Semisupervised Feature Analysis by Mining Correlations Among Multiple Tasks

Xiaojun Chang and Yi Yang

***Abstract*— In this paper, we propose a novel semisupervised feature selection framework by mining correlations among multiple tasks and apply it to different multimedia applications. Instead of independently computing the importance of features for each task, our algorithm leverages shared knowledge from multiple related tasks, thus improving the performance of feature selection. Note that the proposed algorithm is built upon an assumption that different tasks share some common structures. The proposed algorithm selects features in a batch mode, by which the correlations between various features are taken into consideration. Besides, considering the fact that labeling a large amount of training data in real world is both time-consuming and tedious, we adopt manifold learning, which exploits both labeled and unlabeled training data for a feature space analysis. Since the objective function is nonsmooth and difficult to solve, we propose an iteractive algorithm with fast convergence. Extensive experiments on different applications demonstrate that our algorithm outperforms the other state-of-the-art feature selection algorithms.**

***Index Terms*— 3-D motion data analysis, gene pattern recognition, image annotation, multitask feature selection, semisupervised learning.**

I. Introduction

IN MANY computer vision and pattern recognition applications, the dimension of data representation is usually very high. Recent studies have claimed that not all features in the high-dimensional feature space are discriminative and informative, since many features are often noisy or correlated with each other, which will deteriorate the performances of subsequent data analyzing tasks [1]–[4]. Consequently, feature selection is utilized to select a subset of features from the original high-dimensional feature space [5]–[12]. It has twofold functions in enhancing the performances of learning tasks. First, feature selection eliminates noisy and redundant information to get better representation, thus facilitating classification and clustering tasks. Second, the dimension of selected feature space becomes much lower, which makes the subsequent computation more efficient. Inspired by the motivations, researchers have made much progress to feature selection during last few years.

Manuscript received March 26, 2016; accepted June 8, 2016. Date of publication March 26, 2016; date of current version September 15, 2017. This work was supported in part by the Data to Decisions Cooperative Research Centre www.d2dcrc.com.au, and in part by the Australian Research Council Discovery Projects.

The authors are with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: cxj273@gmail.com; yee.i.yang@gmail.com).

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/TNNLS.2016.2582746

According to the availability of class labels of training data, feature selection algorithms fall into two groups, i.e., supervised feature selection and unsupervised feature selection. Supervised feature selection algorithms, for example, Fisher score [13], only use labeled training data for feature selection. With sufficient labeled training samples, supervised feature selection is reliable to train appropriate feature selection functions because of utilization of class labels. However, labeling a large amount of training samples manually is unrealistic in real-world applications. Recent works on semisupervised learning have indicated that it is beneficial to leverage both labeled and unlabeled training data for the data analysis. Motivated by the progress of semisupervised learning, much research attention has been paid to semisupervised feature selection. For example, Zhao *et al.* [50] propose a semisupervised feature selection algorithm based on spectral analysis. A common limitation of the existing supervised and semisupervised feature selection algorithms is that they evaluate the importance of each feature individually, ignoring correlations between different features. To address this problem, some state-of-the-art algorithms are proposed to take feature correlations into consideration for feature selection. For example, [3] and [14] implement their methods in a supervised way and Ma *et al.* [6] design their approach in a semisupervise way.

Another limitation of current feature selection algorithms is that they select features for each task individually, which fails to mine correlations among multiple related tasks. Recent studies have indicated that it is beneficial to learn multiple related tasks jointly [15]–[18]. Motivated by this fact, multitask learning (MTL) has been introduced to the field of multimedia. For instance, Yang *et al.* [7] present a novel feature selection algorithm which leverages shared information from related tasks. Nevertheless, they design their algorithm in a supervised way.

The semisupervised algorithm proposed in this paper combines the strengths of semisupervised feature selection and MTL. Both labeled and unlabeled training data are utilized for feature selection. Meanwhile, correlations between different features are taken into consideration to improve the performance of feature selection.

We illustrate how the proposed algorithm works for video classification. First, we represent all the training and testing videos as feature vectors. Then, sparse coefficients are learned by exploiting relationships among different features and leveraging knowledge from multiple related tasks. After selecting the most representative features, we can apply

the sparse coefficients to the feature vectors of the testing videos for classification.

We name the proposed algorithm semisupervised feature selection by mining correlations among multiple tasks, abbreviated as SFMC.

The main contributions of this paper can be summarized as follows.

1) We propose a novel efficient semisupervised multitask feature selection algorithm, which is able to mine the correlations among multiple tasks when few labeled training data are provided.
2) To explore correlations among the data points, we leverage the benefit of manifold learning into the proposed framework.
3) The proposed framework is generic and can be readily extended to many existing algorithms.
4) We have conducted extensive experiments to validate the effectiveness of the proposed algorithm. The experimental results confirm its superiority.

The rest of this paper is organized as follows. Section II summarizes the overview of related work. A novel semisupervised feature selection by mining correlations among multiple tasks is proposed in Section III. We present our experimental results in Section IV. The conclusion of our work is discussed in Section V.

II. Related Work

In this section, we briefly review the related research on feature selection, semisupervised learning, and MTL.

A. Feature Selection

Previous works have claimed that feature selection is capable of selecting the most representative features, thus facilitating subsequent data analyzing tasks [3], [19]–[21].

Existing feature selection algorithms are designed in various ways. Classical feature selection algorithms, such as Fisher score [13], evaluate the weights of all features, rank them accordingly, and select the most discriminating features one by one [22]. Although these classical feature selection algorithms gain good performances in different applications, they have three main limitations. First, they only use labeled training data to exploit the correlations between features and labels for feature selection. Labeling a large amount of training data consumes a lot of human labor in real-world applications. Second, the most representative features are selected one by one, thus ignoring the correlations among different features. Third, they select features for each task independently, which fails to leverage the knowledge shared by multiple related tasks.

To overcome the aforementioned limitations, researchers have proposed multiple feature selection algorithms. $l_{2,1}$-norm regularization has been widely used in feature selection algorithms for its capability of selecting features across all data points with joint sparsity. For example, Zheng and Liu [23] propose an algorithm which selects features jointly based on spectral regression with $l_{2,1}$-norm constraint. Nie *et al.* [14] adopt $l_{2,1}$-norm on both regularization term and loss function. Yang *et al.* [7] propose to select features by leveraging shared knowledge from multiple related tasks. However, their algorithms are all designed in a supervised way.

B. Semisupervised Learning

Semisupervised learning has shown its promising performance in different applications [24]–[30]. With semisupervised learning, unlabeled training data can be exploited to learn data structure, which can save human labor cost for labeling a large amount of training data [31]–[35]. Hence, semisupervised learning is beneficial in terms of both the human laboring cost and data analysis performance.

Graph Laplacian-based semisupervised learning has gained increasing interest for its simplicity and efficiency [36]. Nie *et al.* [21] propose a manifold learning framework based on graph Laplacian and compared its performance with the other state-of-the-art semisupervised algorithms. Ma *et al.* [6] propose a semisupervised feature selection algorithm built upon manifold learning. Yang *et al.* [37] propose a new semisupervised algorithm based on a robust Laplacian matrix for relevance feedback. Their algorithm has demonstrated its prominent performance. Therefore, we propose to leverage it in our feature selection framework. These previous works, however, independently select features for each task, which fails to consider correlations among multiple-related tasks.

