Web Image Annotation via Subspace-Sparsity Collaborated Feature Selection

Zhigang Ma, Feiping Nie, Yi Yang, Jasper Uijlings, and Nicu Sebe, *Member, IEEE*

Abstract—The number of web images has been explosively growing due to the development of network and storage technology. These images make up a large amount of current multimedia data and are closely related to our daily life. To efficiently browse, retrieve and organize the web images, numerous approaches have been proposed. Since the semantic concepts of the images can be indicated by label information, automatic image annotation becomes one effective technique for image management tasks. Most existing annotation methods use image features that are often noisy and redundant. Hence, feature selection can be exploited for a more precise and compact representation of the images, thus improving the annotation performance. In this paper, we propose a novel feature selection method and apply it to automatic image annotation. There are two appealing properties of our method. First, it can jointly select the most relevant features from all the data points by using a sparsity-based model. Second, it can uncover the shared subspace of original features, which is beneficial for multi-label learning. To solve the objective function of our method, we propose an efficient iterative algorithm. Extensive experiments are performed on large image databases that are collected from the web. The experimental results together with the theoretical analysis have validated the effectiveness of our method for feature selection, thus demonstrating its feasibility of being applied to web image annotation.

Index Terms—Image Annotation, Supervised Learning, Sparse Feature Selection, Shared Subspace Uncovering.

1 Introduction

As digital cameras become very common gadgets in our daily life, we have witnessed an explosive growth of digital images. On the other hand, the popularity of many social networks such as Facebook and Flickr helps boost the sharing of these personal images on the web. In fact, digital images now take up a very large proportion of multimedia contents in the network and are utilized intensively with different purposes. However, it is not straightforward to effectively organize and access these web images because we are facing an overwhelmingly large amount of them. Aiming to manage the images efficiently, automatic image annotation has been proposed as an important technique in multimedia analysis. The key idea for image annotation is to correlate keywords or detailed text descriptions with images to facilitate image indexing, retrieval, organization and management.

The sheer amount of web images itself provides us free and rich image repository for research. Researchers have been developing many automatic image annotation methods by leveraging the web scale databases such as Flickr which consist of a large number of user-generated images annotated with user-defined tags [1]. Appearance-based annotation, which is one popular approach, is generally realized through two processes, namely searching and mining. Similar images of the unannotated images are first found out from the web scale databases through the searching process and then the mining process extracts annotation from the textual information of these retrieved similar images. Research work using this approach has demonstrated promising performance for automatic image annotation [2][3]. Appearance-based image annotation has its effectiveness, but a major problem is that it can be negatively affected when user-generated tags do not reflect the concepts precisely. Learning-based automatic annotation is another effective approach and has gained much research interest. This approach is dependent on certain amount of available annotated images as the training data to learn classifiers for image annotation. Many algorithms have been rendered using learning-based approach these years with varying degrees of success for multimedia semantic analysis [4][5][6][7][8]. Therefore, this paper focuses on exploiting learning based methods for image annotation.

Images are normally represented by multiple features, which can be quite different from each other [9]. As it is inevitable to bring in irrelevant and/or redundant information in the feature representation, feature selection can be used to preprocess the data to facilitate subsequent image annotation task [11]. Hence, it is of great value to propose effective feature selection methods. Existing feature selection algorithms are achieved by different means. For instance, classical feature selection algorithms such as Fisher Score [12] compute the weights of different features, rank them accordingly and then select features one by one. These classical algorithms generally evaluate the importance of each feature individually and neglect the useful information of the correlation between different features. To overcome the disadvantage of selecting features individually,

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

researchers have proposed another approach which selects features jointly across all data points by taking into account the relationship of different features [11][13]. These methods have shown promising performance in different applications. In this paper we propose a feature selection technique which builds upon the latest mathematical advances in sparse, joint feature selection and apply this to automatic image annotation.

Image annotation is basically a classification problem. However, most web images are multi-labeled, that is to say, an image can reflect several semantic concepts. This intrinsic characteristic of web images makes it a complicated problem to classify them. A simple way to annotate multi-label images is to transform the problem to a couple of binary classification problems for each concept respectively. Though it is easy to implement, this approach neglects the correlation between different concept labels which is potentially useful. Therefore, many recent works [15] have proposed to exploit the shared subspace learning for multi-label tasks by incorporating the relational information of concept labels into multi-label learning. Inspired by their success, we apply shared subspace learning to the problem of feature selection.

To summarize, we combine the latest advances in joint, sparse feature selection with multi-label learning to create a novel feature selection technique which uncovers a feature subspace that is shared among classes. We name our method Sub-Feature Uncovering with Sparsity and demonstrate its effectiveness for automatic web image annotation. The main contributions of our work are:

- Our method leverages the prominent joint feature selection with sparsity, which can select the most discriminative features by exploiting the whole feature space.
- Our method considers the correlation between different concept labels to facilitate the feature selection.
- We conduct several experiments on large scale databases collected from the web. The results demonstrate the effectiveness of utilizing sparse feature selection and label correlation simultaneously.

The rest of this paper is organized as follows. We briefly introduce the state of the art on shared feature subspace uncovering, feature selection and automatic image annotation in section II. Then we elaborate the formulation of our method followed by the proposed solution in section III. We conduct extensive experiments in section IV to verify the advantage of our method for web image annotation. The conclusion is drawn in section V.

2 Related Work

Our work is geared towards better image annotation performance by exploiting effective feature selection. In this section, we briefly review the three related topics of our work, *i.e.*, shared feature subspace uncovering, feature selection and automatic image annotation.

2.1 Shared Feature Subspace Uncovering

Let x be a datum represented by a feature vector. The general goal of supervised learning is to predict for the input x an output y. To achieve this objective, learning algorithms usually use training data $\{(x_i, y_i)\}_{i=1}^{n}$ to learn a prediction function f that can correlate x with y. A common approach to obtain f is to minimize the following regularized empirical error:

$$\min_{f} \sum_{i=1}^{n} loss(f(x_i), y_i) + \mu\Omega(f), \tag{1}$$

where $loss\,(\cdot)$ is the loss function and $\mu\Omega(f)$ is the regularization with μ as its parameter.

It is reasonable to assume that multi-label images share certain common attributes. For example, a picture related to "parade", "people" and "street" share the component "people" with another picture related to "party", "people." Intuitively, we can leverage such label correlations for image annotation. In multi-label learning problems, Ando *et al.* assume that there is a shared subspace for the original feature space [17]. The concepts of an image are predicted by its vector representation in the original feature space together with the embedding in the shared subspace, which can be generalized as the following demonstration:

$$f(x) = v^T x + p^T Q^T x, \tag{2}$$

where v and p are the weight vectors and Q is a common subspace shared by all the features.

