Exploiting the Entire Feature Space with Sparsity for Automatic Image Annotation *

Zhigang Ma
DISI, University of Trento, Italy
ma@disi.unitn.it

Yi Yang
SCS, Carnegie Mellon University, USA
yiyang@cs.cmu.edu

Feiping Nie
CSE, University of Texas at Arlington, USA
feipingnie@gmail.com

Jasper Uijlings
DISI, University of Trento, Italy
uijlings@disi.unitn.it

Nicu Sebe
DISI, University of Trento, Italy
sebe@disi.unitn.it

ABSTRACT

The explosive growth of digital images requires effective methods to manage these images. Among various existing methods, automatic image annotation has proved to be an important technique for image management tasks, *e.g.*, image retrieval over large-scale image databases. Automatic image annotation has been widely studied during recent years and a considerable number of approaches have been proposed. However, the performance of these methods is yet to be satisfactory, thus demanding more effort on research of image annotation. In this paper, we propose a novel semi-supervised framework built upon feature selection for automatic image annotation. Our method aims to jointly select the most relevant features from all the data points by using a sparsity-based model and exploiting both labeled and unlabeled data to learn the manifold structure. Our framework is able to simultaneously learn a robust classifier for image annotation by selecting the discriminating features related to the semantic concepts. To solve the objective function of our framework, we propose an efficient iterative algorithm. Extensive experiments are performed on different real-world image datasets with the results demonstrating the promising performance of our framework for automatic image annotation.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Selection Process; I5.2 [**Design Methodology**]: Feature Evaluation and Selection; I4.10 [**Image Respresentation**]: Statistical

General Terms

Algorithms, Experimentation, Theory

*Area chair: Qi Tian

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

Keywords

Image Annotation, Semi-supervised Learning, Sparse Feature Selection, Manifold Learning

1. INTRODUCTION

With the popularity of digital cameras, an overwhelming number of digital images are present in our world. The advancement of computer network and storage technology, on the other hand, has led to many websites such as Flickr for people to share these images. Confronted with these countless images, we have to think about how to organize and utilize them effectively. Image annotation is one technique that can facilitate the management of these image resources by associating keywords or detailed text descriptions with images. Rich and accurate annotation is critical for efficient image indexing, retrieval, organization and management.

However, the sheer amount of images makes it infeasible to annotate all the images manually. Hence, automatic image annotation has received a lot of research interest.

Tagging is one popular way for image annotation by leveraging a great number of image resources from the Web-scale datasets such as Flickr. These user-generated images are annotated with user-defined tags and therefore can be used in the research on image annotation. The tagging approach for annotation is typically realized through two processes, searching and mining. The searching process aims to find the similar images of the unannotated images from the Web-scale datasets while the mining process extracts annotation from the textual information of these retrieved similar images. Research on this topic has already shown progress in automatic image annotation [23][19]. For instance, in [23], Wang *et al.*, have proposed a new image tagging system by leveraging the Web-scale images. The system first searches for similar images on the Web and then mines the images found to obtain their annotations. However, user-tagged images are potentially noisy because the tags may not reflect the concepts but something outside images, thus leading to degraded performance of image annotation.

Another effective way, namely, the labeling approach, has also been studied widely for automatic image annotation. Among the various labeling methods, learning-based automatic annotation has gained the most popularity. Learning-based methods [13][4][29] require pre-labeled samples as the training data to learn the models for image annotation. Since images are usually represented by different features, much work [9][25][24] has focused on optimizing the feature selection process in their annotation frameworks. By finding the discriminative subset of original features and eliminat-