C. Multitask Learning

MTL is an inductive transfer mechanism whose principle goal is to improve generalization performance. It improves generalization by leveraging the domain-specific information contained in the training signals of multiple related tasks. It achieves this goal by training multiple tasks jointly based on a shared representation. MTL has been widely used in many applications with the appealing advantage that it learns multiple related tasks with a shared representation [15], [16], [38], [39]. Recent studies have indicated that learning multiple related tasks jointly always outperform learning them independently. Inspired by the progress of MTL, researchers have introduced it to the field of multimedia and demonstrated its promising performance on multimedia analysis. For example, Yang *et al.* [7] propose a novel multitask feature selection algorithm which improves feature selection performance by leveraging shared information among multiple related tasks. Yang *et al.* [7] apply knowledge adaptation to multimedia event detection and compare its performance with several state-of-the-art algorithms. Despite their good performances, these classical algorithms are all implemented only with labeled training data.

III. Methodology

In this section, we describe the approach of our proposed algorithm in detail.

A. Problem Formulation

Suppose we are going to select features for t tasks. The lth task contains n_l training data with m_l data labeled.

We can formulate the regularized framework for feature selection as follows:

$$\min_{W_l} \sum_{l=1}^{t} (\text{loss}(W_l) + \alpha g(W_l)) + \gamma\, \Omega(W) \tag{1}$$

where W_l is feature selection matrix for the lth task, $W = [W_1, \ldots, W_t]$, $\text{loss}(W_l)$ is the loss function which evaluates consistency between features and labels, $g(W_l)$ is a regularization function, $\Omega(W)$ is a regularization term which is used to encode the common components of different feature selection functions, and α and γ are regularization parameters.

To step further, we first give the definitions of Frobenius norm and trace norm. Given an arbitrary matrix $M \in \mathbb{R}^{a\times b}$, where a and b are arbitrary numbers, its Frobenius norm is defined as $\|M\|_F$. The definition of its trace norm is

$$\|M\|_* = \text{Tr}(MM^T)^{\frac{1}{2}} \tag{2}$$

where $\text{Tr}(\cdot)$ denotes the trace operator. In the literature, there are many approaches to define the loss function. Following the works in [6] and [7], we adopt the least square loss function for its simplicity and efficiency. ℓ_2^2-norm regularizer has been widely used in different applications to avoid the overfitting problem [14], [22]. Motivated by the works in [7] and [40], we propose to leverage shared knowledge among multiple related tasks by minimizing the trace norm of W. The objective function is given by

$$\min_{W_l} \sum_{l=1}^{t} (\text{loss}(W_l) + \alpha \|W_l\|_F) + \gamma\, \|W\|_*. \tag{3}$$

The state-of-the-art feature selection algorithms are implemented through supervised learning and select features for each task independently. In this paper, we want to incorporate MTL and semisupervised learning into (1). We propose to leverage semisupervised learning by adopting the Laplacian proposed in [37]. We adopt this Laplacian because it exploits both manifold structure and local discriminant information of multimedia data, thus resulting in better performance.

To begin with, let us define $X_l = [x_l^1, \ldots, x_l^{n_l}]$ as the training data matrix of the lth task where m_l data are labeled and n_l is the total number of the training data of the lth task. $x_l^i \in \mathbb{R}^d$ is the ith datum of the lth task. $Y_l = [y_l^1, \ldots, y_l^{m_l}, y_l^{m_l+1}, \ldots, y_l^{n_l}]^T \in \{0,1\}^{n_l\times c_l}$ is the label matrix and c_l denotes the class number of the lth task. $y_l^i|_{i=1}^{n_l} \in \mathbb{R}^{c_l}$ is the label vector with c_l classes. $Y_{li,j} = 1$ if x_l^i is in the jth class of the lth task, while $Y_{li,j} = 0$ otherwise. For unlabeled datum x_l^i, y_l^i is set to a zero vector. For any d, we define $\mathbf{1}_d \in \mathbb{R}^d$ as a column vector with all the elements equal to 1, $H_d = I - (1/d)\mathbf{1}_d\mathbf{1}_d^T \in \mathbb{R}^{d\times d}$ as a matrix for centering the data by subtracting the mean of the data. Note that $H_d = H_d^T = H_d H_d$. For each data point x_l^i of the lth task, we construct a local clique $\mathcal{N}_{lk}$ containing x_l^i and its $k-1$ nearest neighbors. Euclidean distance is used to determine whether two given data points are within k nearest neighbors in the original feature space. $G_l^i = \{i_l^0, i_l^1, \ldots, i_l^{k-1}\}$ is index set of samples in $\mathcal{N}_{lk}$. S_{li} denotes selection matrix with its elements $(S_{li})_{pq} = 1$ if $p = G_l^i\{q\}$ and $(S_{li})_{pq} = 0$ otherwise.

Inspired by [37], we construct the Laplacian matrix by exploiting both manifold structure and local discriminant information. Denoting $L_{li} = H_k(X_l^T X_l + \lambda I)^{-1} H_k$, we compute the Laplacian matrix L as follows:

$$\begin{aligned} L_l &= \sum_{i=1}^{n_l} S_{li} L_{li} S_{li}^T \\ &= [S_{l1}, \ldots, S_{ln_l}] \begin{bmatrix} L_{l1} & & \\ & \cdots & \\ & & L_{ln_l} \end{bmatrix} [S_{l1}, \ldots, S_{ln_l}]^T. \end{aligned} \tag{4}$$

Note that manifold regularization is able to explore the data manifold structure [21], [41], [42]. By applying manifold regularization to the loss function in (1), we have

$$\begin{aligned} \arg\min_{W,b} \sum_{l=1}^{t} & \text{Tr}\big(W^T X_l L_l X_l^T W\big) \\ & + \alpha\big(\|W_l\|_F + \beta\|X_{lL}^T W_l + \mathbf{1}_{n_l} b_l^T - Y_{lL}\|_F\big) \\ & + \gamma\, \|W\|_* \end{aligned} \tag{5}$$

where $\text{Tr}(\cdot)$ denotes trace operator, and X_{lL} and Y_{lL} are labeled training data and corresponding ground truth labels of the lth task.

To make all labels of training data contribute to the optimization of W_l, we introduce a predicted label matrix $F_l = [f_{l1}, \ldots, f_{ln_l}] \in \mathbb{R}^{n_l\times c_l}$ for the training data of the lth task. $f_{li} \in \mathbb{R}^{c_l}$ is the predicted label vector of x_{li}. According to [6] and [24], F_l can be obtained as follows:

$$\arg\min_{F_l} \text{Tr}\big(F_l^T L_l F_l\big) + \text{Tr}((F_l - Y_l)^T U_l (F_l - Y_l)) \tag{6}$$

where U_l is the selection diagonal matrix of the lth task. The diagonal element $U_{lii} = \infty$ if x_{li} is labeled and $U_{lii} = 1$ otherwise. In the experiments, 10^6 is used to approximate ∞.

Following the work in [6], we incorporate (6) into (5). At the same time, all the training data and corresponding labels are taken into consideration. Therefore, the objective function finally arrives at

$$\begin{aligned} \min_{F_l,W_l,b_l} \sum_{l=1}^{t} & \big(\text{Tr}[(F_l - Y_l)^T U_l (F_l - Y_l)] + \text{Tr}\big(F_l^T L_l F_l\big) \\ & + \alpha\big(\|W_l\|_F + \beta\|X_l^T W_l + \mathbf{1}_{n_l} b_l^T - F_l\|_F^2\big)\big) \\ & + \gamma\, \|W\|_*. \end{aligned} \tag{7}$$

From (7), we can see that the proposed algorithm is capable of evaluating the informativeness of all features jointly for each task with the $l_{2,1}$-norm and the information from different tasks can be transferred from one to another with the trace norm.

B. Optimization

The proposed function involves the $l_{2,1}$-norm and trace norm, which are difficult to solve in a closed form. We propose to solve this problem in the following steps.