Suppose the images are related to c concepts in multi-label learning and there are m_t training data $\{x_i\}_{i=1}^{m_t}$ belonging to the t-th concept labeled as $\{y_i\}_{i=1}^{m_t}$. Then (1) can be redefined as:

$$\begin{aligned} \min_{f_t, Q} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} loss\left(f_t(x_i), y_i\right) + \mu\Omega(f_t) \right) \\ s.t. \quad Q^T Q = I \end{aligned} \tag{3}$$

Note that the constraint $Q^T Q = I$ in (3) is imposed to make the problem tractable.

By incorporating the shared feature subspace uncovering of (2) into (3), we get:

$$\begin{aligned} \min_{\{v_t, p_t\}, Q} \sum_{t=1}^{c} \left(\frac{1}{m_t} \sum_{i=1}^{m_t} loss\left(\left(v_t + Q p_t\right)^T x_i, y_i \right) + \mu\Omega(\{v_t, p_t\}) \right) \\ s.t. \quad Q^T Q = I \end{aligned} \tag{4}$$

Shared feature subspace learning has received increasing attention for its effectiveness on multi-label data [15]. Its theory has also been applied in multimedia analysis and proved its advantage. For instance, Amores *et al.* have leveraged the idea of sharing feature across multiple classes for object-class recognition and achieved prominent performance [18]. As a result, we adopt shared feature subspace uncovering in our feature selection framework and build our mathematical formulation on (4).

2.2 Feature Selection

Feature selection is widely adopted in many multimedia analysis applications. Its principle is to select the most discriminating features from the original ones while simultaneously eliminate the noise, thus resulting in better performance in practice. Another advantage of feature selection lies in its

attribute that it reduces the dimensionality of the original data, which in turn reduces the computational cost of the classification.

According to the availability of label information, feature selection algorithms can be classified into two groups: supervised and unsupervised. Unsupervised feature selection [19][20][21] is used when there is no label information. An effective way of unsupervised feature selection is to use the manifold structure of the whole feature set to select the most meaningful features [21].

In contrast, supervised feature selection is preferable when there is available label information that can be leveraged by using the correlation between features and labels. In the literature, plenty of supervised feature selection methods have been proposed. For example, Fisher Score [12] and ReliefF [22] are traditional supervised feature selection methods and are exploited widely in multimedia analysis. However, traditional feature selection usually neglects the correlation among different features [21]. Therefore, another approach has been developed recently, namely sparsity-based feature selection [23][13] which can exploit the feature correlation. This approach is built upon the comprehension that many real world data can be sparsely represented, thus rendering the possibility of searching the sparse representation of the data to realize feature selection. The $l_{2,1}$-norm regularization is known to be an effective model for sparse feature selection [24] and has drawn increasing attention [13][16].

The $l_{2,1}$-norm of an arbitrary matrix $W \in \mathbb{R}^{d\times c}$ is defined as:

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum\nolimits_{j=1}^{c} W_{ij}^2} \tag{5}$$

In [13] and [16], $l_{2,1}$-norm is leveraged to conduct feature selection jointly across the entire feature space with promising performance. Their works demonstrate that the $l_{2,1}$-norm of W makes W sparse, meaning that some of its rows shrink to zero. Consequently, W can be viewed as the combination coefficients for the most discriminative features. Feature selection is then realized by W where only the features associated with the non-zero rows in W are selected. Sparsity-based feature selection is efficient as it can select discriminative features jointly across all data points. However, few works have incorporated sparsity-based feature selection and shared feature subspace uncovering into one joint framework.

2.3 Automatic Image Annotation

Image annotation can be viewed as a classification task. It aims to correlate concept labels with specific images by classifying images to different classes. The ultimate goal is that the predicted labels via annotation algorithms can precisely reflect the real semantic contents of images. Nonetheless, the web image resources are countless so it is infeasible to annotate all of them manually. Hence, automatic image annotation becomes an essential tool for handling web scale images for retrieval, index and other management tasks.

Existing automatic image annotation methods have utilized a plethora of techniques [1][3][4][10][25]. Since images are usually represented by different features, much work [10][11][7] has focused on optimizing the feature selection process in their annotation frameworks. By finding the discriminative subset of original features and eliminating the noise, feature selection can help improve image annotation performance. For instance, Ma *et al.* have exploited a sparse selection model to select discriminative features that are closely related to image concepts for image annotation [7].

Thanks to the continuous effort made by researchers, we have witnessed great advance in automatic annotation for web images. However, the performance of automatic image annotation is yet to be satisfactory, thus requiring more research work in this domain. Inspired by the recent advanced techniques of feature selection and shared feature subspace uncovering, we propose a novel framework to extract the most discriminating features to boost the image annotation performance.

3 The Proposed Framework

In this section, we first illustrate the formulation of our Sub-Feature Uncovering with Sparsity (SFUS) framework. Then a detailed approach is rendered to solve the objective problem.

3.1 Problem Formulation

Our method roots from the shared feature subspace uncovering as given by (4).

Denote the training data matrix as $X = [x_1, x_2, ..., x_n]$ where $x_i \in \mathbb{R}^d (1 \leq i \leq n)$ is the i-th datum and n is the total number of the training data. Let $Y = [y_1, y_2, ..., y_n]^T \in \{0,1\}^{n\times c}$ be the label matrix. c stands for the class number and $y_i \in \mathbb{R}^c (1 \leq i \leq n)$ is the label vector with c classes. Denote $V = [v_1, v_2, ..., v_c] \in \mathbb{R}^{d\times c}$ and $P = [p_1, p_2, ..., p_c] \in \mathbb{R}^{sd\times c}$ where sd is the dimension of the shared subspace. We can then present (4) in a more compact way as:

$$\begin{aligned} &\min_{V,P,Q} loss\left((V+QP)^T X, Y\right) + \mu\Omega(V,P) \\ &s.t. \quad Q^T Q = I \end{aligned} \tag{6}$$

By defining $W = V + QP$ where $W \in \mathbb{R}^{d\times c}$, the above function equivalently becomes:

$$\begin{aligned} &\min_{W,V,P,Q} loss\left(W^T X, Y\right) + \mu\Omega(V,P) \\ &s.t. \quad Q^T Q = I \end{aligned} \tag{7}$$