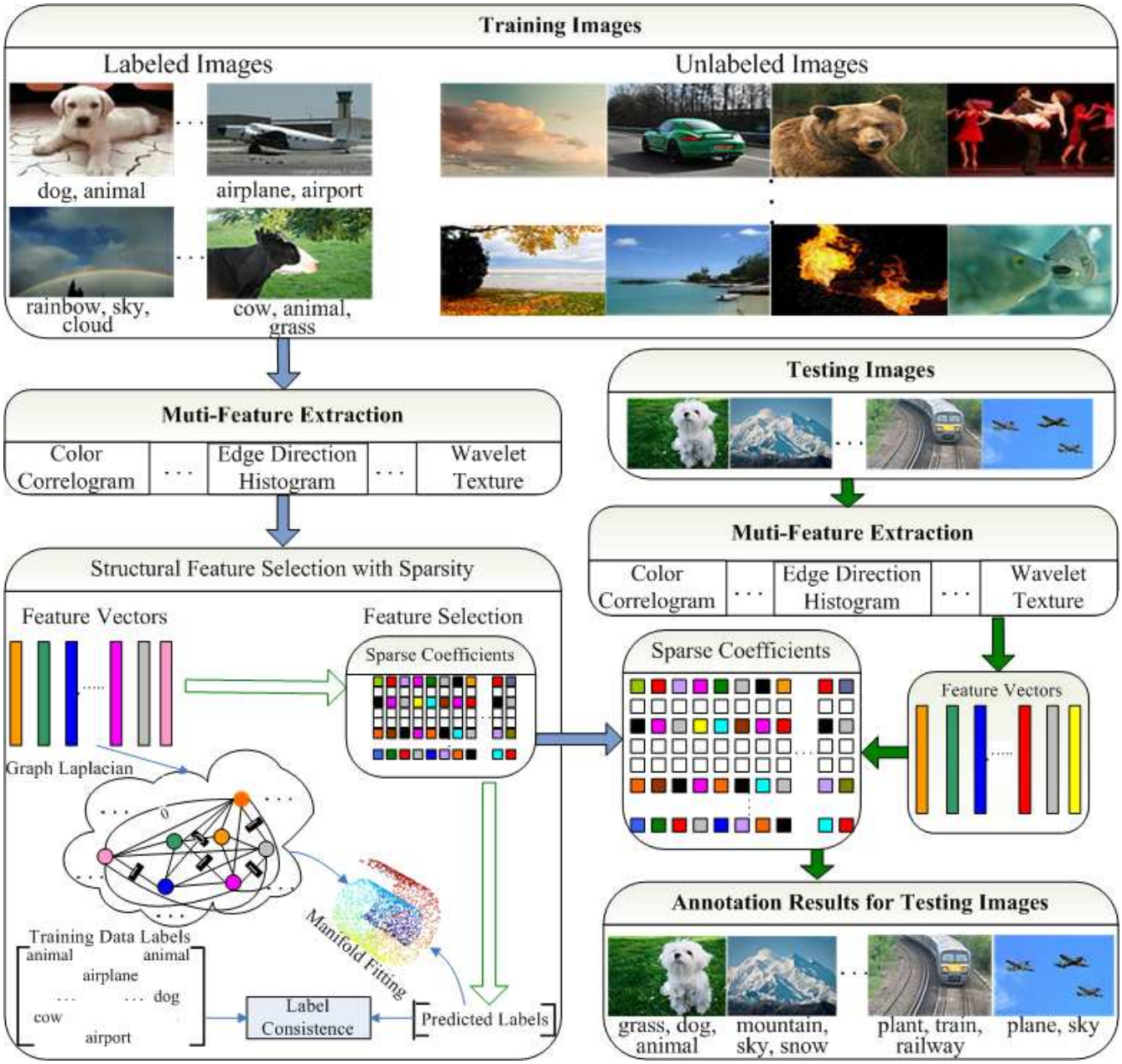


Figure 1: The illustration of our SFSS image annotation framework.

ing the noise, feature selection can help improve image annotation performance. On the other hand, a large body of feature selection algorithms are available to be utilized for image annotation. Yet these methods have certain drawbacks themselves. Classical feature selection algorithms such as Fisher Score [8] 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. Another problem is that they can only use labeled training samples for feature selection, which have excessive cost in human labor. In many practical applications, however, we only have limited labeled data and a large number of unlabeled data. Motivated by this fact, semi-supervised feature selection has been proposed. For example, in [31], Zhao *et al.* have presented an algorithm based on the spectral graph theory but similarly to Fisher Score [8], 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 [25][24][17] have proposed to extract features jointly across all data points. Nonetheless, [25][24][17] still implement their methods in a supervised way.

To progress beyond the current image annotation methods, we propose a novel semi-supervised annotation framework inspired by the recent work in feature selection [17][25][32] and semi-supervised learning [18][20]. As mentioned before, previous feature selection methods have not considered selecting feature jointly across all data points in the semi-supervised scenario. Therefore, our work aims to utilize both labeled and unlabeled data to select features

while simultaneously consider the correlation between them. We call the new method in our framework Structural Feature Selection with Sparsity (SFSS). Figure 1 illustrates how our method works for image annotation. All the training and testing images are first represented by feature vectors. The graph Laplacian is constructed based on both labeled and unlabeled training data. Then sparse feature selection and label prediction are conducted simultaneously by satisfying both label consistence with the training data labels and the manifold fitting on data structure. By selecting the most relevant features, the sparse coefficients can be applied to the feature vectors of the testing images for annotation. The main contributions of our work are as follows:

- Our method is a semi-supervised joint feature selection algorithm with sparsity, which can select the discriminative features by exploiting the whole feature space. Furthermore, our method can simultaneously learn a classifier for image annotation during the feature selection.
- The advantages of both manifold learning and joint feature selection are leveraged together, which is verified by the image annotation results of extensive experiments.
- Our method yields good results even if few labeled samples are available, which makes it attractive for real-world image datasets.

The rest of this paper is organized as follows. In section II, we review related work on feature selection, semi-supervised learning and automatic image annotation. We then illustrate the formulation of our framework and propose the solution and an algorithm for solving the objective function in section III. Experiments are given in section IV and section V concludes this paper.

2. RELATED WORK

Our work is geared towards more satisfactory image annotation performance. As feature selection and semi-supervised learning are both effective ways in multimedia analysis, we combine them in our framework for image annotation. In this section, we briefly review three related research topics, *i.e.*, feature selection, semi-supervised learning and automatic image annotation.

