Discriminating Joint Feature Analysis for Multimedia Data Understanding

Zhigang Ma, Feiping Nie, Yi Yang, Jasper Uijlings, Nicu Sebe, and Alexander G. Hauptmann

Abstract—In this paper, we propose a novel semi-supervised feature analyzing framework for multimedia data understanding and apply it to three different applications: image annotation, video concept detection and 3D motion data analysis. Our method is built upon two advancements of the state of the art: (1) $l_{2,1}$-norm regularized feature selection which can jointly select the most relevant features from all the data points. This feature selection approach was shown to be robust and efficient in literature as it considers the correlation between different features jointly when conducting feature selection; (2) manifold learning which analyzes the feature space by exploiting both labeled and unlabeled data. It is a widely used technique to extend many algorithms to semi-supervised scenarios for its capability of leveraging the manifold structure of multimedia data. The proposed method is able to learn a classifier for different applications by selecting the discriminating features closely related to the semantic concepts. The objective function of our method is non-smooth and difficult to solve, so we design an efficient iterative algorithm with fast convergence, thus making it applicable to practical applications. Extensive experiments on image annotation, video concept detection and 3D motion data analysis are performed on different real-world data sets to demonstrate the effectiveness of our algorithm.

Index Terms—Feature Analysis, Sparsity, Semi-supervised Learning, Image Annotation, Video Concept Detection, 3D Motion Data Analysis.

1 Introduction

The explosive increase of multimedia data, *i.e.*, text, image and video has brought the challenge of how to effectively index, retrieve and organize these resources. A common approach is to analyze the semantic concepts of multimedia data and to correlate concept labels with them for management tasks. Within the realm of multimedia data understanding, image and video concept understanding have obtained increasing research interest as both of them become prevalent with the popularity of the social web sites such as Flickr and YouTube. To effectively index, retrieve and manage these multimedia resources, it is necessary and beneficial to study concept analyzing techniques. Multimedia data are usually represented by different types of features. Previous works have shown that feature selection is able to reduce irrelevant and/or redundant information in the feature representation, thus facilitating subsequent analyzing tasks such as image annotation [1][2].

Existing feature selection algorithms are achieved by different means. For instance, classical feature selection algorithms such as Fisher Score [3] compute the weights of different features and then select features one by one. These classical algorithms generally evaluate the importance of each feature individually but neglect the useful information of the correlation between different features. Another problem is that they only use labeled training samples for feature selection, which have an excessive cost in human labor. Semi-supervised learning has shown to be an effective tool for saving labeling cost by using both labeled and unlabeled data. Motivated by this fact, semi-supervised feature selection has also been proposed. For example, in [4], Zhao *et al.* have presented an algorithm based on the spectral graph theory but similarly to Fisher Score [3], their method selects features one by one. To overcome the disadvantage of selecting features individually, a plethora of state of the art approaches such as [1][2][6] have been proposed to extract features jointly across all data points. Nonetheless, [1][2][6] implement their methods in a supervised way.

Our semi-supervised feature selection method combines the strengths of joint feature selection [6][1][7] and semi-supervised learning [8][9]. It utilizes both labeled and unlabeled data to select features while simultaneously consider the correlation between them. We name the proposed method Structural Feature Selection with Sparsity (SFSS).

In this paper, we apply our method to three different multimedia analyzing tasks, *i.e.*, image annotation, video concept detection and human action analysis from 3D motion data. Image annotation correlates labels that describe semantic concepts to images. It is basically a classification problem as it has to decide which classes an image may belong to. Annotation is realized by exploiting the correspondence between visual features and semantic concepts of the images. Video concept detection is another important tool for multimedia resource management. Similarly to image annotation, it aims to assign different concept labels to videos. We additionally apply SFSS

- *Z. Ma, J. Uijlings and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento. E-mail: ma, uijlings, sebe@disi.unitn.it*
- *F. Nie is with the Department of Computer Science and Engineering, University of Texas at Arlington. E-mail: feipingnie@gmail.com*
- *Y. Yang and A. Hauptmann are with the School of Computer Science, Carnegie Mellon University. E-mail: yiyang, alex@cs.cmu.edu*

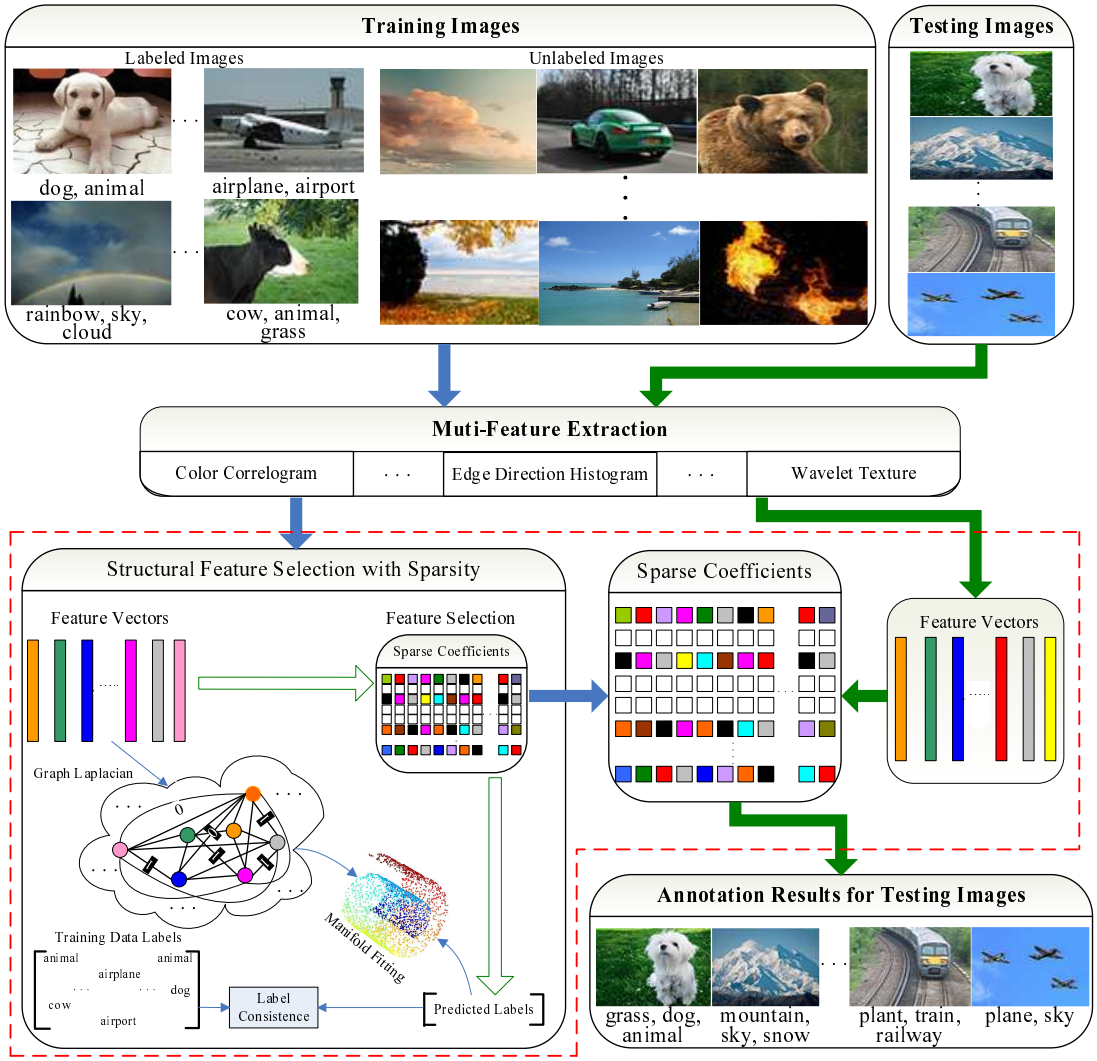


Figure 1. The general process of our method for image annotation. The red frame indicates the core part of our algorithm which analyzes the feature space for practical applications.

to human action analysis from 3D motion data.