By setting the derivative of (7) with respect to b_l to 0, we obtain

$$b_l = \frac{\mathbf{1}}{n_l}\big(F_l - X_l^T W_l\big)^T \mathbf{1}_{n_l}. \tag{8}$$

Substituting b_l in (7) with (8), we obtain

$$\min_{F_l,W_l,b_l}\sum_{l=1}^{t}\Big(\mathrm{Tr}[(F_l-Y_l)^T U_l(F_l-Y_l)]+\mathrm{Tr}\big(F_l^T L_l F_l\big) +\alpha\left(\|W_l\|_F+\beta\|X_l^T W_l+\frac{1}{n_l}\mathbf{1}_{n_l}\mathbf{1}_{n_l}^T \times\big(F_l-X_l^T W_l\big)-F_l\|_F\right)\Big)+\gamma\|W\|_*$$

$$\Rightarrow \min_{F_l,W_l}\sum_{l=1}^{t}\big(\mathrm{Tr}[(F_l-Y_l)^T U_l(F_l-Y_l)]+\mathrm{Tr}\big(F_l^T L_l F_l\big) +\alpha\big(\|W_l\|_{2,1}+\beta\|H_{n_l}X_l^T W_l-H_{n_l}F_l\|_F^2\big)\big)+\gamma\|W\|_* \quad (9)$$

where $H_{n_l}=I_{n_l}-(\mathbf{1}/n_l)\mathbf{1}_{n_l}\mathbf{1}_{n_l}^T$ is a centering matrix. By setting the derivative of (9) with respect to F_l to 0, we have

$$2U_lF_l-2U_lY_l+2L_lF_l+\alpha\beta\big(2H_{n_l}F_l-2H_{n_l}X_l^T W_l\big)=0.$$

Therefore, we have

$$F_l=(\alpha\beta H_{n_l}+U_l+L_l)^{-1}\big(\alpha\beta H_{n_l}X_l^T W_l+U_lY_l\big). \quad (10)$$

Denoting $P_l=(\alpha\beta H_{n_l}+U_l+L_l)^{-1}$ and $Q_l=\alpha\beta H_{n_l}X_l^T W_l+U_lY_l$, we have

$$F_l=P_lQ_l. \quad (11)$$

By substituting F_l into (9) with (11), we can rewrite the objective function as follows:

$$\min_{Q_l,W_l}\sum_{l=1}^{t}\big(\mathrm{Tr}[(P_lQ_l-Y_l)^T U_l(P_lQ_l-Y_l)] +\mathrm{Tr}\big(Q_l^T P_l^T L_lP_lQ_l\big)+\alpha\big(\|W_l\|_F+\beta\big\|H_{n_l}X_l^T W_l -H_{n_l}P_lQ_l\big\|_F\big)\big)+\gamma\|W\|_*. \quad (12)$$

As $\mathrm{Tr}(Q_l^T P_l^T U_lY_l)=\mathrm{Tr}(Y_l^T U_l^T P_lQ_l)$ and $\mathrm{Tr}(\alpha\beta W_l^T X_l H_lP_lQ_l)=\mathrm{Tr}(\alpha\beta Q_l^T P_l^T H_lX_l^T W_l)$, the objective function can be rewritten as follows:

$$\min_{W_l}\sum_{l=1}^{t}\big(\alpha\beta\mathrm{Tr}\big(W_l^T X_lH_{n_l}(I_{n_l}-\alpha\beta P_l)H_{n_l}X_l^T W_l\big) -2\alpha\beta\mathrm{Tr}\big(W_l^T X_lH_{n_l}P_lU_lY_l\big)+\alpha\|W_l\|_F\big)+\gamma\|W\|_*. \quad (13)$$

Denoting $R_l=X_lH_{n_l}(I_{n_l}-\alpha\beta P_l)H_{n_l}X_l^T$, $T_l=X_lH_{n_l}P_lU_lY_l$ and $W_l=[w_l^1,\ldots,w_l^d]$, the objection function becomes

$$\min_{W_l}\sum_{l=1}^{t}\big(\alpha\beta\mathrm{Tr}\big(W_l^T R_lW_l\big)-2\alpha\beta\mathrm{Tr}\big(W_l^T T_l\big)+\alpha\|W_l\|_F\big)+\gamma\|W\|_*. \quad (14)$$

Equation (14) can be rewritten as

$$\min_{W_l}\sum_{l=1}^{t}\big(\alpha\beta\mathrm{Tr}\big(W_l^T R_lW_l\big)-2\alpha\beta\mathrm{Tr}\big(W_l^T T_l\big) +\alpha\mathrm{Tr}\big(W_l^T D_lW_l\big)\big)+\frac{\gamma}{2}\mathrm{Tr}\big(W^T(WW^T)^{-\frac{1}{2}}W\big) \quad (15)$$

Algorithm 1 Optimization Algorithm for SFMC

Data: Training data $X_l|_{l=1}^t\in\mathbb{R}^{d\times n_l}$
Training data labels $Y_l|_{l=1}^t\in\mathbb{R}^{n\times c}$
Parameters γ, α and β
Result:
Feature Selection Matrix $W_l|_{l=1}^t\in\mathbb{R}^{d\times c_l}$

```
l = 1 ;
while l ≤ t do
    Initialise W_l|_{l=1}^t ∈ R^{d×c_l} ;
    Compute the Laplacian matrix L_l|_{l=1}^t ;
    Compute the Selection matrix U_l|_{l=1}^t ;
    H_{n_l} = I_{n_l} − (1/n_l) 1_{n_l} 1_{n_l}^T ;
    P_l = (αβ H_{n_l} + U_l + L_l)^{-1} ;
    R_l = X_l H_{n_l} (I_{n_l} − αβ P_l) H_{n_l} X_l^T ;
    T_l = X_l H_{n_l} P_l U_l Y_l ;
end
Set r = 0 ;
Set W_0 = [W_1, ··· , W_t] ;
repeat
    l = 1 ;
    Compute the diagonal matrix as:
    D̃^r = (1/2)(W_r W_r^T)^{-1/2} ;
    while l ≤ t do
        Compute the diagonal matrix D_l^r according to
        Equation (16) ;
        Update W_l^r by W_l^r = (R_l + (α/β) D_l^r + (γ/αβ) D̃^r)^{-1} T_l ;
        Update F_l^r by
        F_l^r = (αβ H_{n_l} + U_l + L_l)^{-1} (αβ H_{n_l} X_l^T W_l + U_l Y_l) ;
        Update b_l^r by b_l^r = (1/n_l)(F_l − X_l^T W_l)^T 1_{n_l} ;
        l = l + 1 ;
    end
    W_{r+1} = [W_1, ··· , W_t] ;
    r = r + 1 ;
until Convergence;
Return the optimal W_l|_{l=1}^t and b_l|_{l=1}^t.
```

where D_l is a diagonal matrix which is defined as

$$D_l=\begin{bmatrix}\frac{1}{2\|w_l^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|w_l^d\|_2}\end{bmatrix}. \quad (16)$$

By setting the derivative with respect to W_l to 0, we have

$$W_l=\left(R_l+\frac{\alpha}{\beta}D_l+\frac{\gamma}{\alpha\beta}\widetilde{D}\right)^{-1}T_l \quad (17)$$

where $\widetilde{D}=(1/2)(WW^T)^{-1/2}$.

As shown in Algorithm 1, an iterative algorithm is proposed to optimize the objective function (7) based on the above mathematical deduction.