2.1 Feature Selection

Feature selection aims to extract important features and reduce the noise to better represent the original data. As a result, it can help increase accuracy. Furthermore, feature selection can boost computational efficiency on the classification side but also on the feature extraction side by only extracting those that are relevant for classification.

Traditional feature selection algorithms use fully labeled data and leverage the labels for feature selection. For instance, Fisher Score [8] evaluates the relevance of a feature according to the label distribution of the data. Previous works have demonstrated their good performance. However, as they evaluate features one by one the computational cost is very high. In contrast, a recently popular approach, *i.e.*, sparsity-based feature selection extracts features jointly across all data points [32][17][26]. This approach originated from the idea that the underlying representations of many real-world processes are often sparse. Hence, feature selection can be achieved by searching the sparse representation of the data. The most well-known sparse model is the $l1$-norm regularization. In [3], Cai *et al.* use an $l1$-regularized regression model to select features jointly instead of evaluating each feature independently. A family of works have also been rendered to extend the $l1$-norm regularization to the $l1/lq$-norm regularization which can better exploit the pairwise correlation among groups of features. In [32], Zhao *et al.* use spectral regression with $l_{2,1}$-norm constraint to evaluate features jointly and their method can effectively remove redundant features. In [17], Nie *et al.* also leverage joint $l_{2,1}$-norm minimization on both loss function and regularization for feature selection. These methods have shown to be prominent in feature selection and thus can facilitate clustering, classification, annotation and many other applications. Taking into account the benefits of sparse feature selection, we apply it to our framework as well. But different from the state of the art, we extend it to a semi-supervised way.

2.2 Semi-Supervised Learning

Semi-supervised learning uses both labeled and unlabeled data for different tasks in machine learning. The motive lies in the fact that labeled data are expensive to obtain while abundant unlabeled data are easy to acquire and helpful to improve the performance of the learning tasks. In [33], Zhu reviews different approaches for semi-supervised learning in the past. A major paradigm for semi-supervised learning is to construct graph to utilize manifold structure for learning process.

A considerable number of methods based on the graph Laplacian have been proposed so far. In [10], He has proposed to incrementally learn an adaptive subspace by preserving the semantic structure of the image space for image retrieval. Yang *et al.* have proposed a semi-supervised approach for cross media retrieval in [27]. In [22], Wang *et al.* have integrated multiple graphs into a regularization and optimization framework for video annotation. In [18], Nie *et al.* have proposed a Flexible Manifold Embedding framework which utilizes label information from labeled data together with a manifold structure from both labeled and unlabeled data, and applied their framework to dimensionality reduction and showed better results than other state of the art semi-supervised algorithms. From these previous works we learn that semi-supervised learning can leverage the whole data distribution, thus resulting in promising performance for different applications. As a result, we also incorporate the graph Laplacian based semi-supervised learning into our framework to facilitate the feature selection. The improvement of feature selection can consequently lead to better image annotation performance, which is the focused application in our work.

2.3 Automatic Image Annotation

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 usually realized by exploiting the correspondence between visual features and semantic concepts of the images. Since manual annotation is time-consuming and has excessive cost in human labor, the existing research works are mainly focused on automatic image annotation.

Present approaches for automatic image annotation are based on different theories. One approach is to use different probabilistic models to predict the semantic concepts for unlabeled images, *e.g.*, in [30] a probabilistic model connecting the visual features and the textual words has been proposed for automatic image annotation. However, using probabilistic models inevitably brings in parameter estimation, which is a complex process. Another approach for automatic image annotation is based on graph construction by mining the data structure. For example, in [16], Liu *et al.* have proposed to construct an adaptive similarity graph to exploit the data distribution to facilitate image annotation and the experimental re-

sults demonstrate the advantage of their method. Another widely adopted approach for automatic image annotation is applying classifiers to obtain the semantic concepts of images. Many works have been presented using this methodology. For instance, Gao *et al.* scale up SVM classifier by incorporating the feature hierarchy and boosting and then apply the classifier to image annotation in [9].

Researchers have made great progress in feature selection, semi-supervised learning and automatic image annotation respectively. However, few works have been rendered to incorporate the latest advanced techniques of feature selection and semi-supervised learning into one framework for image annotation. Zooming into details, we notice that the current semi-supervised feature selection algorithms usually select features one by one while joint feature selection generally belongs to either supervised or unsupervised genre. Meanwhile, the potential of applying feature selection to improve image annotation remains largely unexplored. To bridge the gap, in this paper we propose a unified framework for image annotation by leveraging both the state of the art feature selection and semi-supervised learning methodology.

3. METHODOLOGY

In this section, we first illustrate the formulation of our Structural Feature Selection with Sparsity (SFSS) framework. Then a detailed approach is rendered to solve the objective function of SFSS.