Taking image annotation as an example, we illustrate the general analyzing process of our method in Figure 1. All the training and testing images are first represented by different types of features, followed by the graph Laplacian construction. Then sparse feature selection and label prediction are conducted simultaneously by satisfying both label consistence with the training data labels and manifold fitting on the data structure. The obtained sparse coefficients can be applied to the feature vectors for selection and be directly leveraged for classification.

The main contributions are as follows:

- We combine the recent advances of feature selection and semi-supervised learning into a single framework.
- The advantage of manifold learning, which is known to be effective in exploring relationship among multimedia data, is incorporated into our framework.
- We apply our method to different applications for which we show promising performance. Our method is especially competitive when few labeled samples are available.
- A fast iterative algorithm is proposed to solve our objective function.

2 Related Work

In this section, we briefly review the research on feature selection and semi-supervised learning.

2.1 Feature Selection

Feature selection is an effective tool in multimedia data understanding by selecting discriminating features and reducing the noise from the original data, resulting in more efficient and accurate multimedia analysis results.

In literature, there are many different feature selection algorithms. Some classical feature selection methods such as Fisher Score [3] evaluate the relevance of a feature according to the label distribution of the data. Although these classical methods have good performance when used in different applications, they have two major drawbacks. First, a lot of human

labor is consumed as they require all the training data to be labeled to exploit the correlation between features and labels for feature selection. Second, their computational cost is high as they evaluate features one by one.

To progress beyond these classical methods, researchers have proposed sparsity-based feature selection to extract features jointly [7][6][10][11], *i.e.*, each feature either has small scores or large scores over all data points, thus facilitating feature selection. Among various methods using this approach, $l_{2,1}$-norm regularization based algorithms have gained increasing interest for the sparsity, joint selection way and the ability to exploit the pairwise correlation among groups of features. For example, Zhao *et al.* use spectral regression with $l_{2,1}$-norm constraint to select features jointly and effectively remove redundant features in [7]. Nie *et al.* exploit joint $l_{2,1}$-norm minimization on both loss function and regularization for feature selection in [6]. Feature selection using $l_{2,1}$ models has shown its prominent performance. Therefore we propose to leverage it in our feature selection framework. However, the state of the art using $l_{2,1}$ models mostly conducts feature selection in a supervised scenario. Since in practice label information is expensive to obtain, we design our $l_{2,1}$-norm based feature selection in a semi-supervised way which can utilize both labeled data and unlabeled data.

2.2 Semi-supervised Learning

Semi-supervised learning is widely used in many applications with the appealing feature that it can use both labeled and unlabeled data [13]. The benefit of utilizing semi-supervised learning is that we can save human labor cost for labeling a large amount of data because it can exploit unlabeled data to learn the data structure. Thus, the human labeling cost and accuracy are both considered which gives semi-supervised learning a great potential to boost the learning performance when properly designed [12].

Among the different methods, graph Laplacian based semi-supervised learning has gained most research interest. Yang *et al.* have proposed a semi-supervised approach for cross media retrieval in [14]. In [8], Nie *et al.* have proposed a Flexible Manifold Embedding framework built upon graph Laplacian and demonstrated its advantage for dimensionality reduction over other state of the art semi-supervised algorithms. In [15], a new semi-supervised algorithm based on a robust Laplacian matrix is proposed for relevance feedback. Semi-supervised learning has proved to be able to bring in promising performance by leveraging the whole data distribution for multimedia data understanding in these previous works [14][8][15].

3 Methodology

In this section, we illustrate the detailed approach of our algorithm.

3.1 Problem Formulation

We aim to select features that are mostly related to the concepts of the training data. Suppose that $X \in \mathbb{R}^{d\times n}$ indicate the training data, $Y \in \mathbb{R}^{n\times c}$ are the labels accordingly. d is the dimension of the original feature, n is the number of the training data, and c is the number of concepts. We propose to use a projection matrix W to correlate X with Y for feature selection. As W is used to select features from the original feature space and it is expected to be related to the semantic concepts, W is a $d \times c$ matrix. The problem is subsequently to design an objective function to obtain W for feature selection. In our method, we propose to exploit the $l_{2,1}$-norm based sparse feature selection due to its efficacy shown in recent works. The $l_{2,1}$-norm based methods select features by exploiting the correlations between different features and select them jointly [7][6][10][11]. The boosted feature selection performance can consequently facilitate other applications. $l_{2,1}$-norm based algorithms can be generalized as the following objective function:

$$\min_{W} loss(W) + \gamma \left\|W\right\|_{2,1}, \tag{1}$$

where W is a projection matrix used for feature selection and $loss(W)$ is the loss function. γ is a regularization parameter. The definition of $\left\|W\right\|_{2,1}$ is:

$$\left\|W\right\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum\nolimits_{j=1}^{c} W_{ij}^2}. \tag{2}$$

The regularization term $\left\|W\right\|_{2,1}$ in the above function makes the optimal W sparse, according to [6][10]. As a result, W can be regarded as the combination coefficients for the most discriminative features to achieve feature selection.

Our goal is to design a robust loss function of Eq. (1) through which we obtain the W for feature selection. In literature, most works built upon Eq. (1), *e.g.*, [16][7][6], are realized through supervised learning. However, we want to incorporate semi-supervised learning into Eq. (1) as it is known to be an effective tool for saving cost while simultaneously maintaining or enhancing the learning performance when properly designed [12]. To this end, we propose to leverage semi-supervised learning by using the widely adopted graph Laplacian.

To begin with, we have following notations. $X = [x_1, x_2, ..., x_n]$ is the training data matrix where m data are labeled. $x_i \in \mathbb{R}^d (1 \le i \le n)$ is the i-th datum and n is the total number of the training data. $Y = [y_1, y_2, ..., y_m, y_{m+1}, ..., y_n]^T \in \{0, 1\}^{n\times c}$ is the label matrix and c indicates the class number. $y_i \in \mathbb{R}^c (1 \le i \le n)$ is the label vector with c classes. Y_{ij} denotes the j-th datum of y_i and $Y_{ij} = 1$ if x_i is in the j-th class, while $Y_{ij} = 0$ otherwise. If x_i is not labeled, y_i is set to a vector with all zeros, *i.e.*, $\forall i > m$, $y_i|_{i=(m+1)}^{n} = 0^{c\times 1}$.

A typical way to construct the graph Laplacian is as follows: First, we define a matrix G whose element G_{ij} weighs the similarity between x_i and x_j as

$$G_{ij} = \begin{cases} 1 & x_i \text{ and } x_j \text{ are } k \text{ nearest neighbors;} \\ 0 & \text{otherwise.} \end{cases}$$

In Eq. (3), we use the Euclidean distance to evaluate whether the two samples x_i and x_j are within the k nearest neighbors in the original feature space. Second, a diagonal matrix D is formulated with $D_{ii} = \sum_{j=1}^{n} G_{ij}$. Finally, the graph Laplacian L is constructed through $L = D - G$.