IV. Experiments

In this section, experiments are conducted to evaluate the performance of the proposed algorithm on video classification,

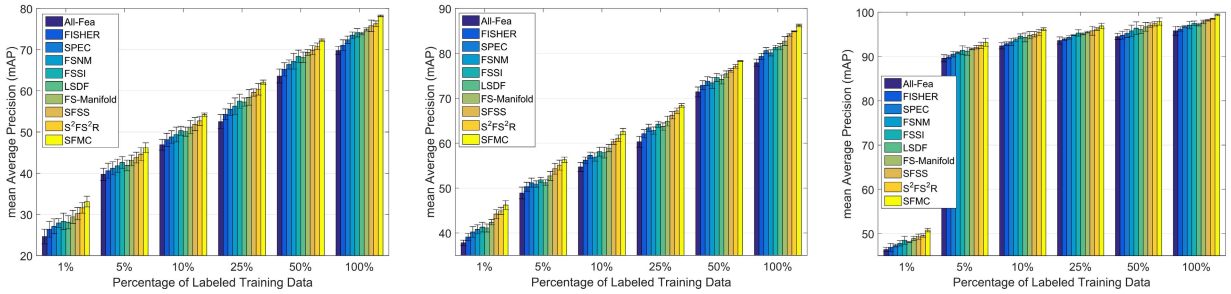


Fig. 1. Performance comparison of video classification (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color. (a) Subject 1. (b) Subject 2. (c) Subject 3.

image annotation, human motion recognition, 3-D motion recognition, and gene pattern recognition, respectively. We also conduct statistical tests on all the five tasks to ensure that the improvement of the proposed algorithm is statistically significant. To step further, we conduct experiments to study the proposed algorithm's ability of overcoming the lack of label information with the use of unlabeled training data. In addition, experiments are conducted to study the performance with respect to the influence of number of selected features and parameter sensitivity.

A. Experiment Setup

We use four different data sets in the experiment, including one video data sets, the Columbia Consumer Video (CCV) data set [43], one image data sets, the NUSWIDE data set [44], one human motion data set, the HMDB data set [45], one 3-D motion skeleton data set, the HumanEva data set [46], and three gene expression data sets [47]–[49]. In order to demonstrate the advantages of our algorithm, we compare its performance with the following approaches.

1) *All Features:* We directly use the original features without feature selection as a baseline.
2) *Fisher Score:* This is a classical feature selection method, which evaluates importance of features and selects the most discriminating features one by one [13].
3) *Feature Selection via Joint $l_{2,1}$-Norms Minimization:* Joint $l_{2,1}$-norm minimization is utilized on both loss function and regularization for joint feature selection [14].
4) *SPEC:* It uses spectral graph theory to conduct feature selection [23].
5) *Feature Selection With Shared Information Among Multiple Tasks (FSSI):* It simultaneously learns multiple feature selection functions of different tasks in a joint framework [7]. Hence, it is capable to utilize shared knowledge between multiple tasks to facilitate decision making.
6) *Locality Sensitive Semisupervised Feature Selection:* This is a semisupervised feature selection based on two graph constructions, i.e., within-class graph and between-class graph [50].
7) *Discriminative Semisupervised Feature Selection via Manifold Regularization:* It selects features through maximizing the classification margin between different classes and simultaneously exploiting the data geometry by the manifold regularization [51].
8) *Structural Feature Selection With Sparsity (SFSS):* It combines the strengths of joint feature selection and semisupervised learning into a single framework [6]. Labeled and unlabeled training data are both utilized for feature selection. Meanwhile, correlations between different features are taken into consideration.
9) *Semisupervised Feature Selection via Spline Regression:* The discriminative information between labeled training videos and the local geometry structure of all the training videos are well preserved by the combined semisupervised scatters [52].

In the experiments, a fivefold cross validation is employed to split the data into training and testing splits. During training, we learn the optimal parameters α, β, and γ through another fivefold cross validation. We independently repeat the experiments 30 times and report the average results. There is another parameter k, which specifies k nearest neighbors used to compute graph Laplacian. It is fixed at 15 as suggested in [6]. Following [53], we use a nonlinear SVM with χ^2-kernel. Mean average precision (MAP) is used to evaluate the performance.

B. Video Classification

First, we compare the performances of different algorithms in terms of video classification task using the CCV [43]. We refer the event category as Subject 1, the scene category as Subject 2 and the object category as Subject 3. Since the original videos of this data set have not been available on the Internet, we directly use the STIP features with 5000-D bag-of-word (BoW) representation provided in [43]. We set the number of selected features as $\{2500, 3000, \ldots, 4500, 5000\}$ for all the algorithms and report the best results.

We present the experimental results in terms of video classification when various percentages of labeled training data are used in Fig. 1. From the experimental results, we can see that the proposed algorithm performs better than the other alternatives when different percentages of labeled training data are utilized. We have the following observations.

1) In general, feature selection algorithms achieve better performance than the baseline method that all features

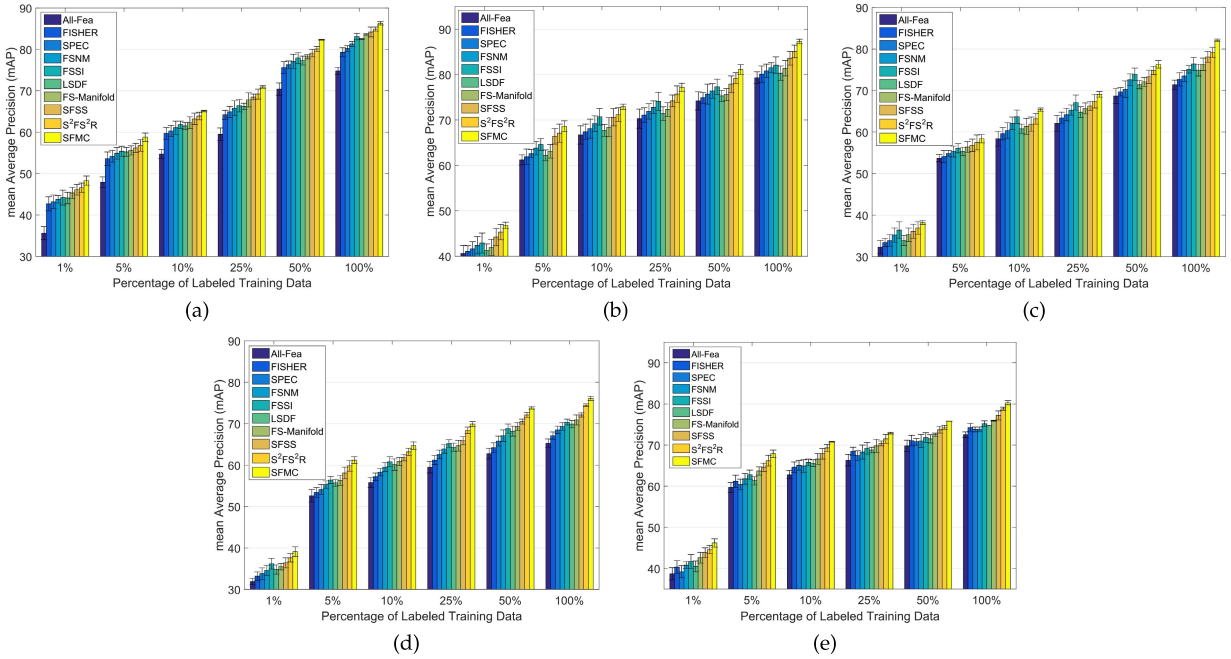


Fig. 2. Performance comparison of human motion data analysis (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color. (a) Subject 1. (b) Subject 2. (c) Subject 3. (d) Subject 4. (e) Subject 5.

are utilized, which demonstrates the necessity of feature selection for video classification.