3.1 Formulation of Proposed Framework

Most traditional feature selection algorithms suffer from the limitation of evaluating features one by one, *i.e.*, they select features independently without considering the correlations between different features [8]. Aiming to exploit relational information among the features and select them jointly, recent works have applied $l_{2,1}$-norm based models to their algorithms [32][17][26]. These algorithms can select the most discriminating features with joint sparsity, thus boosting the feature selection performance. The algorithms can be generalized as to solve the following problem:

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

where W is a projection matrix used for feature selection and $loss(W)$ is the loss function. γ is a regularization parameter. Suppose $W \in \mathbb{R}^{d\times c}$, then its $l_{2,1}$-norm is defined as:

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

As shown in previous works [17][26], regularization using the $l_{2,1}$-norm of W makes the optimal W corresponding to (1) sparse, which means 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.

Various algorithms have been developed by applying different loss functions [1][32][17]. Despite the efficacy, these $l_{2,1}$-norm based feature selection algorithms generally work in the supervised way, in other words, they utilize labeled data to evaluate features. If very few labeled data are leveraged, it may induce over-fitting. On the contrary, if these methods use a large number of labeled data, the cost will be quite high.

Semi-supervised learning can exploit both the labeled data and unlabeled data, which makes it appropriate in practical applications where only a few labels are available. Besides, previous works have shown that semi-supervised learning has great potential to boost the learning performance when properly designed [7]. Considering the benefits of semi-supervised learning, we therefore extend the objective function in (1) to semi-supervised scenario.

To begin with, we briefly introduce the fundamental principle of the graph Laplacian, which is widely used in semi-supervised learning and adopted in our method.

Denote $X = [x_1, x_2, ..., x_n]$ as 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. Let $Y = [y_1, y_2, ..., y_m, y_{m+1}, ..., y_n]^T \in \{0,1\}^{n\times c}$ be the label matrix. c stands for the class number and $y_i \in \mathbb{R}^c (1 \le i \le n)$ is the label vector with c classes. Let Y_{ij} denote the j-th datum of y_i then $Y_{ij} = 1$ 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, namely, $\forall i > m$, $y_i|_{i=(m+1)}^{n} = 0^{c\times 1}$. Define a weight matrix G, whose element G_{ij} reflects the similarity between x_i and x_j as

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

The graph Laplacian is then constructed through $L = D - G$ where L is the graph Laplacian matrix and D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n} G_{ij}$.

Manifold Regularization [2] is the most well-known approach based on the graph Laplacian to extend many algorithms to semi-supervised way. Its popularity lies in the facts that multimedia data normally possess a manifold structure, which has been shown in many works [28][15] and that Manifold Regularization can explore such manifold structure. We therefore apply it to the loss function in (1) and get the following objective function:

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

A brief explanation for the above objective function is as follows. 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 parameters to balance the regression and the $l_{2,1}$-norm of W. Though the objective in (3) is semi-supervised, the features selected through it are merely best related to the known ground truth labels Y_l through the regression $\mu \left\| X_l^T W + 1_n b^T - Y_l \right\|_{\mathrm{F}}^2$. Following the recent transductive classification algorithm [34][33], we expect the selected features to accurately reflect all the labels of the training data where many of them are unlabeled. Therefore, we define 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. Note that $f_i \in \mathbb{R}^c (1 \le i \le n)$ is the predicted label vector of $x_i \in X$. F is supposed to be consistent with the ground truth labels of the training data and be smooth on the manifold structure so it can be optimized through the following objective function [34][33]:

$$\min_{F} \sum_{l=1}^{c} \left[\frac{1}{2} \sum_{i,j=1}^{n} (F_{il} - F_{jl})^2 G_{ij} + \sum_{i=1}^{n} U_{ii} (F_{il} - Y_{il})^2 \right], \tag{4}$$

where F_{il} is the l-th element of f_i and U is a diagonal matrix whose diagonal element $U_{ii} = \infty$ if x_i is labeled and $U_{ii} = 1$ otherwise. For later usage, we name U as a decision rule matrix for convenience. (4) can be rewritten as:

$$\min_{F} Tr\left(F^T L F\right) + Tr\left((F-Y)^T U (F-Y)\right), \tag{5}$$

where $Tr(\cdot)$ denotes the trace operator. L in the above function can also be constructed by some other sophisticated methods, *e.g.*, the one in [27] to exploit the manifold structure. The framework shown

in (5) has been widely applied to multimedia content analysis such as video annotation [21], cross media retrieval [27], *etc.*, where different methods were adopted to compute the matrix L.

Further inspired by the work in [18] which has proved the advantage of incorporating predicted labels F into Manifold Regularization, we propose our ultimate objective function as:

$$\begin{aligned}\arg\min_{F,W,b} & Tr\left(F^T LF\right) + Tr\left((F-Y)^T U(F-Y)\right) \\ & +\mu\left\|X^T W + 1_n b^T - F\right\|_{\mathrm{F}}^2 + \gamma\|W\|_{2,1}. \end{aligned} \tag{6}$$

Our framework can solve for F, W and b at the same time. Since W selects the features most related to the class labels, it can be used directly for image annotation. For any unseen images denoted by X', their label matrix F'=$X'^T W + 1_n b^T$.