The graph Laplacian is the basis of semi-supervised learning. We further leverage Manifold Regularization [17] built upon the graph Laplacian to extend our framework to a semi-supervised scenario. Manifold Regularization is adopted because multimedia data has been normally shown to possess a manifold structure [18][19] and Manifold Regularization can explore it. Consequently, by applying Manifold Regularization to the loss function in Eq. (1) we obtain:

$$\arg\min_{W,b} Tr\left(W^T X L X^T W\right) + \mu \left\| X_l^T W + 1_n b^T - Y_l \right\|_{\mathrm{F}}^2 + \gamma \left\| W \right\|_{2,1}. \quad (3)$$

where $Tr(\cdot)$ denotes the trace operator. X_l and Y_l denote the labeled training data and their ground truth labels respectively. $b \in \mathbb{R}^c$ is the bias term and $1_n \in \mathbb{R}^n$ denotes a column vector with all its n elements being 1. μ and γ are regularization parameters.

As can be seen, the optimal W obtained from Eq. (3) is affected by the known ground truth labels Y_l. However, inspired by the transductive classification algorithm proposed in [20][13], we expect all the labels of the training data to contribute to the optimization of W. To achieve this goal, we denote a predicted label matrix as $F = [f_1, \ldots f_n]^T \in \mathbb{R}^{n \times c}$ for all the training data in X. $f_i \in \mathbb{R}^c (1 \leq i \leq n)$ is the predicted label vector of $x_i \in X$. According to [8], F should satisfy the smoothness on both the ground truth labels of the training data and the manifold structure. Hence, it can be obtained as follows [20][13]:

$$\arg\min_{F} Tr\left(F^T L F\right) + Tr\left((F-Y)^T U (F-Y)\right). \quad (4)$$

In the above function, we define a selecting diagonal matrix U whose diagonal element $U_{ii} = \infty$ if x_i is labeled and $U_{ii} = 1$ otherwise. This definition is to make the predicted labels F consistent with the ground truth labels Y. In practice, we can use a very large value, *e.g.* 10^{10} to approximate ∞.

Following the methodology in [8], we incorporate Eq. (4) into Eq. (3) and meanwhile consider all the training data with their labels (note that now we use X and F instead of X_l and Y_l respectively). Consequently, our objective function becomes:

$$\arg\min_{F,W,b} Tr\left(F^T L F\right) + Tr\left((F-Y)^T U (F-Y)\right) + \mu \left\| X^T W + 1_n b^T - F \right\|_{\mathrm{F}}^2 + \gamma \left\| W \right\|_{2,1}. \quad (5)$$

From Eq. (5) we can see that we are able to get F, W and b simultaneously. Additionally, the optimal W obtained through Eq. (5) can be utilized directly for classification as W selects the features most related to the class labels.

3.2 Solution

Our objective function involves the $l_{2,1}$-norm which is non-smooth. Hence, it is not straightforward to optimize it. We propose to solve the problem as follows.

By setting the derivative of Eq. (5) *w.r.t.* b to zero, we obtain:

$$b^T = \frac{1}{n}(1_n^T F - 1_n^T X^T W). \quad (6)$$

Substituting b^T in Eq. (5) with Eq. (6), the problem becomes:

$$\arg\min_{F,W} Tr\left(F^T L F\right) + Tr\left((F-Y)^T U (F-Y)\right) + \mu \left\| (I - \frac{1}{n} 1_n 1_n^T) X^T W - (I - \frac{1}{n} 1_n 1_n^T) F \right\|_F^2 + \gamma \left\| W \right\|_{2,1}, \quad (7)$$

where I is an identity matrix. Let H represent $I - \frac{1}{n} 1_n 1_n^T$, the objective becomes:

$$\arg\min_{F,W} Tr\left(F^T L F\right) + Tr\left((F-Y)^T U (F-Y)\right) + \mu \left\| H X^T W - H F \right\|_F^2 + \gamma \left\| W \right\|_{2,1}. \quad (8)$$

Note that $H = H^T = H^2$. By setting the derivative of Eq. (8) *w.r.t.* F to zero, we have:

$$F = PQ, \quad (9)$$

where $P = (L + U + \mu H)^{-1}$ and $Q = UY + \mu H X^T W$. Substituting F in Eq. (8) with Eq. (9), we arrive at:

$$\arg\min_{W} Tr\left(Q^T P^T (L+U) P Q - Q^T P^T U Y - Y^T U P Q + \mu W^T X H X^T W - \mu W^T X H P Q - \mu Q^T P^T H X^T W + \mu Q^T P^T H P Q\right) + \gamma \left\| W \right\|_{2,1}. \quad (10)$$

As $Tr(Q^T P^T U Y) = Tr(Y^T U P Q)$ and $Tr(\mu W^T X H P Q) = Tr(\mu Q^T P^T H X^T W)$, Eq. (10) becomes:

$$\arg\min_{W} Tr\left(Q^T P^T Q - 2 Q^T P^T Q + \mu W^T X H X^T W\right) + \gamma \left\| W \right\|_{2,1}.$$

By substituting $Q = UY + \mu H X^T W$ in the above function, we get:

$$\arg\min_{W} Tr\left(W^T (X H (\mu I - \mu^2 P) H X^T) W - 2\mu Y^T U P H X^T W\right) + \gamma \left\| W \right\|_{2,1}.$$

Denoting $A = XH(\mu I - \mu^2 P)HX^T$ and $B = \mu XHPUY$, the objective function becomes:

$$\arg\min_{W} Tr\left(W^T A W\right) - 2Tr\left(B^T W\right) + \gamma \left\| W \right\|_{2,1}. \quad (11)$$

3.3 Algorithm

Eq. (11) is a quadratic problem. First we have the following lemma to show that it is solvable.

Lemma 1: The objective of our framework is convex.

Proof: To prove *Lemma 1* is actually to prove that for any non-zero X, A defined in Eq. (11) is positive semi-definite. We therefore prove as follows:

$$\begin{aligned} A &= XH(\mu I - \mu^2 P)HX^T \\ &= \mu XHX^T - 2\mu^2 XHPHX^T + \mu^2 XHPP^{-1}PHX^T \\ &= \mu XHX^T - 2\mu^2 XHPHX^T \\ &+ \mu^2 XHP(L+U+\mu H)PHX^T \\ &= \mu\left((X^T - \mu PHX^T)^T H (X^T - \mu PHX^T) \right. \\ &+ \left. \mu XHP(L+U)PHX^T \right) \\ &= \mu\left(M^T H M + \mu X N X^T\right) \end{aligned} \quad (12)$$

where $M = X^T - \mu PHX^T$, $N = HP(L+U)PH$. As H and N are both larger than zero, we can easily draw the conclusion that $\mu M^T HM + \mu^2 XNX^T$ is greater than zero. Thus, $A = XH(\mu I - \mu^2 P)HX^T$ is positive semi-definite, demonstrating that the problem of our framework is convex. □

Algorithm 1: The optimization algorithm for SFSS.