2) As the percentage of labeled training data increases, the performance of all the compared algorithms improves. This indicates that if more labeled training data are provided, better performance is expected.
3) When the percentage of labeled training data is small (e.g.,, 1%), the semisupervised feature selection algorithms outperform the supervised alternatives by a large margin. For example, the performance of SFSS versus FISHER is 44.2% versus 39.1%. This result demonstrates the benefit of exploiting unlabeled training data for subsequent video classification.
4) FSSI generally achieves better classification results, which confirms the effectiveness of sharing information among different tasks.
5) The advantage of the proposed algorithm over the other alternatives is especially visible with only few labeled training data, i.e.,, 1% or 5%.

C. Image Annotation

We use the NUS-WIDE data set [44] to test the performance of our algorithm. This data set includes 269 648 images of 81 concepts. A 500 dimension BoW feature based on an SIFT descriptor is used in this experiment. We take each concept as a separate annotation task, thus resulting in 81 tasks. It is difficult to report all the results of these 81 tasks, so the average result is reported. In this experiment, we set the number of selected features as $\{250, 275, \ldots, 475, 500\}$ and report the best results.

We report the detailed experimental results when 1%, 5%, 10%, 25%, 50%, and 100% training data are shown in Fig. 2. From the results shown in Fig. 2, we can observe the following.

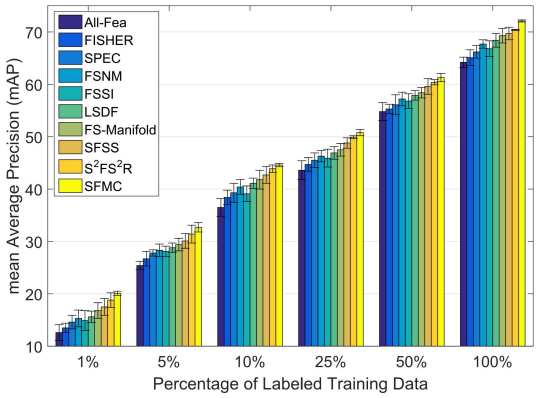
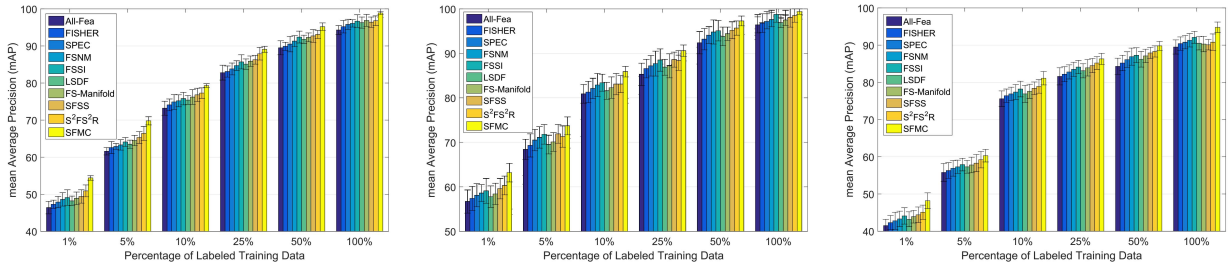


Fig. 3. Performance comparison of image annotation (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color.

1) When few training samples are labeled, semisupervised feature selection algorithms generally have much better performance than supervised alternatives, thanks to the exploration of unlabeled training data.
2) As the number of labeled training samples increases, the performance obviously improves.
3) Compared with other supervised feature selection algorithms, FSSI generally has better performance. We attribute this improvement to the effectiveness of sharing information among different tasks.
4) When the ratio of labeled training examples is very low, e.g.,, 1%, the proposed algorithm outperforms all the compared algorithms, which shows a better property of semisupervised feature selection.

D. Human Motion Recognition

We use the HMDB video data set [45] to compare the algorithms in terms of human motion recognition. HMDB data



(a) (b) (c)

Fig. 4. Performance comparison of 3-D motion data analysis (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color. (a) Subject 1. (b) Subject 2.

set consists of 6766 videos which are associated with 51 distinct action categories. These categories can be categorized into five groups: 1) general facial actions; 2) facial actions with object manipulation; 3) general body movements; 4) body movements with object interaction; and 5) body movements for human interaction. Therefore, in this experiment, the five groups are considered as five different tasks. Wang and Schmid [54] claim that motion boundary histograms are an efficient way to suppress camera motion, and thus, it is used to process the videos. A 2000 dimension BoW feature is generated to represent the original data. The number of selected features is tuned in the range of $\{1000, 1200, \ldots, 1800, 2000\}$ for all the algorithms.

The detailed experimental results in terms of human motion recognition are shown in Fig. 3. From the experimental results, we have similar observations that the proposed method outperforms other alternatives by a large margin. This experiment further provides evidence that the proposed algorithm is more advantageous with insufficient number of labeled training data.

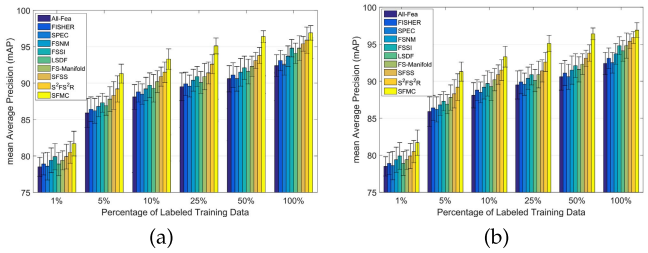
E. 3-D Motion Data Analysis

We evaluate the performance of our algorithm in terms of 3-D motion data analysis using the Human-Eva 3-D motion database. There are five different types of actions in this database. Following the work in [37] and [55], we randomly select 10 000 samples of two subjects (5000 per subject). We encode each action as a collection of 16 joint coordinates in the 3-D space and obtain a 48-D feature vector. Joint relative features between different joints are computed on top of that resulting a feature vector with 120 dimensions. We combine the two kinds of feature vectors and get a 168-D feature. In this experiment, we consider the two subjects as two different tasks. The number of selected features is tuned from $\{100, 110, \ldots, 160\}$.

Fig. 4 shows the detailed experimental results in terms of 3-D motion data analysis when 1%, 5%, 10%, 25%, 50%, and 100% training data are labeled. We can observe that our algorithm consistently outperforms the other feature selection alternatives and obtains much better improvement when a small number of training data are labeled.

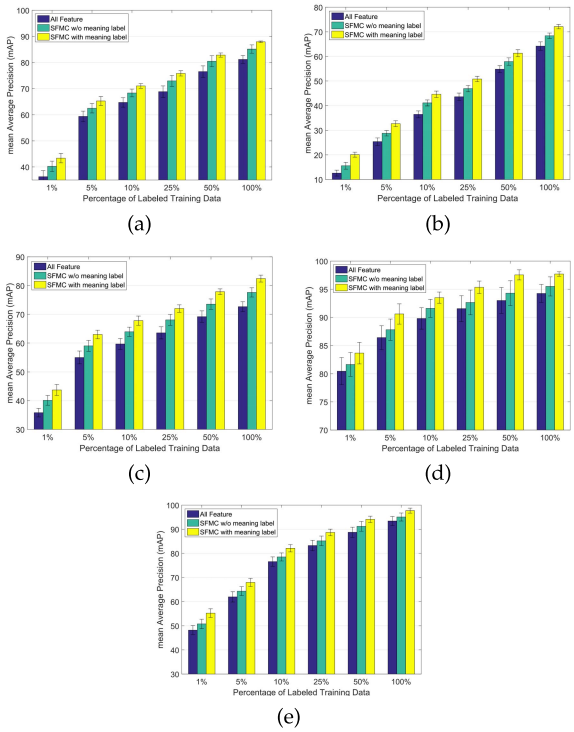
F. Gene Pattern Recognition

To further validate the performance of the proposed approach, we use three gene expression data sets:



(a) (b)

Fig. 5. Performance comparison of gene expression (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color. (a) Subject 1. (b) Subject 2. (c) Subject 3.