It is worth mentioning that although (6) looks similar to the objective function $\arg\min_{F,W,b} Tr\left(F^T LF\right) + Tr\left((F-Y)^T U(F-Y)\right)$ $+\mu\left\|X^T W + 1_n b^T - F\right\|_{\mathrm{F}}^2 + \gamma\|W\|_{\mathrm{F}}^2$ in [18], the $l_{2,1}$-norm regularization $\|W\|_{2,1}$ in our framework makes it different from [18]. First of all, our framework is able to select the discriminative features across all the data points with joint sparsity. It can be applied to feature selection while the method in [18] is dedicated to classification and dimensionality reduction. Since our method is able to select a denoised and compact feature subset, it can enhance the accuracy as well as efficiency. On top of that, our framework further differs from [18] in the way to obtain W. In [18], W is calculated easily in a closed form. In contrast, the $l_{2,1}$-norm in our framework is non-smooth which makes it difficult to be solved. To deal with this problem, we propose an efficient algorithm to obtain W and elaborate it subsequently.

3.2 Solution

The objective problem of (6) can be solved as follows.

First, by setting the derivative of (6) *w.r.t* b to zero, we have:

$$\begin{aligned} & 2\mu(X^T W + 1_n b^T - F) = 0 \\ \Rightarrow\ & 1_n^T 1_n b^T = 1_n^T F - 1_n^T X^T W \\ \Rightarrow\ & b^T = \frac{1}{n}(1_n^T F - 1_n^T X^T W) \end{aligned} \tag{7}$$

Substituting b^T in (6) with (7), the problem becomes:

$$\begin{aligned} & \arg\min_{F,W} Tr\left(F^T LF\right) + Tr\left((F-Y)^T U(F-Y)\right) \\ & +\mu\left\|X^T W + \frac{1}{n}1_n 1_n^T F - \frac{1}{n}1_n 1_n^T X^T W - F\right\|_F^2 + \gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_{F,W} Tr\left(F^T LF\right) + Tr\left((F-Y)^T U(F-Y)\right) \\ & +\mu\left\|(I - \frac{1}{n}1_n 1_n^T)X^T W - (I - \frac{1}{n}1_n 1_n^T)F\right\|_F^2 + \gamma\|W\|_{2,1}, \end{aligned} \tag{8}$$

where I is an identity matrix. Using H to represent $I - \frac{1}{n}1_n 1_n^T$, we arrive at:

$$\begin{aligned} & \arg\min_{F,W} Tr\left(F^T LF\right) + Tr\left((F-Y)^T U(F-Y)\right) \\ & +\mu\left\|HX^T W - HF\right\|_F^2 + \gamma\|W\|_{2,1}. \end{aligned} \tag{9}$$

Note that $H = H^T = H^2$. By setting the derivative of (9) *w.r.t* F to zero, we have:

$$\begin{aligned} & 2LF + 2U(F-Y) - 2\mu H(HX^T W - HF) = 0 \\ \Rightarrow\ & (L + U + \mu H)F = UY + \mu HX^T W \\ \Rightarrow\ & F = PQ, \end{aligned} \tag{10}$$

where $P = (L+U+\mu H)^{-1}$ and $Q = UY + \mu HX^T W$. Substituting F in (9) with (10), the problem becomes:

$$\begin{aligned} & \arg\min_W Tr\left(Q^T P^T LPQ + Q^T P^T UPQ - Q^T P^T UY\right. \\ & \left. -Y^T UPQ + \mu(W^T X - Q^T P^T)H^T H(X^T W - PQ)\right) \\ & +\gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_W Tr\left(Q^T P^T (L+U)PQ - Q^T P^T UY - Y^T UPQ\right. \\ & +\mu W^T XHX^T W - \mu W^T XHPQ - \mu Q^T P^T HX^T W \\ & \left. +\mu Q^T P^T HPQ\right) + \gamma\|W\|_{2,1} \end{aligned} \tag{11}$$

Since $Tr\left(Q^T P^T UY\right) = Tr\left(Y^T UPQ\right)$ and $Tr\left(\mu W^T XHPQ\right)$ $= Tr\left(\mu Q^T P^T HX^T W\right)$, (11) becomes:

$$\begin{aligned} & \arg\min_{F,W} Tr\left(Q^T P^T (L+U+\mu H)PQ - 2Q^T P^T UY\right. \\ & \left. +\mu W^T XHX^T W - 2\mu Q^T P^T HX^T W\right) + \gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_W Tr\left(Q^T P^T P^{-1} PQ - 2Q^T P^T (UY + \mu HX^T W)\right. \\ & \left. +\mu W^T XHX^T W\right) + \gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_W Tr\left(Q^T P^T Q - 2Q^T P^T Q + \mu W^T XHX^T W\right) \\ & +\gamma\|W\|_{2,1} \end{aligned} \tag{12}$$