Input:

The training data $X \in \mathbb{R}^{d\times n}$;
The training data labels $Y \in \mathbb{R}^{n\times c}$;
Parameters μ and γ.

Output:

Converged $W \in \mathbb{R}^{d\times c}$.

1: Construct the graph Laplacian matrix $L \in \mathbb{R}^{n\times n}$;
2: Compute the selecting matrix $U \in \mathbb{R}^{n\times n}$;
3: $H = I - \frac{1}{n}1_n 1_n^T$;
4: $P = (L + U + \mu H)^{-1}$;
5: $A = XH(\mu I - \mu^2 P)HX^T$;
6: $B = \mu XHPUY$;
7: Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d\times c}$ randomly;
8: **repeat**
Compute the diagonal matrix D_t as:
$D_t = \begin{bmatrix} 2\left\|w_t^1\right\|_2 & & \\ & \ldots & \\ & & 2\left\|w_t^d\right\|_2 \end{bmatrix}$;
Update W_{t+1} as: $W_{t+1} = (D_t A + \gamma I)^{-1} D_t B$;
$t = t + 1$.
until *Convergence*;
9: Return W.

To solve Eq. (11), we first reformulate it with the Lagrangian function as:

$$\mathcal{L}(W) = Tr\left(W^T AW\right) - 2Tr\left(B^T W\right) + \gamma \left\|W\right\|_{2,1}. \quad (13)$$

Denoting $W = [w^1, ..., w^d]^T$ with w^i as its i-th row, we define a diagonal matrix D whose diagonal elements $D_{ii} = 2\left\|w^i\right\|_2$. Then by setting the derivative of Eq. (13) *w.r.t.* W to zero, we obtain:

$$\begin{aligned} &2AW - 2B + 2\gamma D^{-1} W = 0 \\ \Rightarrow\ &W = (A + \gamma D^{-1})^{-1} B = (DA + \gamma I)^{-1} DB. \end{aligned} \quad (14)$$

According to the mathematical deduction aforementioned, we propose an iterative approach to solve the problem in Eq. (11). The iterative algorithm is illustrated in Algorithm 1 and it converges. We briefly discuss the computational complexity. Computing the graph Laplacian is $\mathcal{O}(n^2)$. During the training, learning W involves calculating the inverse of a few matrices, among which the most complex part is $\mathcal{O}(n^3)$. Denote n_{te} as the number of testing data. Once we get W, it takes $c \times d \times n_{te}$ multiplications to predict the categories of the testing data. For large scale data sets $n_{te} \gg c$ and $n_{te} \gg d$. Thus, the classification complexity is approximately linear *w.r.t.* n_{te}, which is very efficient.

The convergence of Algorithm 1 can be proved following the work in [6][10][32].

4 Experiments

We evaluate our method on image annotation, video concept detection and 3D motion data analysis respectively. Additional analyzing experiments are also performed to assess the overall performance of our method. These include a parameter sensitivity study and a convergence study.

4.1 Compared Algorithms

To evaluate the advantage of our method for multimedia data understanding, we compare it with the following algorithms:

- Fisher Score (FISHER) [3]: a classical method. It selects the most discriminative features by evaluating the importance of each feature individually.
- Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR) [31]: a sparsity based state of the art method. It realizes sparse feature selection by using a Laplace prior.
- Group Lasso with Logistic Regression (GLLR) [1]: a recently proposed method based on a sparse model. It utilizes group lasso extended with logistic regression to select both sparse and discriminative groups of homogeneous features.
- Feature Selection via Joint $l_{2,1}$-Norms Minimization (FSNM) [6]: a recent sparse feature selection algorithm. It employs joint $l_{2,1}$-norm minimization on both loss function and regularization for joint feature selection.
- Semi-supervised Feature Selection via Spectral Analysis (sSelect) [4]: a semi-supervised feature selection method based on spectral analysis.
- Locality sensitive semi-supervised feature selection (LSDF) [5]: a semi-supervised approach based on two graph construction, *i.e.*, within-class graph and between-class graph.

We use the regularized least square regression for classification after FISHER, SBMLR, FSNM, sSelect and LSDF finish the feature selection. In contrast, GLLR and SFSS can learn the classifiers directly when performing feature selection.

Table 1 illustrates the different properties of each method used in our experiments.

Table 1
A brief comparison between the different methods.

Method	SS[a]	S[b]	J-FS[c]	I-FS[d]	One-Step[e]
FISHER [3]		√		√	
SBMLR [31]		√	√		
GLLR [1]		√	√		√
FSNM [6]		√	√		
sSelect [4]	√			√	
LSDF [5]	√			√	
SFSS	√		√		√

[a] semi-supervised.
[b] supervised.
[c] feature selection across all data points.
[d] feature selection one by one.
[e] simultaneous classifier learning.

4.2 Experimental Data Sets

4.2.1 Image Annotation

Three data sets, *i.e.*, Corel-5K [21][22], MSRA-MM [23] and NUS-WIDE [24] are used in our experiments. The following is a brief description of the three data sets.

Corel-5K: In our experiment, we use the standard data set used in [21][22]. Corel-5K consists of 5,000 images from 50 different categories. Three types of color features (color histogram, color moment, and color coherence) and three types of texture features (Tamura coarseness histogram, Tamura directionality, and MSRSAR texture) are used to represent the images.

MSRA-MM: The data set used in our experiments is a subset of the original MSRA-MM 2.0 data set, which includes 50,000 images related to 100 concepts. However, 7,734 images within it are not associated with any labels. We have removed these images and obtained a subset of 42,266 labeled images. Three feature types used in [1], namely Color Correlogram, Edge Direction Histogram and Wavelet Texture are combined in our experiments.

NUS-WIDE: It consists of 209,347 labeled real-world images collected from Flickr which are associated with 81 concepts. The images are also represented by the combination of Color Correlogram, Edge Direction Histogram and Wavelet Texture.

4.2.2 Video Concept detection

We choose the Kodak consumer video data set [25] and the CareMedia data set [26].

Kodak: It consists of 1,358 consumer video clips and 5,166 key-frames are extracted accordingly. Among these key-frames, 3590 ones are annotated. We use all the annotated key-frames belonging to 22 concepts in our experiments for video concept detection. Color Correlogram, Edge Direction Histogram and Wavelet Texture are used to represent the key-frames.

CareMedia: The video data set was collected by Carnegie Mellon University to provide useful statistics to help doctors' diagnosis and patients' heath status assessment. 15 geriatric patients' activities in public spaces were recorded in a nursing home [26]. We test the performance by annotating the following 5 concepts which are concerned with patients' detailed behaviors: Pose and/or Motor Action (*e.g.* Tremors), Positive (*e.g.* Smiles and Dancing), Physically Aggressive (*e.g.* Punching), Physically Non-aggressive (*e.g.* Eating), and Staff Activities (*e.g.* Feeding). The MoSIFT feature [27] is used to represent each video sequence. In this experiment, we use a subset consisting of 3913 video sequences recorded by one camera in the dining room.

4.2.3 3D Motion Data Analysis

We choose the HumanEva 3D motion database [28]. There are five types of actions, namely boxing, gesturing, jogging, walking and throw-catch performed by different subjects in this database. We randomly sample 10,000 data of two subjects (5,000 per subject) similarly to [29][30] in our experiment. The action of the two subjects is considered to be different. We simultaneously recognize the identities and actions, which comes to 10 semantic categories in total. Each action is encoded as a collection of 16 joint coordinates in 3D space, thus resulting in a 48 dimensional feature vector. On top of that, we compute the Joint Relative Features between different joints and get a feature vector with 120 dimensions. The two kinds of feature vectors are further combined to generate a 168 dimensional feature.