It can be seen from the above function that by applying a different loss function and regularization, we can realize shared feature subspace uncovering in different ways. The least square loss has been widely used in research which can be illustrated as $\left\|X^T W - Y\right\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. By utilizing the least square loss, Ji *et al.* [15] have proposed to achieve shared subspace learning in the following way:

$$\begin{aligned} &\min_{W,P,Q} \left\|X^T W - Y\right\|_F^2 + \alpha\|W\|_F^2 + \beta\|W - QP\|_F^2 \\ &s.t. \quad Q^T Q = I \end{aligned} \tag{8}$$

In the above function, $\alpha\|W\|_F^2 + \beta\|W - QP\|_F^2$ is the regularization term. The first part regulates the information to

each specific label and the second part controls the complexity of the objective function. This approach is mathematically tractable and can be easily implemented. However, there are two issues worthy of further consideration. First, the least square loss is very sensitive to outliers, thus demanding a more robust loss function. Second, as we aim to conduct effective feature selection, it is advantageous to exert the sparse feature selection models on the regularization term. In [13], Nie *et al.* have proved that $l_{2,1}$-norm based models can handle both the aforementioned issues.

We therefore propose the following objective function as our foundation to realize feature selection:

$$\begin{aligned} &\arg\min_{W,P,Q} \left\|X^TW-Y\right\|_{2,1} + \alpha\left\|W\right\|_{2,1} + \beta\left\|W-QP\right\|_F^2 \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{9}$$

The loss function in our objective, that is to say, $\left\|X^TW-Y\right\|_{2,1}$ is robust to outliers as indicated in [13]. At the same time, $\left\|W\right\|_{2,1}$ in the regularization term guarantees that W is sparse to achieve feature selection across all data points [16][13].

3.2 Solution

As can be seen in (9), our problem involves the $l_{2,1}$-norm which is non-smooth and cannot be solved in a closed form. As a result, we propose to solve it as follows.

By denoting $X^TW-Y=[z^1,...,z^n]^T$ and $W=[w^1,...,w^d]^T$, the objective in (9) is equivalent to:

$$\begin{aligned} &\arg\min_{W,P,Q} Tr\left((X^TW-Y)^T\tilde{D}(X^TW-Y)\right) \\ &+\alpha Tr\left(W^TDW\right)+\beta\left\|W-QP\right\|_F^2 \\ &s.t. \quad Q^TQ=I, \end{aligned} \tag{10}$$

where $\tilde{D}$ and D are two matrices with their diagonal elements $\tilde{D}_{ii}=\frac{1}{2\|z^i\|_2}$ and $D_{ii}=\frac{1}{2\|w^i\|_2}$ respectively.

Note that for any arbitrary matrix A, $\left\|A\right\|_F^2=Tr\left(A^TA\right)$. Thus, (10) becomes:

$$\begin{aligned} &\arg\min_{W,P,Q} Tr\left((X^TW-Y)^T\tilde{D}(X^TW-Y)\right) \\ &+\alpha Tr\left(W^TDW\right)+\beta Tr\left((W-QP)^T(W-QP)\right) \\ &s.t. \quad Q^TQ=I, \end{aligned} \tag{11}$$

By setting the derivative of (11) *w.r.t* P to zero, we have:

$$\begin{aligned} &\beta(2Q^TQP-2Q^TW)=0 \\ \Rightarrow \quad &P=Q^TW \end{aligned} \tag{12}$$

Substituting P in (11) with (12) we have:

$$\begin{aligned} &\arg\min_{W,Q} Tr\left((X^TW-Y)^T\tilde{D}(X^TW-Y)\right) \\ &+\alpha Tr\left(W^TDW\right)+\beta Tr\left((W-QQ^TW)^T(W-QQ^TW)\right) \\ \Rightarrow &\arg\min_{W,Q} Tr\left((X^TW-Y)^T\tilde{D}(X^TW-Y)\right) \\ &+\alpha Tr\left(W^TDW\right)+\beta Tr\left(W^T(I-QQ^T)(I-QQ^T)W\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{13}$$

Since $(I-QQ^T)(I-QQ^T)=(I-QQ^T)$, the problem becomes:

$$\begin{aligned} &\arg\min_{W,Q} Tr\left((X^TW-Y)^T\tilde{D}(X^TW-Y)\right) \\ &+Tr\left(W^T(\alpha D+\beta I-\beta QQ^T)W\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{14}$$

By setting the derivative of (14) *w.r.t* W to zero, we get:

$$\begin{aligned} &2X\tilde{D}X^TW-2X\tilde{D}Y+2(\alpha D+\beta I-\beta QQ^T)W=0 \\ \Rightarrow \quad &(X\tilde{D}X^T+\alpha D+\beta I-\beta QQ^T)W=X\tilde{D}Y \\ \Rightarrow \quad &W=(M-\beta QQ^T)^{-1}X\tilde{D}Y \\ \Rightarrow \quad &W=N^{-1}X\tilde{D}Y, \end{aligned} \tag{15}$$

where $M=X\tilde{D}X^T+\alpha D+\beta I$, $N=(M-\beta QQ^T)^{-1}$ and $N=N^T$.

Note that (14) can be rewritten as:

$$\begin{aligned} &\arg\min_{W,Q} Tr\left(W^TX\tilde{D}X^TW\right)-2Tr\left(W^TX\tilde{D}Y\right) \\ &+Tr\left(Y^T\tilde{D}Y\right)+Tr\left(W^T(\alpha D+\beta I-\beta QQ^T)W\right) \\ &\Rightarrow \arg\min_{W,Q} Tr\left(W^T(X\tilde{D}X^T+\alpha D+\beta I-\beta QQ^T)W\right) \\ &-2Tr\left(W^TX\tilde{D}Y\right)+Tr\left(Y^T\tilde{D}Y\right) \\ &\Rightarrow \arg\min_{W,Q} Tr\left(W^T(M-\beta QQ^T)W\right)-2Tr\left(W^TX\tilde{D}Y\right) \\ &+Tr\left(Y^T\tilde{D}Y\right) \\ &\Rightarrow \arg\min_{W,Q} Tr\left(W^TNW\right)-2Tr\left(W^TX\tilde{D}Y\right)+Tr\left(Y^T\tilde{D}Y\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{16}$$