As $Q = UY + \mu HX^T W$, we thus have:

$$\begin{aligned} & \arg\min_W Tr\left(\mu W^T XHX^T W - (UY + \mu HX^T W)^T P^T\right. \\ & \left. (UY + \mu HX^T W)\right) + \gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_W Tr\left(\mu W^T XHX^T W - \mu Y^T UPHX^T W\right. \\ & \left. -\mu W^T XHP^T UY - \mu^2 W^T XHP^T HX^T W\right) + \gamma\|W\|_{2,1} \\ \Rightarrow\ & \arg\min_W Tr\left(W^T (XH(\mu I - \mu^2 P)HX^T)W\right. \\ & \left. -2\mu Y^T UPHX^T W\right) + \gamma\|W\|_{2,1}. \end{aligned} \tag{13}$$

Denoting $XH(\mu I - \mu^2 P)HX^T$ by A and $\mu XHPUY$ by B respectively, we obtain the following quadratic problem to solve:

$$\arg\min_W Tr\left(W^T AW\right) - 2Tr\left(B^T W\right) + \gamma\|W\|_{2,1}. \tag{14}$$

3.3 Algorithm

To solve the problem in (14), we use the Lagrangian function and rewrite it as:

$$Tr\left(W^T AW\right) - 2Tr\left(B^T W\right) + \gamma\|W\|_{2,1}. \tag{15}$$

Denote $W = [w^1, ..., w^d]^T$ with w^i as its i-th row. By setting the

Algorithm 1: The SFSS algorithm.

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: Compute the graph Laplacian matrix $L \in \mathbb{R}^{n\times n}$;
2: Compute the decision rule 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:
$D_t = \begin{bmatrix} 2\left\|w_t^1\right\|_2 & & \\ & \cdots & \\ & & 2\left\|w_t^d\right\|_2 \end{bmatrix}$;
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.

derivative of (15) *w.r.t* W to zero, we have:

$$\begin{aligned} &2AW - 2B + 2\gamma D^{-1}W = 0 \\ \Rightarrow\ & W = (A+\gamma D^{-1})^{-1}B = (DA+\gamma I)^{-1}DB, \end{aligned} \tag{16}$$

where D is a diagonal matrix with its diagonal element $D_{ii} = 2\left\|w^i\right\|_2$. Therefore, we propose the detailed iterative approach in Algorithm 1 to solve the objective problem in (14). Next, we show that Algorithm 1 converges, which mainly follows our previous work in [17][26].

THEOREM 1. *The objective shown in (14) monotonically decreases until converging to the global optimum using the iterative approach in Algorithm 1.*

PROOF. The step to compute W_{t+1} in Step 8 of Algorithm 1 performs the following operation:

$$\begin{aligned} &\arg\min_W Tr\left(W^TAW\right) - 2Tr\left(B^TW\right) + \gamma Tr\left(W^T D_t^{-1} W\right) \\ \Rightarrow\ &\arg\min_W Tr\left(W^TAW\right) - 2Tr\left(B^TW\right) + \gamma\sum_i \frac{\left\|w^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \end{aligned} \tag{17}$$

It can be inferred from (17) that W_{t+1} is the minimum of $Tr\left(W^TAW\right) - 2Tr\left(B^TW\right) + \gamma\sum_i \frac{\left\|w^i\right\|_2^2}{\left\|w_t^i\right\|_2}$. Therefore we have:

$$\begin{aligned} &Tr\left(W_{t+1}^TAW_{t+1}\right) - 2Tr\left(B^TW_{t+1}\right) + \gamma\sum_i \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \\ &\le Tr\left(W_t^TAW_t\right) - 2Tr\left(B^TW_t\right) + \gamma\sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \end{aligned} \tag{18}$$

If

$$\sum_i \left\|w_{t+1}^i\right\|_2 - \sum_i \frac{\left\|w_{t+1}^i\right\|_2^2}{2\left\|w_t^i\right\|_2} \le \sum_i \left\|w_t^i\right\|_2 - \sum_i \frac{\left\|w_t^i\right\|_2^2}{2\left\|w_t^i\right\|_2}, \tag{19}$$

by incorporating it to (18) we can get:

$$\begin{aligned} &Tr\left(W_{t+1}^TAW_{t+1}\right) - 2Tr\left(B^TW_{t+1}\right) + \gamma\sum_i \left\|w_{t+1}^i\right\|_2 \\ &\le Tr\left(W_t^TAW_t\right) - 2Tr\left(B^TW_t\right) + \gamma\sum_i \left\|w_t^i\right\|_2, \end{aligned} \tag{20}$$

which proves that the objective function value in (14) monotonically decreases until convergence. Note that it is easy to prove (19) as follows. It is known that $2ab \le a^2 + b^2$. If $b \ne 0$, we have [17]:

$$\begin{aligned} &a \le \frac{a^2}{2b} + \frac{b^2}{2b} \\ \Rightarrow\ &a - \frac{a^2}{2b} \le b - \frac{b^2}{2b} \end{aligned} \tag{21}$$

By substituting a and b with $\sum_i \left\|w_{t+1}^i\right\|_2$ and $\sum_i \left\|w_t^i\right\|_2$ respectively in (21), we obtain (19). Consequently (20) also holds. It can be proved that the objective of our framework is convex, we therefore can conclude that the proposed approach converges to the global optimum. □

4. EXPERIMENTS

In this section, we conduct extensive experiments to study the performance of our framework for automatic image annotation. We also compare the annotation results of our method with four other algorithms.