Table 2
The settings of the training sets.

	Size (n)	Labeled Percentage (m)[1]
Corel-5K	2500	2, 5, 10, 25, 50, 100
MSRA-MM	10000	1, 5, 10, 25, 50, 100
NUS-WIDE	10000	1, 5, 10, 25, 50, 100
Kodak	1000	2, 5, 10, 25, 50, 100
CareMedia	1000	1, 5, 10, 25, 50, 100
HumanEva	3000	1, 5, 10, 25, 50, 100

Table 3
Performance comparison of image annotation (MAP±Standard Deviation) when 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE) training data are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.090±0.008**	**0.047±0.002**	**0.065±0.002**
FISHER [3]	0.069±0.006	0.041±0.002	0.058±0.003
GLLR [1]	0.066±0.008	0.032±0.008	0.046±0.007
FSNM [6]	0.078±0.007	0.043±0.002	0.059±0.002
SBMLR [31]	0.052±0.004	0.040±0.002	0.056±0.003

4.3 Experimental Setup

First, a training set for each data set is generated randomly consisting of n samples, among which $m\%$ samples are labeled. The detailed settings are given in Table 2. The remaining data of each data set work as the corresponding testing set. We generate the training and testing sets 5 times and report the average results with standard deviation.

In the experiments, we have to tune two types of parameters. One is the parameter k that specifies the k nearest neighbors used to compute the graph Laplacian. We fix it at 15 following the setting in our previous work [32]. The other one is the regularization parameters which are represented as μ and γ in Eq. (5). We tune them from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and report the best results.

To evaluate the classification performance, we use Mean Average Precision (MAP) as the evaluation metric for its stability and discriminating capability.

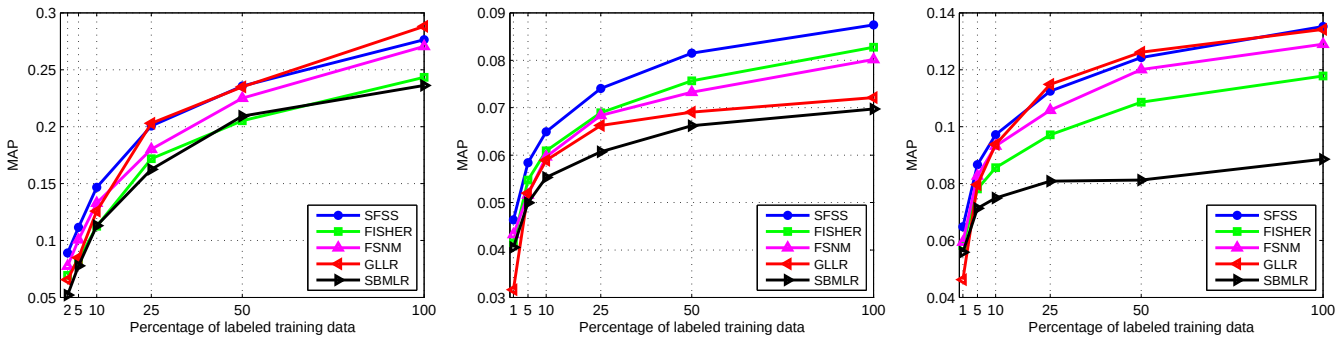
4.4 Multimedia Understanding Performance

In this section, we report the experimental results on image annotation, video concept detection and 3D motion data analysis respectively.

4.4.1 Image Annotation

Figure 2 shows the annotation results when different percentages of data are labeled. Table 3 to Table 5 show the results when 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE), 5% and 10% of the training data are labeled. We have the following observations from the experimental results: 1) As the

1. Note that the settings of the labeled training data on Corel-5K and Kodak are slightly different from others to guarantee that each concept class has at least one labeled training data.



(a) Corel-5K (b) MSRA-MM (c) NUS-WIDE

Figure 2. Performance comparison of image annotation *w.r.t.* the percentage of labeled training data. When 10% or less of the data are labeled our method outperforms all other algorithms. When 25% or more of the data are labeled, our method yields top performance or, on the MSRA-MM data set significantly better performance.

Table 4
Performance comparison of image annotation (MAP±Standard Deviation) when 5% training data are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.112±0.009**	**0.059±0.002**	**0.087±0.003**
FISHER [3]	0.083±0.007	0.055±0.002	0.078±0.002
GLLR [1]	0.085±0.010	0.052±0.001	0.079±0.001
FSNM [6]	0.101±0.007	0.051±0.002	0.082±0.002
SBMLR [31]	0.078±0.005	0.050±0.002	0.071±0.003

Table 5
Performance comparison of image annotation (MAP±Standard Deviation) when 10% training data are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.147±0.009**	**0.065±0.001**	**0.097±0.002**
FISHER[3]	0.113±0.003	0.061±0.002	0.086±0.003
GLLR [1]	0.126±0.015	0.059±0.001	0.094±0.002
FSNM [6]	0.133±0.009	0.060±0.001	0.093±0.003
SBMLR [31]	0.113±0.013	0.055±0.002	0.075±0.007

Table 6
Performance comparison of video concept detection (MAP±Standard Deviation) *w.r.t.* 2%, 5% and 10% labeled data on Kodak data set.

	2% labeled	5% labeled	10% labeled
SFSS	**0.259±0.015**	**0.303±0.023**	**0.346±0.027**
FISHER[3]	0.185±0.021	0.230±0.009	0.298±0.022
GLLR [1]	0.220±0.028	0.249±0.015	0.283±0.024
FSNM [6]	0.210±0.025	0.240±0.009	0.291±0.019
SBMLR [31]	0.189±0.029	0.222±0.009	0.269±0.026

Table 7
Performance comparison of video concept detection (MAP±Standard Deviation) *w.r.t.* 1%, 5% and 10% labeled data on CareMedia data set.

	1% labeled	5% labeled	10% labeled
SFSS	**0.257±0.018**	**0.293±0.009**	**0.301±0.014**
FISHER[3]	0.235±0.017	0.279±0.012	0.286±0.014
GLLR [1]	0.220±0.017	0.276±0.017	0.286±0.011
FSNM [6]	0.236±0.014	0.278±0.011	0.286±0.014
SBMLR [31]	0.202±0.003	0.227±0.004	0.249±0.007

number of labeled training data increases, the performance increases. 2) Our method is the only one which has consistently high scores on all three data sets. Other methods have varying degrees of success on each data set. 3) When 25% or more of the training data are labeled, our method is competitive with the best algorithms compared or better. Yet the more labeled data is available, the smaller our advantage is over other supervised algorithms. On the Corel-5K data set GLLR [1] slightly outperforms our method; on the NUS-WIDE data set our method is competitive with GLLR [1]; on the MSRA-MM data set our method outperforms all other methods. 4) Finally, when less than 25% of the data are labeled, our method consistently outperforms other methods on all three data sets. This is especially visible on the Corel-5K and MSRA-MM data sets.