By incorporating the W obtained with (15) into the above function, we have:

$$\begin{aligned} &\arg\min_{Q} Tr\left(Y^T\tilde{D}X^TN^{-1}NN^{-1}X\tilde{D}Y\right) \\ &-2Tr\left(Y^T\tilde{D}X^TN^{-1}X\tilde{D}Y\right)+Tr\left(Y^T\tilde{D}Y\right) \\ &\Rightarrow \arg\min_{Q} Tr\left(Y^T\tilde{D}Y\right)-Tr\left(Y^T\tilde{D}X^TN^{-1}X\tilde{D}Y\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{17}$$

The above problem is equivalent to the following one:

$$\begin{aligned} &\arg\max_{Q} Tr\left(Y^T\tilde{D}X^TN^{-1}X\tilde{D}Y\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{18}$$

According to Sherman-Woodbury-Morrison formula, $N^{-1}=(M-\beta QQ^T)^{-1}=M^{-1}+\beta M^{-1}Q(I-\beta Q^TM^{-1}Q)^{-1}Q^TM^{-1}$. Thus, (18) becomes:

$$\begin{aligned} &\arg\max_{Q} Tr\left(Y^T\tilde{D}X^TM^{-1}X\tilde{D}Y+\beta Y^T\tilde{D}X^TQ\right. \\ &\left.(I-\beta Q^TM^{-1}Q)^{-1}Q^TM^{-1}X\tilde{D}Y\right) \\ &s.t. \quad Q^TQ=I \end{aligned} \tag{19}$$

which is equivalent to:

$$\begin{aligned}
&\arg\max_{Q} Tr\left(Y^T\tilde{D}X^TM^{-1}Q(I-\beta Q^TM^{-1}Q)^{-1}\right.\\
&\left.Q^TM^{-1}X\tilde{D}Y\right)\\
&\Rightarrow \arg\max_{Q} Tr\left(Y^T\tilde{D}X^TM^{-1}Q(Q^TQ-\beta Q^TM^{-1}Q)^{-1}\right.\\
&\left.Q^TM^{-1}X\tilde{D}Y\right)\\
&\Rightarrow \arg\max_{Q} Tr\left(Y^T\tilde{D}X^TM^{-1}Q[Q^T(I-\beta M^{-1})Q]^{-1}\right.\\
&\left.Q^TM^{-1}X\tilde{D}Y\right)\\
&\qquad s.t.\quad Q^TQ=I
\end{aligned} \tag{20}$$

As for any arbitrary matrices A, B and C, $Tr(ABC) = Tr(BCA)$, the above function becomes:

$$\begin{aligned}
&\arg\max_{Q} Tr\left([Q^T(I-\beta M^{-1})Q]^{-1}Q^TM^{-1}X\tilde{D}Y\right.\\
&\left.Y^T\tilde{D}X^TM^{-1}Q\right)\\
&\Rightarrow \arg\max_{Q} Tr\left((Q^TAQ)^{-1}Q^TBQ\right)\\
&\qquad s.t.\quad Q^TQ=I,
\end{aligned} \tag{21}$$

where $A = I - \beta M^{-1}$ and $B = M^{-1}X\tilde{D}YY^T\tilde{D}X^TM^{-1}$.

Equation (21) can be easily solved by the eigen-decomposition of $A^{-1}B$. However, as the solving of Q requires the input of $\tilde{D}$ and D which are related to W, it is still not straightforward to get Q and W. To solve this problem, we propose an iterative approach demonstrated in Algorithm 1. The complexity of the proposed algorithm is briefly discussed as follows. The complexity of calculating the inverse of a few matrices is $\mathcal{O}(d^3)$. To obtain Q, we need to conduct eigen-decomposition of $A^{-1}B$, which is also $\mathcal{O}(d^3)$ in complexity.

The proposed iterative approach in Algorithm 1 can be verified to converge to the optimal W by the following theorem.

Theorem 1: The objective function value shown in (9) monotonically decreases in each iteration until convergence using the iterative approach in Algorithm 1.

Proof: See Appendix A. □

Algorithm 1: The algorithm for solving the SFUS objective function.

Input:
The training data $X \in \mathbb{R}^{d\times n}$;
The training data labels $Y \in \mathbb{R}^{n\times c}$;
Parameters α and β.

Output:
Optimized $W \in \mathbb{R}^{d\times c}$.

1: Set $t = 0$ and initialize $W_0 \in \mathbb{R}^{d\times c}$ randomly;
2: **repeat**
Compute $[z_t^1, ..., z_t^n]^T = X^TW_t - Y$;
Compute the diagonal matrix $\tilde{D}_t$ as:

$$\tilde{D}_t = \begin{bmatrix} \frac{1}{2\|z_t^1\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|z_t^n\|_2} \end{bmatrix};$$

Compute the diagonal matrix D_t as:

$$D_t = \begin{bmatrix} \frac{1}{2\|w_t^1\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|w_t^d\|_2} \end{bmatrix};$$

Compute $M_t = X\tilde{D}_tX^T + \alpha D_t + \beta I$;
Compute $A_t = I - \beta M_t^{-1}$;
Compute $B_t = M_t^{-1}X\tilde{D}_tYY^T\tilde{D}_tX^TM_t^{-1}$;
Obtain Q_t by the eigen-decomposition of $A_t^{-1}B_t$;
Update W_{t+1} according to (15);
$t = t + 1$.
until *Convergence*;
3: Return W.

4 EXPERIMENTS

To validate the efficacy of our method when applied to automatic image annotation, we conduct several experiments particularly on image databases that are collected from the web image resources.

4.1 Compared Methods

We compare our method with one baseline and several feature selection algorithms on automatic image annotation to understand how our method progresses towards better annotation performance. The compared methods are enumerated as follows.

- Using all features (All-Fea): our baseline. It means that we use the original data without feature selection for annotation.
- Fisher Score (F-score) [12]: 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) [14]: a sparsity based state of the art method. It realizes sparse feature selection by using a Laplace prior.
- Spectral feature selection (SPEC) [26]: a state of the art method using spectral regression. It selects features one by one by leveraging the work of spectral graph theory. The supervised implementation is used in our experiments for fair comparison.
- Group Lasso with Logistic Regression (GLRR) [11]: 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) [13]: a latest sparse feature selection algorithm. It employs joint $l_{2,1}$-norm minimization on both loss function and regularization for joint feature selection.

As our framework is expanded upon regularized least square regression, we use it as the classifier for all the compared approaches.