4.1 Image Datasets

We choose three image datasets, *i.e.*, Corel-5K [12][11], MSRA-MM [14] and NUS-WIDE [6] in our experiments. The following is a brief description of the three datasets.

Corel-5K: This dataset is comprised of real-world images from COREL image CDs. In our experiment, we use the standard dataset used in [12][11]. Corel-5K consists of 5,000 images from 50 different categories, *i.e.*, there are 50 concepts in total.

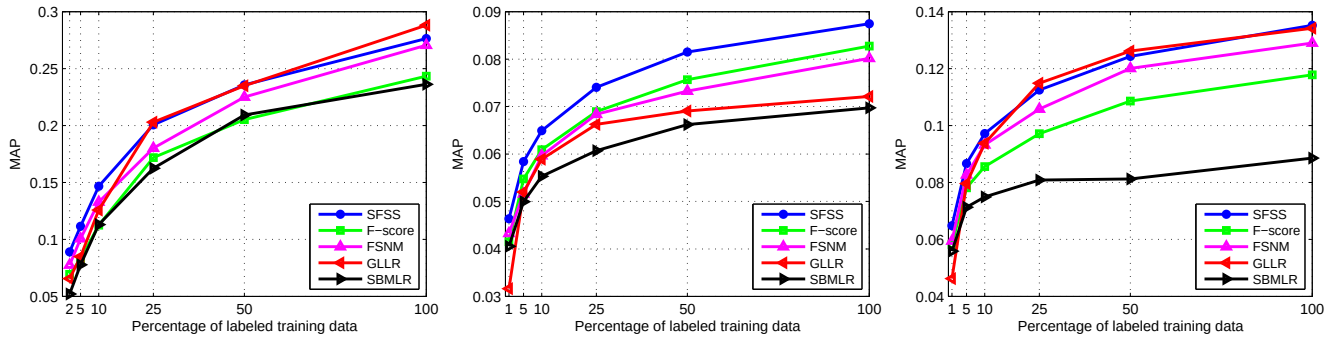
MSRA-MM: The dataset used in our experiments is a subset of the original MSRA-MM 2.0 dataset, which includes 50,000 images related to 100 concepts. However, 7,734 images within it are not associated with any labels. We have removed these images and obtained a subset of 42,266 labeled images. Three feature types used in [25], namely Color Correlogram, Edge Direction Histogram and Wavelet Texture are combined in our experiments to represent the images considering the computational efficiency.

NUS-WIDE: It consists of 269,000 real-world images collected from Flickr. Since 59,653 images of this dataset are not labeled, we have removed them and used the remaining 209,347 labeled images associated with 81 concepts in our experiments. The images are also represented by the combination of Color Correlogram, Edge Direction Histogram and Wavelet Texture.

4.2 Compared Methods

In our experiments, we compare our SFSS framework with one classical and three state of the art methods for automatic image annotation. Detailed information of them is given as follows.

- Fisher Score (F-score) [8]: It depends on fully labeled training data to select features with the best discriminating ability.
- Group Lasso with Logistic Regression (GLLR) [25]: It selects both sparse and discriminative groups of homogeneous features by group lasso extended with logistic regression.
- Feature Selection via Joint $l_{2,1}$-Norms Minimization (FSNM) [17]: It employs joint $l_{2,1}$-norm minimization on both loss



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

Figure 2: Performance variance *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, in the MSRA-MM dataset significantly better performance.

Table 1: A brief comparison between the different methods.

Method	Supervision	Feature Selection with Sparsity	Classifier Generation
SFSS	Semi	Yes	Yes
F-score [8]	Full	No	No
GLLR [25]	Full	Yes	Yes
FSNM [17]	Full	Yes	No
SBMLR [5]	Full	Yes	No

Table 2: Performance comparison (MAP±Standard Deviation) when 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE) training images are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.090±0.008**	**0.047±0.002**	**0.065±0.002**
F-score [8]	0.069±0.006	0.041±0.002	0.058±0.003
GLLR [25]	0.066±0.008	0.032±0.008	0.046±0.007
FSNM [17]	0.078±0.007	0.043±0.002	0.059±0.002
SBMLR [5]	0.052±0.004	0.040±0.002	0.056±0.003

Table 3: Performance comparison (MAP±Standard Deviation) when 5% training images are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.112±0.009**	**0.059±0.002**	**0.087±0.003**
F-score [8]	0.083±0.007	0.055±0.002	0.078±0.002
GLLR [25]	0.085±0.010	0.052±0.001	0.079±0.001
FSNM [17]	0.101±0.007	0.051±0.002	0.082±0.002
SBMLR [5]	0.078±0.005	0.050±0.002	0.071±0.003

Table 4: Performance comparison (MAP±Standard Deviation) when 10% training images are labeled.