4.4.2 *Video Concept Detection*

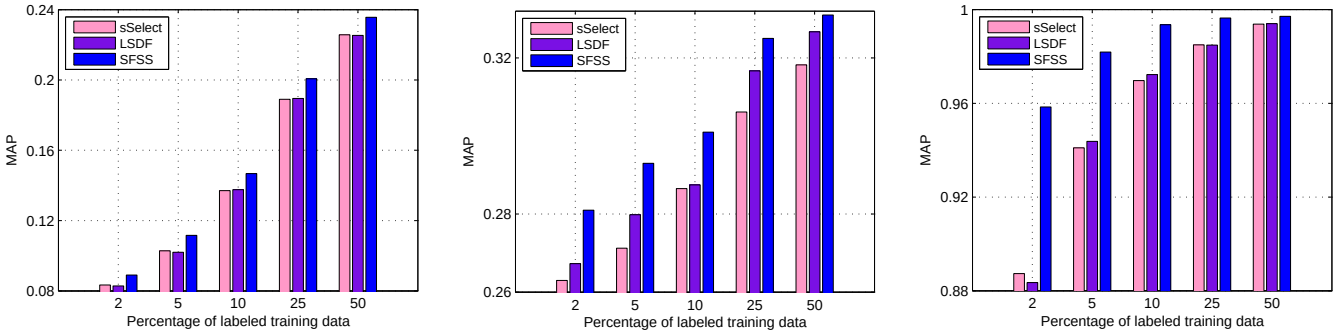
We illustrate the video concept detection results in Figure 3, Table 6 and Table 7. It can be seen from Figure 3 that our method has the top one performance over other algorithms. Table 6 and Table 7 give the detailed results when 2% or 1%, 5% and 10% training data are labeled. We observe that our method is especially competitive when few training data are labeled.

4.4.3 *3D Motion Data Analysis*

The results of 3D motion data analysis are illustrated in Table 8 and Figure 4. From Table 8 and Figure 4 we observe that our method gains huge advantage over other compared approaches. We also notice that SFSS gets satisfactory performance when only 5% training data are labeled and it shows nearly perfect performance (close to 1 in terms of MAP) when over 10% training data are labeled. Intuitively, this indicates that the exploitation of the manifold structure has contributed considerably to the whole analyzing performance.

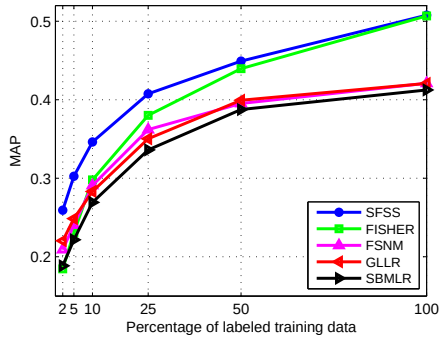
4.5 Comparison with Other Semi-supervised Feature Selection Methods

In this section, we compare SFSS with two state of the art semi-supervised feature selection algorithms, namely sSelect and LSDF. The experiments are conducted on Corel-5K,

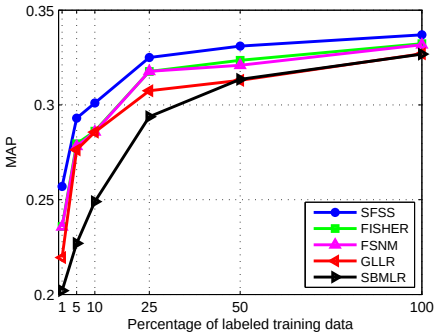


(a) Corel-5K (b) CareMedia (c) HumanEva

Figure 5. Performance comparison with semi-supervised approaches on different applications *w.r.t.* the percentage of labeled training data. Our method outperforms sSelect and LSDF for all settings and has much advantage when few training data (2% and 5%) are labeled.



(a) Kodak



(b) CareMedia

Figure 3. Performance comparison of video concept detection *w.r.t.* the percentage of labeled training data. Our method is consistently better than other compared methods.

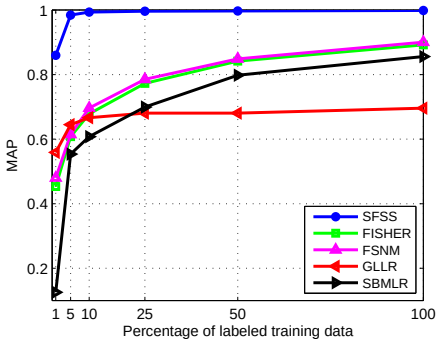


Figure 4. Performance comparison of 3D motion data analysis *w.r.t.* the percentage of labeled training data. Our method has much advantage over other algorithms. Good performance can be achieved even when very few training data are labeled.

Table 8
Performance comparison of 3D motion data analysis (MAP±Standard Deviation) *w.r.t.* 1%, 5% and 10% labeled data.

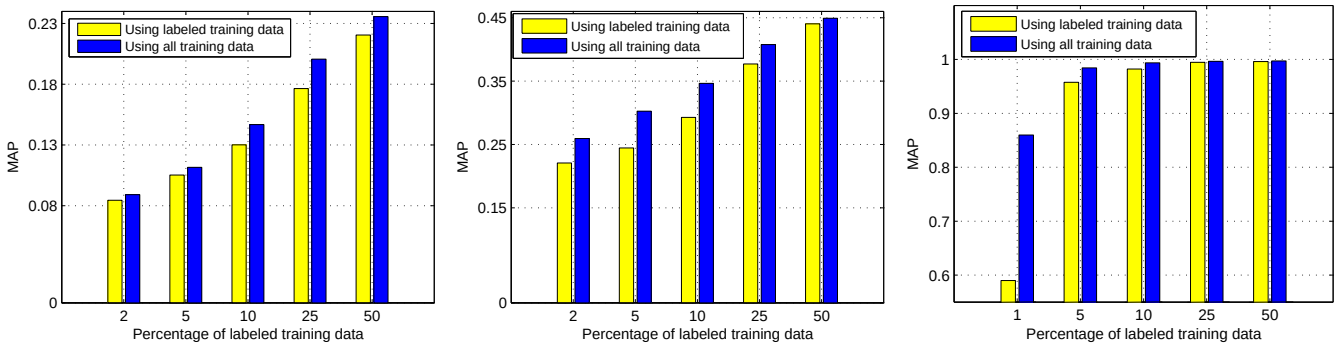
	1% labeled	5% labeled	10% labeled
SFSS	**0.860±0.021**	**0.984±0.015**	**0.994±0.012**
FISHER[3]	0.453±0.016	0.608±0.022	0.678±0.019
GLLR [1]	0.559±0.037	0.645±0.024	0.666±0.013
FSNM [6]	0.480±0.013	0.615±0.024	0.696±0.018
SBMLR [31]	0.126±0.055	0.554±0.022	0.608±0.024

CareMedia and HumanEva data sets for different applications. To be consistent, 2%, 5%, 10%, 25% and 50% training data are labeled in this experiment for all data sets. The results are shown in Figure 5. It can be observed that our method consistently outperforms both sSelect and LSDF. The advantage is especially visible when only few training data are labeled, *i.e.*, 2% or 5%. Semi-supervised methods are used for the cases when we only have limited number of labeled training data. We thus conclude that SFSS is much better than sSelect and LSDF as it has much higher accuracy when only few labeled training data are available.