4.2 Image Databases

Web images cover almost all the concepts people are interested in, thus justifying their advantage to be used as research corpus

Table 1
Performance comparison (±Standard Deviation) when $10 \times c$ images work as training data.

Dataset	Criteria	All-Fea	F-score [12]	SBMLR [14]	SPEC [26]	FSNM [13]	GLRR [11]	SFUS
MSRA	MAP	0.062±0.001	0.060±0.002	0.056±0.002	0.058±0.001	0.061±0.002	0.060±0.001	**0.063±0.001**
	MicroAUC	0.840±0.001	0.861±0.005	0.869±0.003	0.852±0.002	0.875±0.002	0.846±0.001	**0.878±0.002**
	MacroAUC	0.655±0.006	0.655±0.003	0.643±0.006	0.650±0.004	0.658±0.006	0.653±0.005	**0.662±0.005**
NUS	MAP	0.081±0.002	0.080±0.002	0.072±0.008	0.078±0.002	0.092±0.001	0.082±0.002	**0.094±0.003**
	MicroAUC	0.842±0.003	0.851±0.003	0.871±0.005	0.847±0.003	0.869±0.002	0.853±0.002	**0.877±0.002**
	MacroAUC	0.726±0.003	0.728±0.004	0.718±0.028	0.722±0.003	0.753±0.002	0.732±0.003	**0.756±0.003**

Table 2
Performance comparison (±Standard Deviation) when $20 \times c$ images work as training data.

Dataset	Criteria	All-Fea	F-score [12]	SMBLR [14]	SPEC [26]	FSNM [13]	GLRR [11]	SFUS
MSRA	MAP	0.067±0.004	0.066±0.002	0.059±0.001	0.066±0.001	0.068±0.001	0.067±0.001	**0.070±0.001**
	MicroAUC	0.859±0.011	0.876±0.004	0.883±0.004	0.868±0.001	0.887±0.002	0.866±0.002	**0.888±0.002**
	MacroAUC	0.676±0.013	0.680±0.004	0.666±0.004	0.679±0.002	0.687±0.002	0.680±0.002	**0.690±0.002**
NUS	MAP	0.099±0.001	0.098±0.004	0.073±0.007	0.094±0.001	0.105±0.003	0.105±0.002	**0.108±0.002**
	MicroAUC	0.874±0.001	0.880±0.005	0.887±0.006	0.875±0.001	0.888±0.003	0.885±0.003	**0.891±0.003**
	MacroAUC	0.767±0.001	0.770±0.006	0.733±0.024	0.763±0.001	0.785±0.004	0.780±0.001	**0.789±0.003**

for automatic image annotation. For the sake of the study on multimedia analysis, researchers have also managed to collect and process the web images to create good image databases for experimental purpose.

In our experiments, we select two large scale databases which are both made up of web images. The first one is the MSRA-MM 2.0 database which was created by Microsoft Research Asia [27]. This database was collected from the web through a commercial search engine and consists of 50,000 images belonging to 100 concepts. However, 7,734 images of the original database are not associated with any labels, we thus have removed these images and obtained a subset of 42,266 labeled images. In 2009, the Lab for Media Search in National University of Singapore proposed another large scale image database, *i.e.*, NUS-WIDE where all images are from Flickr [28]. NUS-WIDE includes 269,000 real-world images. The very large amount of NUS-WIDE, from our perspective, can well validate the scalability of our framework for real world annotation tasks. Hence, we choose this database in our experiments as well. Nonetheless, 59,653 images within NUS-WIDE are unlabeled, we therefore have removed them and used the remaining 209,347 labeled images related to 81 concepts as experimental corpus.

Considering the computational efficiency, we combine three feature types, *i.e.*, Color Correlogram, Edge Direction Histogram and Wavelet Texture provided by the authors to represent the images of the two databases. As a consequence, the corresponding feature dimensions for MSRA-MM 2.0 and NUS-WIDE are 347 and 345 respectively [27][28].

4.3 Experiment Setup

The procedure of our experiments can be generalized as follows. We first randomly generate a training set comprised of $m \times c$ images for each database similarly to the experimental setting in [29]. The remaining images are used as testing sets. To understand the performance variation *w.r.t* the number of training data, we set m as 10 and 20 respectively and report the corresponding results. We generate the training and testing sets for 5 times and report the average results for fair comparison with other methods.

Note that our objective function in (9) involves two parameters α and β. We tune both of them from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and report the best results. The number of the selected features ranges from $\{100, 150, 200, 250, 300\}$ and we use the corresponding feature subset to represent the images. Then the regularized least square regression is applied as the classifier for image annotation.

To evaluate the annotation performance, we use three evaluation metrics, *i.e.*, Mean Average Precision (MAP), MicroAUC and MacroAUC which are all widely used for multi-label classification tasks [30][11][31][32].

4.4 Performance on Image Annotation

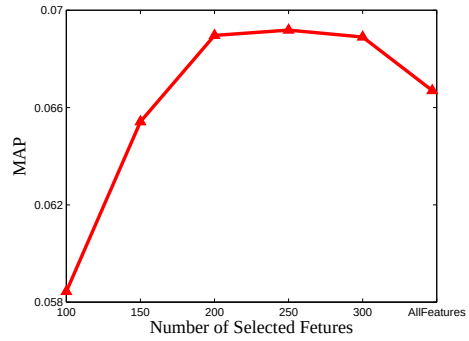
Table 1 and Table 2 show the annotation results when using $10 \times c$ and $20 \times c$ training data respectively. The results in bold indicate the best performance using the corresponding evaluation metric. According to the annotation results, we observe that our method demonstrates consistently superior performance on both databases.

Take MAP as an example. First, our method is better than All-Fea, *i.e.*, not using feature selection for annotation on both data sets. In particular, SFUS obtains notable improvement over All-Fea on NUS-WIDE. Second, our method has better annotation performance than the compared feature selection methods. Using $10 \times c$ training data, SFUS outperforms the second best feature selection method by about 2.6% and it is better than other feature selection algorithms for both data sets; using $20 \times c$ training data, SFUS is better than the second best feature selection method by about 1.6% and 3% on MSRA-MM 2.0 and NUS-WIDE respectively and it demonstrates good advantage over other algorithms. Hence, we conclude that our algorithm is a good feature selection mechanism for web image annotation.