(a) (b) (c) (d) (e)

Fig. 6. Influence of the unlabeled data (mAP ± STD). Best view in color. (a) Video classification. (b) Image annotation. (c) Human motion Recognition. (d) 3-D motion recognition. (e) Gene pattern Recognition.

all/aml [47], lymphoma [48], and global cancer map [49]. In this experiment, we take each data set as a separate recognition task, thus resulting in three tasks.

We report the detailed experimental results in Fig. 5. From these experimental results, we have similar observations. In particular, as the number of labeled training

TABLE I

p-VALUE OF PAIR-WISED WILCOXON TEST WITH RESPECT TO 1%, 5%, 10%, 25%, 50%, AND 100% LABELED TRAINING DATA FOR VIDEO CLASSIFICATION

	1% labeled	5% labeled	10% labeled	25% labeled	50% labeled	100% labeled
All Features	2e-11	1e-11	2e-11	2e-11	3e-11	1e-11
FISHER	2e-11	2e-11	1e-11	3e-11	1e-11	3e-11
SPEC	2e-11	1e-11	3e-11	1e-11	2e-11	2e-11
FSNM	2e-11	1e-11	2e-11	3e-11	1e-11	2e-11
FSSI	3e-11	2e-11	3e-11	3e-11	4e-11	2e-11
LSDF	2e-11	3e-11	2e-11	2e-11	2e-11	1e-11
FS-Manifold	2e-11	3e-11	2e-11	3e-11	3e-11	2e-11
SFSS	5e-11	3e-11	2e-11	3e-11	1e-11	2e-11
S^2FS^2R	1e-11	2e-11	3e-11	2e-11	3e-11	2e-11

TABLE II

p-VALUE OF PAIR-WISED WILCOXON TEST WITH RESPECT TO 1%, 5%, 10%, 25%, 50%, AND 100% LABELED TRAINING DATA FOR IMAGE ANNOTATION

	1% labeled	5% labeled	10% labeled	25% labeled	50% labeled	100% labeled
All Features	2e-11	2e-11	1e-11	2e-11	4e-11	2e-11
FISHER	3e-11	3e-11	2e-11	3e-11	1e-11	3e-11
SPEC	2e-11	3e-11	2e-11	3e-11	3e-11	2e-11
FSNM	2e-11	3e-11	1e-11	3e-11	3e-11	2e-11
FSSI	3e-11	3e-11	5e-11	4e-11	3e-11	2e-11
LSDF	4e-11	3e-11	2e-11	4e-11	3e-11	1e-11
FS-Manifold	2e-11	3e-11	5e-11	4e-11	3e-11	4e-11
SFSS	4e-11	3e-11	4e-11	3e-11	4e-11	3e-11
S^2FS^2R	4e-11	3e-11	4e-11	2e-11	4e-11	3e-11

TABLE III

p-VALUE OF PAIR-WISED WILCOXON TEST WITH RESPECT TO 1%, 5%, 10%, 25%, 50%, AND 100% LABELED TRAINING DATA FOR HUMAN MOTION RECOGNITION

	1% labeled	5% labeled	10% labeled	25% labeled	50% labeled	100% labeled
All Features	4e-11	2e-11	5e-11	2e-11	3e-11	1e-11
FISHER	4e-11	3e-11	5e-11	4e-11	2e-11	3e-11
SPEC	4e-11	1e-11	5e-11	3e-11	2e-11	2e-11
FSNM	5e-11	3e-11	2e-11	3e-11	4e-11	2e-11
FSSI	3e-11	3e-11	4e-11	3e-11	2e-11	4e-11
LSDF	2e-11	4e-11	2e-11	7e-11	3e-11	3e-11
FS-Manifold	2e-11	3e-11	4e-11	2e-11	2e-11	4e-11
SFSS	4e-11	4e-11	2e-11	3e-11	2e-11	4e-11
S^2FS^2R	2e-11	4e-11	4e-11	3e-11	2e-11	4e-11

data increases, the performance of all the compared algorithms increases. The proposed algorithm is the only one which has consistently better performance on all three gene data sets. When 10% or less of the training data are labeled, the proposed algorithm consistently outperforms other methods for all the three gene data sets by a large margin.

G. Statistical Test

To ensure the improvement of the proposed algorithm is statistically significant, nonparametric pair-wised Wilcoxon test is conducted on the five tasks. The experimental results are reported in Tables I–V. We compute the p-value of the proposed algorithm against each other algorithm and set the level of significance $\alpha = 0.05$. From the results shown in Tables I–V, we can see that the proposed algorithm achieves statistically significant improvements.

H. Influence of the Unlabeled Data

To study the proposed algorithm's ability of overcoming the lack of label information with the use of unlabeled data, we conduct an experiment on the five tasks accordingly. We artificially hide the meaningful labels from the training algorithm and only use the unlabeled training data to conduct

TABLE IV

p-VALUE OF PAIR-WISED WILCOXON TEST WITH RESPECT TO 1%, 5%, 10%, 25%, 50%, AND 100% LABELED TRAINING DATA FOR 3-D MOTION RECOGNITION

	1% labeled	5% labeled	10% labeled	25% labeled	50% labeled	100% labeled
All Features	2e-11	4e-11	3e-11	4e-11	3e-11	2e-11
FISHER	3e-11	2e-11	4e-11	2e-11	4e-11	2e-11
SPEC	4e-11	3e-11	2e-11	3e-11	2e-11	3e-11
FSNM	5e-11	3e-11	1e-11	3e-11	3e-11	1e-11
FSSI	3e-11	5e-11	2e-11	5e-11	3e-11	2e-11
LSDF	4e-11	2e-11	3e-11	5e-11	4e-11	2e-11
FS-Manifold	5e-11	3e-11	4e-11	2e-11	5e-11	4e-11
SFSS	4e-11	3e-11	4e-11	6e-11	3e-11	2e-11
S^2FS^2R	4e-11	3e-11	4e-11	5e-11	4e-11	3e-11

TABLE V

p-VALUE OF PAIR-WISED WILCOXON TEST WITH RESPECT TO 1%, 5%, 10%, 25%, 50%, AND 100% LABELED TRAINING DATA FOR GENE RECOGNITION

	1% labeled	5% labeled	10% labeled	25% labeled	50% labeled	100% labeled
All Features	4e-11	2e-11	5e-11	4e-11	3e-11	2e-11
FISHER	2e-11	3e-11	2e-11	4e-11	2e-11	3e-11
SPEC	4e-11	3e-11	5e-11	3e-11	2e-11	5e-11
FSNM	4e-11	2e-11	4e-11	3e-11	3e-11	4e-11
FSSI	3e-11	4e-11	5e-11	4e-11	3e-11	5e-11
LSDF	2e-11	5e-11	3e-11	4e-11	5e-11	3e-11
FS-Manifold	3e-11	5e-11	4e-11	5e-11	2e-11	4e-11
SFSS	3e-11	2e-11	4e-11	4e-11	2e-11	5e-11
S^2FS^2R	4e-11	3e-11	2e-11	4e-11	5e-11	3e-11