4.6 Influence of the Unlabeled Data

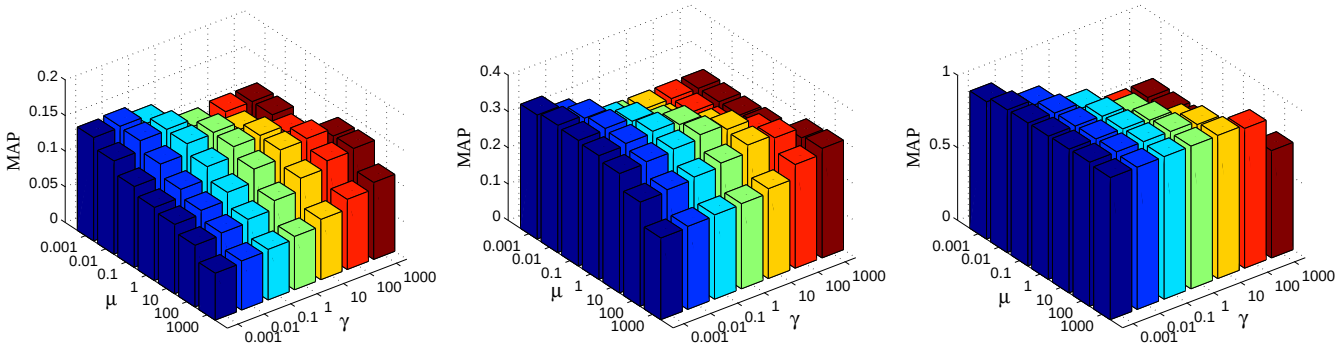
To study the influence of unlabeled training data on the multimedia understanding performance, we conduct an experiment correspondingly.

The unlabeled data in the training set are left out and we only use labeled training data to conduct feature analysis. Then we compare the results with the ones that are achieved by using the entire training set including both labeled and unlabeled data. The experiment is performed on Corel-5K, Kodak and HumanEva data sets for each application respectively. 2% (Corel-5K, Kodak) or 1% (HumanEva), 5%, 10%, 25% and 50% training data are labeled as different settings. Figure 6



(a) Corel-5K (b) Kodak (c) HumanEva

Figure 6. The influence of unlabeled data on different multimedia analyzing tasks. The blue bar stands for the performance of SFSS. The yellow bar indicates the results that are obtained by using only labeled data (no unlabeled data). The comparisons between the two approaches show that using unlabeled data improves the analyzing performance.



(a) Corel-5K (b) Kodak (c) HumanEva

Figure 7. Performance variance *w.r.t.* μ and γ. The figure displays different results when using different μ and γ.

illustrates the comparisons.

It can be seen that using unlabeled data besides the labeled data yields better results over using the labeled data alone. When 10% of the data are labeled, by also using unlabeled data we obtain relative improvements of 13% on the Corel-5K data set and 18% on the Kodak data set. Yet the situation is different for the HumanEva data set. The largest improvement, 45%, is obtained when only 1% of the data are labeled. However, as the percentage of labeled training data grows, the performance by using only labeled training data increases dramatically. The reason could be that the HumanEva data set is clean and easy to analyze. Moreover, the MAP closes in on 1 after 5% training data are labeled, which makes the contribution of the unlabeled data on the performance limited. The improvements in semi-supervised learning are due to the learning of the manifold structure. In theory, the more data points that one has, the better the manifold structure that can be learned. This saturates with enough data. The Corel-5K data set still has huge benefits from using all data instead of 50% for learning the manifold structure. For the HumanEva data set the manifold structure is very important as without this manifold the performance is much lower in general (see Figure 4). Figure 6(c) shows that this manifold is learned well using 25% of the data, after which performance is close to optimal for both the fully supervised and semi-supervised settings.

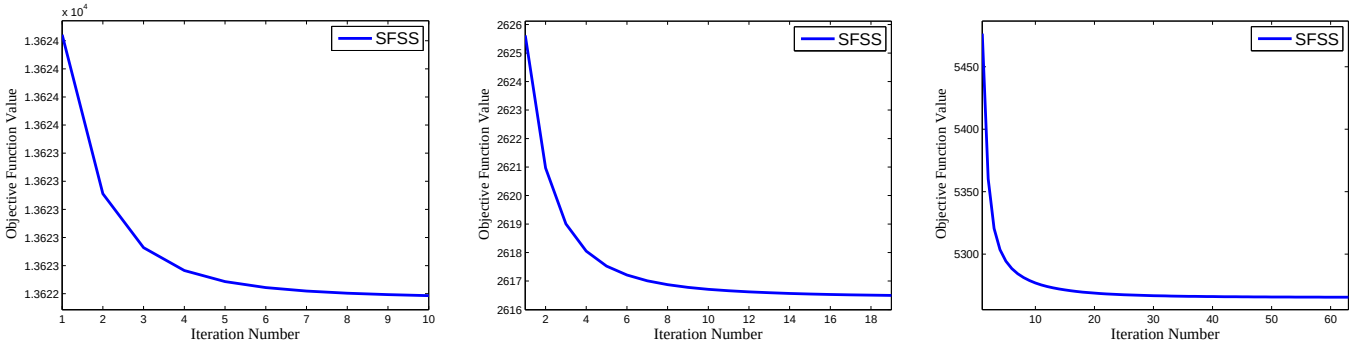
4.7 Parameter Sensitivity Study

In Figure 7, we show the influence of the two parameters μ and γ on the performance of different applications using Corel-5K, Kodak and HumanEva data sets when 10% training data are labeled. It can be seen that the MAP is generally higher when μ and γ are comparable for Corel-5K and Kodak data sets. In contrast, there is no analogous rule identifiable about when the optimal results are obtained for HumanEva data set. The phenomenon demonstrates that the parameter sensitivity is presumably related to the properties of the different data sets.

4.8 Convergence Study

In the previous section, we have proved that the objective function in Eq. (5) converges by using the proposed algorithm. For practical applications it is interesting how fast our algorithm converges.

Figure 8 shows the convergence curves of our optimization algorithm *w.r.t.* the objective function value in Eq. (5) on Corel-5K, Kodak and HumanEva when μ and γ are fixed at 1. It can be seen that our algorithm converges within as few as 10-20 iterations.



(a) Corel-5K (b) Kodak (c) HumanEva

Figure 8. Convergence curves of the objective function value in Eq. (5) using Algorithm 1. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

5 Conclusion

We have proposed a new multimedia analyzing method built upon feature analysis. The method takes advantage of joint feature selection with sparsity, manifold regularization and transductive classification. Additionally, to solve the non-smooth objective function of our algorithm, we have proposed an iterative approach. Our method is general and can be applied to different applications. In this paper, we evaluate its performance on image annotation, video concept detection and 3D motion data analysis. The experimental results have demonstrated that our method consistently outperforms the other compared algorithms for different analyzing tasks. Our method considers the characteristic of multimedia data, the labeling cost, the computational efficiency and the adaptability. Therefore, it is suitable for real-world multimedia understanding applications.

6 Acknowledgments

The work of Z. Ma, J. Uijlings, and N. Sebe was partially supported by the European Commission under the contract FP7-248984 GLOCAL. The work of Y. Yang and A. Hauptmann was partially supported by the National Science Foundation under Grant No. IIS-0917072, and by the National Institutes of Health (NIH) Grant No. 1RC1MH090021-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

[1] F. Wu, Y. Yuan, and Y. Zhuang. Heterogeneous Feature Selection by Group Lasso with Logistic Regression. In *ACM Multimedia*, 2010.