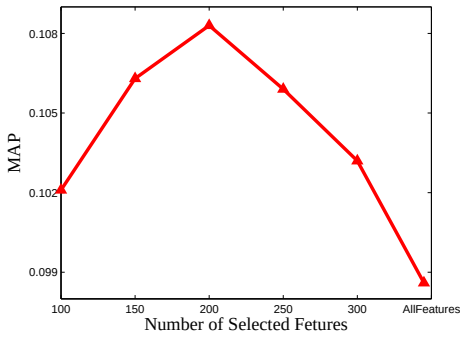
The good performance of SFUS for image annotation can be attributed to the appealing property that it can select features

Table 3
Performance comparison (±Standard Deviation) using Color Correlogram & Wavelet Texture on MSRA-MM when $10 \times c$ training data are labeled.

Criteria	All-Fea	F-score [12]	SBMLR [14]	SPEC [26]	FSNM [13]	GLRR [11]	SFUS
MAP	0.059±0.001	0.059±0.001	0.053±0.003	0.058±0.001	0.059±0.001	0.060±0.001	**0.061±0.001**
MicroAUC	0.848±0.002	0.861±0.006	0.874±0.004	0.854±0.003	0.872±0.002	0.858±0.002	**0.883±0.002**
MacroAUC	0.652±0.006	0.651±0.003	0.636±0.006	0.648±0.004	0.655±0.005	0.652±0.004	**0.659±0.005**

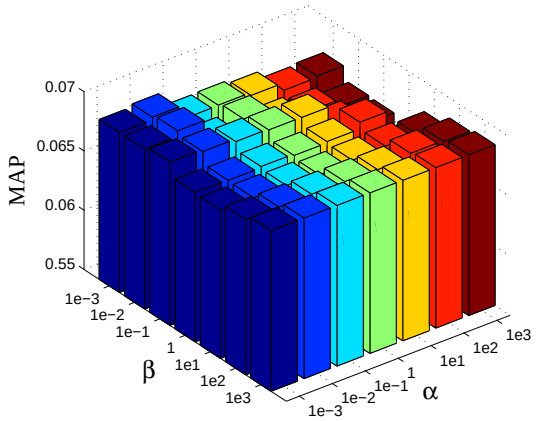


(a) MSRA-MM

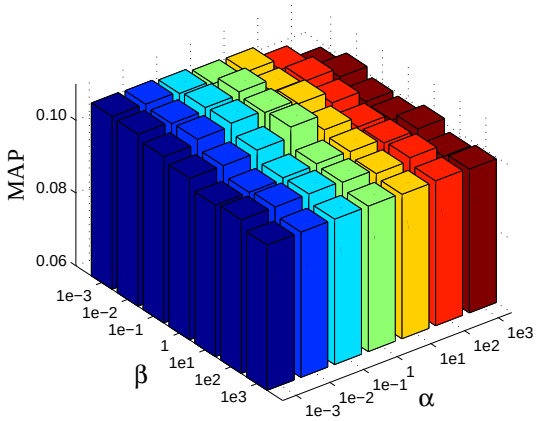


(b) NUS-WIDE

Figure 1. Performance variation *w.r.t* to the number of selected features using our feature selection algorithm.



(a) MAP-MSRA



(b) MAP-NUS

Figure 2. Performance variation *w.r.t* α and β when we fix the number of selected features at 200 for annotation. The figure shows different annotation results when using different values of α and β. With this setting, we get the best results when $\alpha = \beta = 10^{-2}$ for MSRA-MM 2.0 and when $\alpha = 1$ and $\beta = 10^{-2}$ for NUS-WIDE.

jointly across the whole feature space while simultaneously considering the correlation of multiple labels by exploring the shared feature subspace. The incorporation of the sparse model and shared subspace uncovering facilitates the feature selection by finding the most discriminative features, which can be used subsequently in annotation process.

4.5 Influence of Feature Type

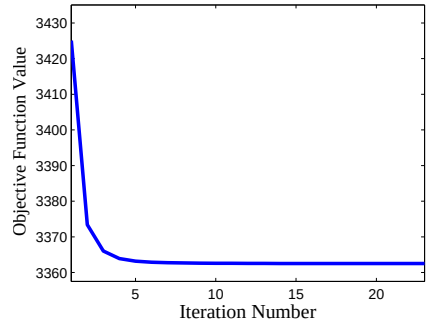
To evaluate the effectiveness of our method, we use a different original feature set, *i.e.*, only Color Correlogram and Wavelet Texture are combined to represent the images and we present the corresponding annotation results. The experiment is conducted on the MSRA-MM dataset with the results shown in Table 3.

It can be seen that our method still outperforms other feature selection algorithms when the images are represented by color histogram and wavelet texture. The results demonstrate that our algorithm is robust for the variance of the original feature set.

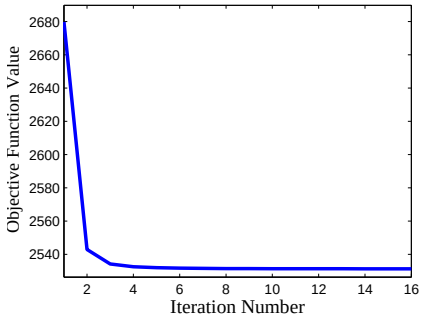
4.6 Influence of Selected Features

As feature selection is aimed at both accuracy and computational efficiency, we perform an experiment to study how the number of selected features can affect the annotation performance using $20 \times c$ training data. This experiment can present us the general trade-off between performance and computational efficiency for the two image databases.

Figure 1 shows the performance variation *w.r.t* the number of selected features in terms of MAP. We have the following observations: 1) When the number of selected features is too small, MAP is not competitive with using all features for annotation, which could be attributed to too much information



(a) MSRA-MM



(b) NUS-WIDE

Figure 3. Convergence curves of the objective function value in (9) using Algorithm 1. The figure shows that the objective function value monotonically decreases until convergence by applying the proposed algorithm.

loss. For instance, when using less than 150 features of MSRA-MM 2.0, MAP is worse than using all features for annotation. 2) MAP increases as the number of selected features increases up to 200. 3) MAP arrives at the peak level when using 200 features. 4) MAP keeps stable from using 200 features to using 300 features for MSRA-MM 2.0 while drops for NUS-WIDE. The different variance shown on the two datasets are supposed to be related to the properties of the datasets. 5) After all the features are selected, in other words, without feature selection, MAP is lower than selecting 200 features for MSRA-MM 2.0 and 100 features for NUS-WIDE. We conclude that, as MAP improves on both databases, our method reduces noise.