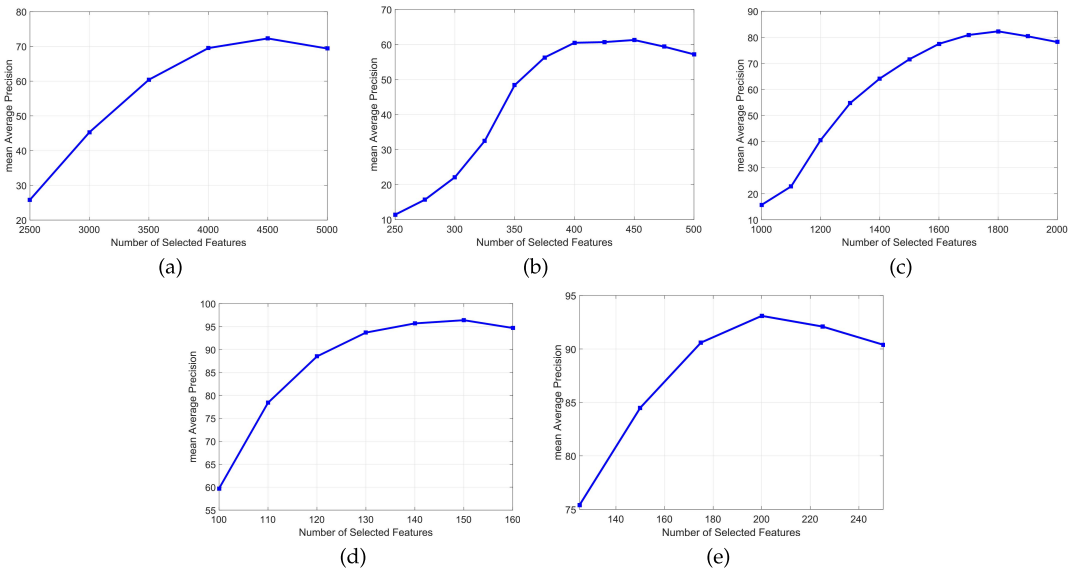


Fig. 7. Influence of selected feature number (mAP ± STD). Best view in color. (a) Video classification. (b) Image annotation. (c) Human motion recognition. (d) 3-D motion recognition. (e) Gene pattern recognition.

feature analysis. Then, we compare the results with the ones that are achieved by using all features and by using both labeled and unlabeled data. The comparisons are shown in Fig. 6. It can be seen that only using unlabeled training data still yields better results over using all the features, demonstrating the benefit of exploring unlabeled training data for feature analysis. This benefit is especially visible when few training exemplars are labeled, i.e.,, 1% and 5%.

I. Influence of Selected Feature Number

In this section, experiments are conducted to study the influence of the number of selected features on all the data sets. For space limitation, we only report the performance when 50% training data are labeled for all the data sets. The experimental results are shown in Fig. 7.

We have similar observations on all the data sets. It can be seen from the experimental results that mAP varies with

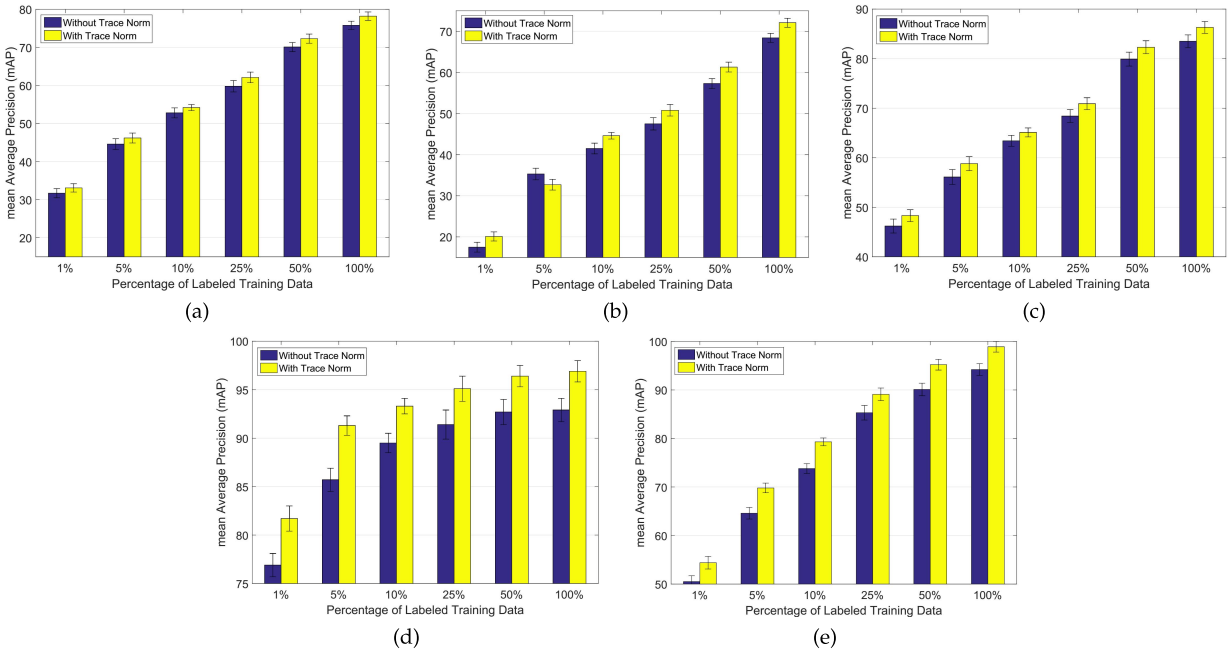


Fig. 8. Influence of trace norm (mAP ± STD) with respect to 1%, 5%, 10%, 25%, 50%, and 100% labeled training data. Best view in color. (a) Video classification. (b) Image annotation. (c) Human motion recognition. (d) 3-D motion recognition. (e) Gene pattern recognition.

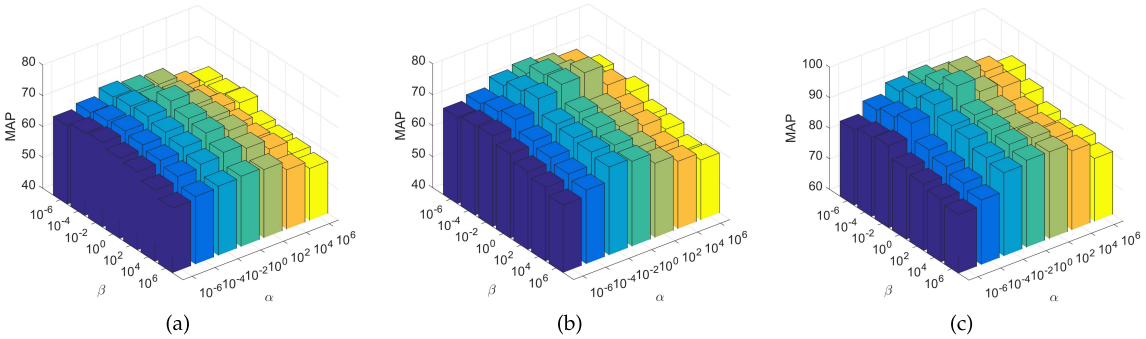


Fig. 9. Parameter sensitivity. (mAP ± STD). Best view in color. (a) Subject 1. (b) Subject 2. (c) Subject 3.

respect to the number of selected features. When we select features sequentially according to the feature importance, the performance (mAP) increases as the number of selected feature increases, indicating that the top ranked features are discriminate for subsequent tasks, i.e.,, classification. When mAP arrives at its peak, the performance becomes stable or even drops if we continue selecting the remaining features. This phenomenon indicates that the remaining features are redundant or even noisy. Based on the above observations, we conclude that it is necessary to select a subset of the features that best preserves the original feature space, thus facilitating subsequent analyzing tasks.

J. Influence of the Trace Norm

To study the influence of the trace norm, we compare the performance of the proposed objective function with the trace norm and without trace norm when 50% training data are labeled. We show the experimental results on all the data sets in Fig. 8.

It can be seen from the experimental results that the performance with trace norm outperforms the one without trace norm. This result indicates that mining information among multiple tasks by incorporating the trace norm can boost the performance of subsequent classification.

