Prof. S. Yan. He has authored or co-authored more than 20 papers over a series of research topics. His current research interests include computer visions, machine learning, and large-scale image retrieval.

Shuicheng Yan (M'06–SM'09) is currently an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, and the Lead Founder of the Learning and Vision Research Group. He has authored or co-authored more than 250 technical papers over a wide range of research topics. His current research interests include computer visions, multimedia, and machine learning.

Prof. Yan is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) and the *ACM Transactions on Intelligent Systems and Technology*. He is currently the Guest Editor of the special issues for the IEEE TRANSACTIONS ON MULTIMEDIA and *Computer Vision and Image Understanding*. He was a recipient of the Best Paper Awards from PCM'11, ACM MM'10, ICME'10, and ICIMCS'09, the Classification Task Prize in both PASCAL VOC'10 and PASCAL VOC'11, the Honorable Mention Prize of the Detection Task in PASCAL VOC'10, the TCSVT Best Associate Editor Award in 2010, the Young Faculty Research Award in 2010, the Singapore Young Scientist Award in 2011, the NUS Young Researcher Award in 2012, and the co-author of the Best Student Paper Awards of PREMIA'09, PREMIA'11, and PREMIA'12.