[2] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *ACM Multimedia*, 2010.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd ed.)*. Wiley-Interscience, New York, USA, 2001.

[4] Z. Zhao and H. Liu. Semi-supervised feature selection via spectral analysis. In *Proceedings of the SIAM International Conference on Data Mining*, 2007.

[5] J. Zhao, K. Lu and X. He. Locality sensitive semi-supervised feature selection. *Neurocomputing* 71:1842-1849, 2008.

[6] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and Robust Feature Selection via Joint L21-Norms Minimization. In *NIPS*, 2010.

[7] Z. Zhao, L. Wang, and H. Liu. Efficient Spectral Feature Selection with Minimum Redundancy. In *AAAI*, 2010.

[8] F. Nie, D. Xu, T. Hung and C. Zhang. Flexible Manifold Embedding: A Framework for Semi-supervised and Unsupervised Dimension Reduction. *IEEE Transactions on Image Processing*,19:1921-1932, 2010.

[9] V. Sindhwani, P. Niyogi and M. Belkin. Linear manifold regularization for large scale semi-supervised learning. In *Workshop on Learning with Partially Classified Training Data, International Conference on Machine Learning*, 2005.

[10] Y. Yang, H. Shen, Z. Ma, Z. Huang and X. Zhou. L21-Norm Regularized Discriminative Feature Selection for Unsupervised Learning. In *IJCAI*, 2011.

[11] Zhigang Ma and Feiping Nie and Yi Yang and Jasper Uijlings and Nicu Sebe. Web Image Annotation via Subspace-Sparsity Collaborated Feature Selection. *IEEE Transactions on Multimedia*, 2012.

[12] I. Cohen, F. Cozman, N. Sebe, M. Cirelo and T. Huang. Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,26(12):1553-1567, 2004.

[13] X. Zhu. Semi-supervised learning literature survey. In *Technical Report 1530, University of Wisconsin, Madison*, 2007.

[14] Y. Yang, D. Xu, F. Nie, J. Luo and Y. Zhuang. Ranking with Local Regression and Global Alignment for Cross Media Retrieval. In *ACM Multimedia*, 2009.

[15] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang and Y. Pan. A Multimedia Retrieval Framework based on Semi-Supervised Ranking and Relevance Feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,34(5):723-742, 2012.

[16] A. Argyriou, T. Evgeniou and M. Pontil. Multi-task feature learning. In *NIPS*, 2007.

[17] M. Belkin, P. Niyogi and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 12:2399-2434, 2006.

[18] Y. Yang, Y. Zhuang, F. Wu and Y. Pan. Harmonizing Hierarchical Manifolds for Multimedia Document Semantics Understanding and Cross-Media Retrieval. *IEEE Transactions on Multimedia*, 10(3):437-446, 2008.

[19] Y. Lin, T. Liu and H. Chen. Semantic Manifold Learning for Image Retrieval. In *ACM Multimedia*, 2005.

[20] X. Zhu, Z. Ghahramani and J. Lafferty. Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML*, 2003.

[21] S. Hoi, M. Lyu, and R. Jin. A Unified Log-based Relevance Feedback Scheme for Image Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):509-524, 2006.

[22] S. Hoi, W. Liu, and S. Chang. Semi-Supervised Distance Metric Learning for Collaborative Image Retrieval and Clustering. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(3):18:1-18:26, 2010.

[23] H. Li, M. Wang, and X. Hua. MSRA-MM 2.0: A Large-Scale Web Multimedia Dataset. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, 2006.

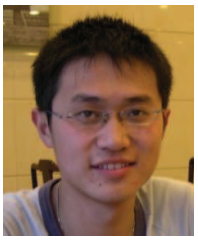
[24] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *CIVR*, 2009.

[25] A. Loui, J. Luo, S. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee and A. Yanagawa. Kodak's Consumer Video Benchmark Data Set: Concept

Definition and Annotation. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, 2007.
[26] http://www.informedia.cs.cmu.edu/caremedia/index.html
[27] M. Chen and A. Hauptmann. MoSIFT: Recognizing Human Actions in Surveillance Videos. In *Technical Report CMU-CS-09-161, Carnegie Mellon University*, 2009.
[28] L. Sigal, and M. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. In *Technical Report CS-06-08, Brown University, Department of Computer Science*, 2006.
[29] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang. Image Clustering using Local Discriminant Models and Global Integration. *IEEE Transactions on Image Processing*, 10:2761-2773, 2010.
[30] H. Ning, W. Xu, Y. Gong, and T. Huang. Discriminative Learning of Visual Words for 3D Human Pose Estimation. In *CVPR*, 2008.
[31] G. Cawley, N. Talbot, and M. Girolami. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In *NIPS*, 2006.
[32] Z. Ma, Y. Yang, F. Nie, J. Uijlings and N. Sebe. Exploiting the Entire Feature Space with Sparsity for Automatic Image Annotation. In *ACM Multimedia*, 2011.

Zhigang Ma received the B.S. and M.S. both from Zhejiang University, Hangzhou, China in 2004 and 2006 respectively, and is currently working toward the PhD degree from the University of Trento, Trento, Italy.

His research interests include machine learning and its application to computer vision and multimedia analysis.



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

He had been a postdoctoral research fellow at the University of Queensland from 2010 to May, 2011. After that, he joined Carnegie Mellon University. He is now a Postdoctoral Research Fellow at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g. multimedia indexing and retrieval, image annotation, video semantics understanding, etc.

Feiping Nie received the B.S. degree in Computer Science from North China University of Water Conservancy and Electric Power, Zhengzhou, China, in 2000, the M.S. degree in Computer Science from Lanzhou University, Lanzhou, China, in 2003, and the Ph.D. degree in Computer Science from Tsinghua University, Beijing, China, in 2009.

Currently, he is a Research Assistant Professor at the University of Texas, Arlington. His research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing and information retrieval.

Jasper R. R. Uijlings received the M.Sc. degree in artificial intelligence at the University of Amsterdam, The Netherlands, in 2006. In 2011 he received a Ph.D. degree at the ISIS Lab in the University of Amsterdam on the topic of Object Recognition in Computer Vision.

Currently he is working as a Postdoctoral Research Fellow at the University of Trento, Italy. His research interests include Computer Vision, Image Retrieval, and statistical pattern recognition.

Nicu Sebe (M'01-SM'11) received the Ph.D. in computer science from Leiden University, Leiden, The Netherlands, in 2001.

Currently, he is with the Department of Information Engineering and Computer Science, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human-computer interaction in computer vision applications. He was involved in the organization of the major conferences and workshops addressing the computer vision and human-centered aspects of multimedia information retrieval, among which as a General Co-Chair of the IEEE Automatic Face and Gesture Recognition Conference, FG 2008, ACM International Conference on Image and Video Retrieval (CIVR) 2007 and 2010, and WIAMIS 2009 and as one of the initiators and a Program Co-Chair of the Human-Centered Multimedia track of the ACM Multimedia 2007 conference. He is the general chair of ACM Multimedia 2013 and was a program chair of ACM Multimedia 2011. He is a senior member of IEEE and of ACM.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, the degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991.

He is currently with the faculty of the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include several different areas: man-machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning. From 1984 to 1994, he worked on speech and machine translation, when he joined the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-on-demand applications.