K. Parameter Sensitivity

We study the influences of the four parameters α, β, γ, and the number of selected features with 10% labeled training data. For space limitation, we only use the CCV data set in this experiment. First, we fix γ and the number of selected features at 1 and 3500, respectively, which are the median values of the tuned range of the parameters. The experimental results are shown in Fig. 9. It can be seen that the performance of our algorithm varies when the parameters (α and β) change. More specifically, MAP is higher when α and β are comparable. Then, α and β are fixed. Fig. 10 shows the parameter sensitivity results. Note that the shared information among multiple feature selection functions $\{W_1, \ldots, W_t\}$ by the parameter γ.

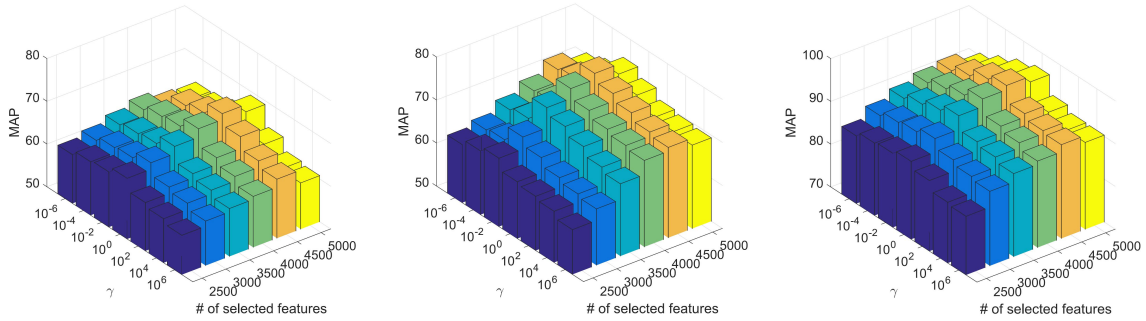


Fig. 10. Parameter sensitivity. (mAP ± STD). Best view in color. (a) Subject 1. (b) Subject 2. (c) Subject 3.

From this figure, we can see that mining correlations between multiple related tasks is beneficial to improve the performance. We can also notice that better performances are gained when the number of features is around 4000 and 4500.

V. Conclusion

In this paper, we have presented a new semisupervised feature analysis method. The proposed algorithm is able to mine correlations between different features and leverage shared information between multiple related tasks when a limited number of training data are labeled. Since the proposed objective function is nonsmooth and difficult to solve, we propose an iterative and efficient algorithm. To validate the effectiveness of the proposed algorithm, we apply it to different applications, including video classification, image annotation, human motion recognition, 3-D motion data analysis, and gene pattern recognition. The experimental results indicate that the proposed method outperforms the other compared algorithms for different applications.

References

[1] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. ICML*, 2003, pp. 1–8.

[2] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering," in *Proc. IDA*, 2005, pp. 440–451.

[3] S. HongYang and B.-G. Hu, "Discriminative feature selection by nonparametric Bayes error minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1422–1434, Aug. 2012.

[4] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, "Compound rank-k projections for bilinear analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1502–1513, Jul. 2015.

[5] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[6] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[7] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[8] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[9] J.-B. Yang and C.-J. Ong, "An effective feature selection method via mutual information estimation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1550–1559, Dec. 2012.

[10] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan, and A. G. Hauptmann, "Harnessing lab knowledge for real-world action recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 60–73, 2014.

[11] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[12] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2012.

[14] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. NIPS*, 2010, pp. 1–9.

[15] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

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

[17] A. Argyriou and T. Evgeniou, "Multi-task feature learning," in *Proc. NIPS*, 2007, pp. 1–8.

[18] S. Wang, X. Chang, X. Li, Q. Z. Shen, and W. Chen, "Multi-task support vector machines for feature selection with shared knowledge discovery," *Signal Process.*, vol. 120, pp. 746–753, Mar. 2016.

[19] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proc. AAAI*, 2010, pp. 673–678.

[20] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[21] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[22] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. IJCAI*, 2011, pp. 1589–1594.

[23] Z. Zheng and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. ICML*, 2007, pp. 1151–1157.

[24] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. TR 1530, 2006.

[25] R. G. F. Soares, H. Chen, and X. Yao, "Semisupervised classification with cluster regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1779–1792, Nov. 2012.

[26] F. Wang, "Semisupervised metric learning by maximizing constraint margin," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 4, pp. 931–939, Aug. 2011.

[27] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang, "Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 12, pp. 1553–1566, Dec. 2004.

[28] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. CVPR*, 2010, pp. 902–909.

[29] X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semi-supervised multi-label feature selection," in *Proc. AAAI*, 2014, pp. 1171–1177.

[30] F. Nie, D. Xu, X. Li, and S. Xiang, "Semisupervised dimensionality reduction and classification through virtual label regression," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 3, pp. 675–685, Jun. 2011.
[31] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM Multimedia*, 2009, pp. 175–184.
[32] Y. Liu, F. Nie, J. Wu, and L. Chen, "Efficient semi-supervised feature selection with noise insensitive trace ratio criterion," *Neurocomputing*, vol. 105, pp. 12–18, Apr. 2012.
[33] X. Chang, H. Shen, S. Wang, J. Liu, and X. Li, "Semi-supervised feature analysis for multimedia annotation by mining label correlation," in *Proc. PAKDD*, 2014, pp. 74–85.
[34] T. Jebara, J. Wang, and S.-F. Chang, "Graph construction and *b*-matching for semi-supervised learning," in *Proc. ICML*, 2009, pp. 441–448.
[35] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.
[36] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2014.
[37] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.
[38] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. ACM SIGKDD*, 2011, pp. 42–50.
[39] F. Wu, Y. Han, Q. Tian, and Y. Zhuang, "Multi-label boosting for image annotation by structural grouping sparsity," in *Proc. ACM MM*, 2010, pp. 15–24.
[40] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statist. Comput.*, vol. 20, no. 2, pp. 231–252, Apr. 2010.
[41] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.
[42] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen, "Semantic manifold learning for image retrieval," in *Proc. ACM Multimedia*, 2005, pp. 249–258.
[43] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ICMR*, 2011, Art. no. 29.
[44] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world Web image database from National University of Singapore," in *Proc. ACM Conf. Image Video Retr. (CIVR)*, Santorini, Greece, Jul. 2009, Art. no. 48.
[45] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. ICCV*, 2011, pp. 2556–2563.
[46] S. Leonid and M. J. Black, "HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion," Dept. Comput. Sci., Brown Univ., Providence, RI, USA, Tech. Rep. CS-06-08, 2006.
[47] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
[48] M. A. Shipp *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Med.*, vol. 8, no. 1, pp. 68–74, 2002.
[49] S. Ramaswamy *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci.*, vol. 98, no. 26, pp. 15149–15154, 2001.
[50] J. Zhao, K. Lu, and X. He, "Locality sensitive semi-supervised feature selection," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1842–1849, Jun. 2008.
[51] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.
[52] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 252–264, Feb. 2015.
[53] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, 2011, pp. 3169–3176.
[54] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. ICCV*, 2013, pp. 3551–3558.
[55] H. Ning, W. Xu, Y. Gong, and T. Huang, "Discriminative learning of visual words for 3D human pose estimation," in *Proc. CVPR*, 2008, pp. 1–8.

Xiaojun Chang is currently pursuing the Ph.D. degree with the Center for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW, Australia.

He has been a Visiting Student with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, since 2014. His current research interests include machine learning, data mining, and computer vision.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010.

He was a Post-Doctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Associate Professor with the University of Technology Sydney, Ultimo, NSW, Australia. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video content understanding.