4.7 Parameter Sensitivity Study

Our method involves two regularization parameters, which are denoted as α and β in (9). To learn how they affect the feature selection and consequently the performance on image annotation, we conduct an experiment on the parameter sensitivity. Following the above experiment, we use $20 \times c$ training data for image annotation. MAP is used here to reflect the performance variation.

Figure 2 demonstrates the MAP variation *w.r.t* α and β on the two databases. From Figure 2 we notice that the annotation performance changes corresponding to different combinations of α and β. The impact of different values of the regularization parameters is supposed to be related to the trait of the database. On our experimental datasets, better results are generally obtained when α and β are comparable in value.

4.8 Convergence Study

As mentioned before, the proposed iterative approach monotonically decreases the objective function value in (9) until convergence. We conduct an experiment to validate our claim and to understand how the iterative approach works. Following the above experiments, we use $20 \times c$ training data in this experiment. The two parameters α and β are both fixed at 1 as that is the median value of the range from which the parameters are tuned.

Figure 3 shows the convergence curves of our algorithm according to the objective function value in (9). It can be observed that the objective function value converges quickly. We also calculate the convergence time which is 17.6 and 10.9 seconds for MSRA-MM 2.0 and NUS-WIDE respectively on a personal PC with Intel Core 2 Quad 2.83GHz CPU. The convergence experiment demonstrates the efficiency of our algorithm.

5 Conclusion

In this paper we have proposed a novel feature selection method and applied it to web image annotation. Our work integrates two state of the art innovations from shared feature subspace uncovering and joint feature selection with sparsity, thus endowing our method the following appealing properties. First, our method jointly selects the most discriminative features across the entire feature space. Additionally, our method considers the correlation between different labels, which has proved to be an effective way in multi-label learning tasks.

To validate the efficacy of our method for web image annotation, we conducted experiments on two popular image databases consisting of web images. It can be seen from the experimental results that our method outperforms classical and state of the art algorithms for image annotation. Therefore, we conclude that our method is a robust feature selection method and its feature subspace sharing foundation makes it particularly suitable for web images which are usually multi-labeled.

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 F. Nie was partially supported by the National Basic Research Program of China (2012CB316400). The work of Y. Yang was partially supported by the National Science Foundation under Grants No. IIS-0917072, and CNS-0751185. 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.

APPENDIX A
PROOF OF THEOREM 1

Proof: According to Algorithm 1, it can be inferred from (11) that:

$$W_{t+1} = \arg\min Tr\left((X^TW - Y)^T\tilde{D}(X^TW - Y)\right) + \alpha Tr\left(W^TDW\right) + \beta\left\|W - QP\right\|_F^2 \quad s.t. \quad Q^TQ = I$$

Therefore, we have

$$\begin{aligned}
&Tr\left((X^TW_{t+1} - Y)^T\tilde{D}_t(X^TW_{t+1} - Y)\right) + \alpha Tr\left(W_{t+1}^TD_tW_{t+1}\right) \\
&+\beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&\leq Tr\left((X^TW_t - Y)^T\tilde{D}_t(X^TW_t - Y)\right) + \alpha Tr\left(W_t^TD_tW_t\right) \\
&+\beta\left\|W_t - Q_tP_t\right\|_F^2 \\
&\Rightarrow \sum_{i=1}^{n}\frac{\left\|x_i^TW_{t+1} - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2} + \alpha\sum_{i=1}^{d}\frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \\
&+\beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&\leq \sum_{i=1}^{n}\frac{\left\|x_i^TW_t - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2} + \alpha\sum_{i=1}^{d}\frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} + \beta\left\|W_t - Q_tP_t\right\|_F^2 \\
&\Rightarrow \sum_{i=1}^{n}\left\|x_i^TW_{t+1} - y_i\right\|_2 - \sum_{i=1}^{n}\left\|x_i^TW_{t+1} - y_i\right\|_2 \\
&+\sum_{i=1}^{n}\frac{\left\|x_i^TW_{t+1} - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2} + \alpha\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 - \alpha\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 \\
&+\alpha\sum_{i=1}^{d}\frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} + \beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&\leq \sum_{i=1}^{n}\left\|x_i^TW_t - y_i\right\|_2 - \sum_{i=1}^{n}\left\|x_i^TW_t - y_i\right\|_2 + \sum_{i=1}^{n}\frac{\left\|x_i^TW_t - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2} \\
&+\alpha\sum_{i=1}^{d}\left\|w_t^i\right\|_2 - \alpha\sum_{i=1}^{d}\left\|w_t^i\right\|_2 + \alpha\sum_{i=1}^{d}\frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} + \beta\left\|W_t - Q_tP_t\right\|_F^2 \\
&\Rightarrow \sum_{i=1}^{n}\left\|x_i^TW_{t+1} - y_i\right\|_2 + \beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&+\alpha\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 - (\sum_{i=1}^{n}\left\|x_i^TW_{t+1} - y_i\right\|_2 - \sum_{i=1}^{n}\frac{\left\|x_i^TW_{t+1} - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2}) \\
&-\alpha(\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 - \sum_{i=1}^{d}\frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2}) \\
&\leq \sum_{i=1}^{n}\left\|x_i^TW_t - y_i\right\|_2 + \alpha\sum_{i=1}^{d}\left\|w_t^i\right\|_2 + \beta\left\|W_t - Q_tP_t\right\|_F^2 \\
&-(\sum_{i=1}^{n}\left\|x_i^TW_t - y_i\right\|_2 - \sum_{i=1}^{n}\frac{\left\|x_i^TW_t - y_i\right\|_2^2}{2\left\|x_i^TW_t - y_i\right\|_2}) \\
&-\alpha(\sum_{i=1}^{d}\left\|w_t^i\right\|_2 - \sum_{i=1}^{d}\frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2})
\end{aligned}$$

It has been shown in [13][16] that for any non-zero vectors $v_t^i|_{i=1}^r$:

$$\sum_i\left\|v_{t+1}^i\right\|_2 - \sum_i\frac{\left\|v_{t+1}^i\right\|_2^2}{2\left\|v_t^i\right\|_2} \leq \sum_i\left\|v_t^i\right\|_2 - \sum_i\frac{\left\|v_t^i\right\|_2^2}{2\left\|v_t^i\right\|_2}$$

where r is an arbitrary number. Thus, we can easily get the following inequality:

$$\begin{aligned}
&\sum_{i=1}^{n}\left\|x_i^TW_{t+1} - y_i\right\|_2 + \alpha\sum_{i=1}^{d}\left\|w_{t+1}^i\right\|_2 + \beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&\leq \sum_{i=1}^{n}\left\|x_i^TW_t - y_i\right\|_2 + \alpha\sum_{i=1}^{d}\left\|w_t^i\right\|_2 + \beta\left\|W_t - Q_tP_t\right\|_F^2 \\
&\Rightarrow \left\|X^TW_{t+1} - Y\right\|_{2,1} + \alpha\left\|W_{t+1}\right\|_{2,1} + \beta\left\|W_{t+1} - Q_{t+1}P_{t+1}\right\|_F^2 \\
&\leq \left\|X^TW_t - Y\right\|_{2,1} + \alpha\left\|W_t\right\|_{2,1} + \beta\left\|W_t - Q_tP_t\right\|_F^2
\end{aligned}$$

which indicates that the objective function value of (9) monotonically decreases until converging to the optimal W through the proposed approach in Algorithm 1. □

REFERENCES

[1] A. Ulges, M. Worring, and T. Breuel, Learning Visual Contexts for Image Annotation from Flickr Groups. *IEEE Transactions on Multimedia*, 13(2):330-341, 2009.

[2] X. Wang, L. Zhang, X. Li, and W. Ma, Annotating Images by Mining Image Search Results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1919-1932, 2008.

[3] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1-3):157-173, 2008.

[4] Y. Lu, and Q. Tian. Discriminant Subspace Analysis: an Adaptive Approach for Image Classification. *IEEE Transactions on Multimedia*, 11(7):1289-1300, 2009.

[5] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and crossmedia retrieval. *IEEE Transactions on Multimedia*, 10(3):437-446, 2008.

[6] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221-229, 2008.

[7] Z. Ma, Y. Yang, F. Nie, J. Uijlings, and N. Sebe. Exploiting the entire feature space with sparsity for automatic image annotation. In *ACM MM*, 2011.

[8] 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 on Pattern Analysis and Machine Intelligence*, 2011.

[9] Y. Yang, Y. Zhuang, D. Xu, Y. Pan, D. Tao, and S. Maybank. Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning. In *ACM MM*, 2009.

[10] Y. Gao, J. Fan, X. Xue, and R. Jain. Automatic Image Annotation by Incorporating Feature Hierarchy and Boosting to Scale up SVM Classifiers. In *ACM MM*, 2006.

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

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

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

[14] G. Cawley, N. Talbot, and M. Girolami. Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation. In *NIPS*, 2006.

[15] S. Ji, L. Tang, S. Yu, and J. Ye. A Shared-subspace Learning Framework for Multi-label Classification. *ACM Transactions on Knowledge Discovery from Data*, 2(1):8(1-29), 2010.

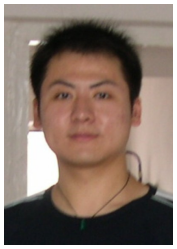
[16] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-Norm Regularized Discriminative Feature Selection for Unsupervised Learning. In *IJCAI*, 2011.

[17] R. Ando, and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817-1853, 2005.

[18] J. Amores, N. Sebe, and P. Radeva. Context-Based Object-Class Recognition and Retrieval by Generalized Correlograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1818-1833, 2007.

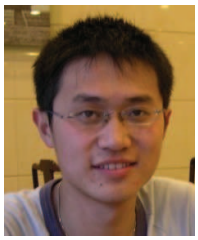
[19] M. Law, M. Figueiredo, and A. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154-1166, 2004.

[20] H. Wei, and S. Billings. Feature Subset Selection and Ranking for Data Dimensionality Reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):162-166, 2007.
[21] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD*, 2010.
[22] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *ECML*, 1994.
[23] B. Krishnapuram, A. Hartemink, L. Carin, and M. Figueiredo. A Bayesian approach to joint feature selection and classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1105-1111, 2004.
[24] Z. Zhao, L. Wang, and H. Liu. Efficient Spectral Feature Selection with Minimum Redundancy. In *AAAI*, 2010.
[25] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394-410, 2007.
[26] Z. Zhao, and H. Liu. Spectral Feature Selection for Supervised and Unsupervised Learning. In *ICML*, 2007.
[27] 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.
[28] T. 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.
[29] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang. Learning with L1-Graph for Image Analalysis. *IEEE Transactions on Image Processing*, 19(4):858-866, 2010.
[30] S. Nowak, A. Llorente, E. Motta, and S. Rueger. The effect of semantic relatedness measures on multi-label classification evaluation. In *ACM MIR*, 2010.
[31] M. Wang, X. Hua, J. Tang, and R. Hong. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE Transactions on Multimedia*, 11(3):465-476, 2009.
[32] Y. Han, F. Wu, J. Jia, Y. Zhuang, B. Yu. Multi-task Sparse Discriminant Analysis (MtSDA) with Overlapping Categories. In *AAAI*, 2010.



Zhigang Ma received his B.S. and M.S. both from Zhejiang University, China in 2004 and 2006 respectively. He is currently studying for his PhD degree at University of Trento, Italy. His research interest mainly includes machine learning and its application to computer vision and multimedia analysis.

Feiping Nie received his B.S. degree in Computer Science from North China University of Water Conservancy and Electric Power, China in 2000, received his M.S. degree in Computer Science from Lanzhou University, China in 2003, and received his Ph.D. degree in Computer Science from Tsinghua University, China in 2009. Currently, He is a research assistant professor at the University of Texas, Arlington, USA. His research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing and information retrieval.



Yi Yang received his Ph.D degree in Computer Science from Zhejiang University, in 2010. He had been a postdoctoral research fellow in the University of Queensland from 2010 to May, 2011. After that, he joined Carnegie Mellon University. He is now a postdoctoral research fellow in School of Computer Science at Carnegie Mellon University. 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.

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 Post-Doc at the University of Trento, Italy. His research interests include Computer Vision, Image Retrieval, and statistical pattern recognition.

Nicu Sebe 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 has served as the guest editor for several special issues in IEEE Computer, Computer Vision and Image Understanding, Image and Vision Computing, Multimedia Systems, and ACM TOMCCAP. He has been a visiting professor in Beckman Institute, University of Illinois at Urbana-Champaign and in the Electrical Engineering Department, Darmstadt University of Technology, Germany. He is the co-chair of the IEEE Computer Society Task Force on Human-centered Computing and is an associate editor of Machine Vision and Applications, Image and Vision Computing, Electronic Imaging and of Journal of Multimedia. He is a senior member of IEEE and of ACM.