	Corel-5K	MSRA-MM	NUS-WIDE
SFSS	**0.147±0.009**	**0.065±0.001**	**0.097±0.002**
F-score[8]	0.113±0.003	0.061±0.002	0.086±0.003
GLLR [25]	0.126±0.015	0.059±0.001	0.094±0.002
FSNM [17]	0.133±0.009	0.060±0.001	0.093±0.003
SBMLR [5]	0.113±0.013	0.055±0.002	0.075±0.007

function and regularization to realize feature selection across all data points.

- Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation (SBMLR) [5]: It exploits sparsity by using a Laplace prior and is used for multi-class pattern recognition. It can also be utilized for feature selection.

Following [17], for FSNM we first select features and then conduct classification with the regularized least square regression. Similarly, F-score and SBMLR are used to perform feature selection and we use regularized least square regression subsequently to learn classifiers from the data represented by the selected features.

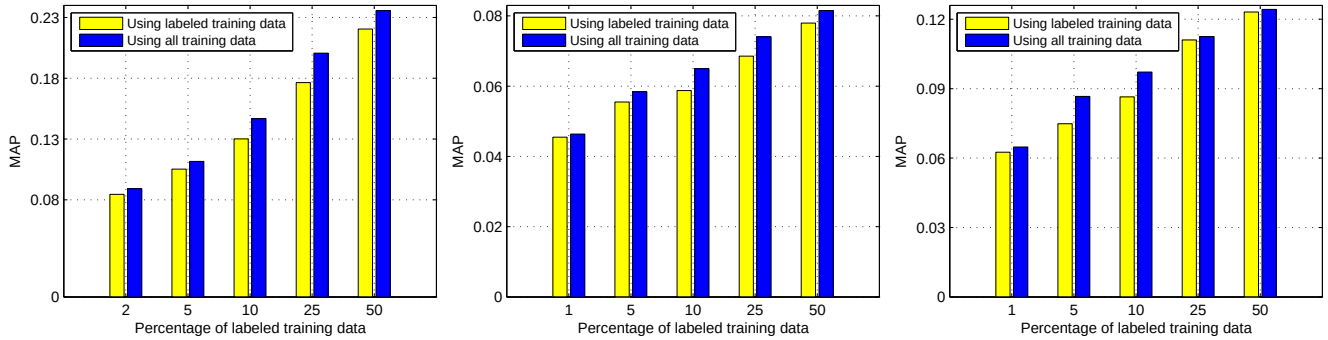
A brief comparison between the properties of our method and the other ones is given in Table 1.

4.3 Experiments Setup

To begin with, we randomly generate a training set for each of the three datasets. The training set comprises n samples, among which m samples are labeled. We set n as 2500 for the Corel-5K dataset and 10000 for both the MSRA-MM and NUS-WIDE datasets. The rest of the images of each dataset work as the corresponding testing set which is used to evaluate the annotation performance. For MSRA-MM and NUS-WIDE, we set m as 1%, 5%, 10%, 25%, 50% and 100% of the total training images respectively. However, to cover all the semantic concepts in Corel-5K, at least 50 labeled training data are required. Thus, slightly different from the settings of the other two datasets, m is set as 2%, 5%, 10%, 25%, 50% and 100% of the total training images for Corel-5K. The results with different m are reported accordingly. The generation of training and testing sets is conducted 5 times for each dataset and the average results are reported.

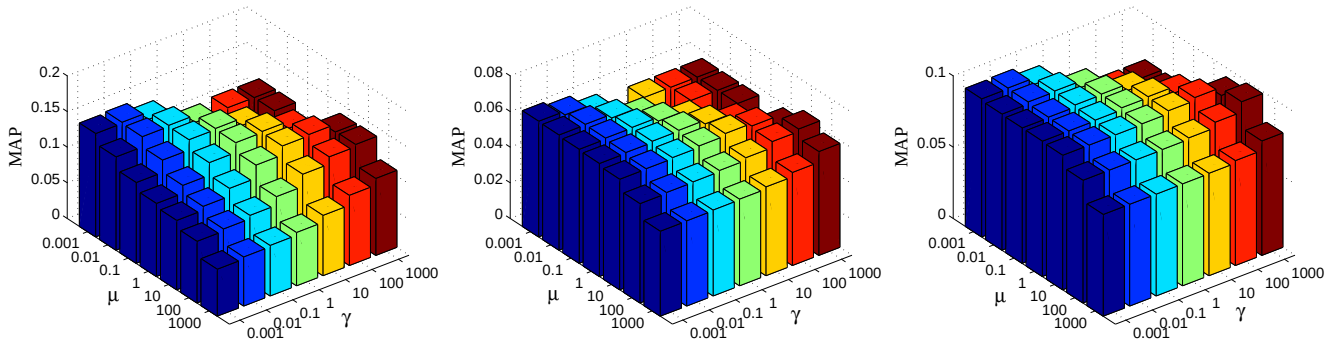
There are two types of parameters to be tuned in our experiments. The first one is the parameter k that specifies the k nearest neighbors used to compute the graph Laplacian matrix L and it is set to 15 for all three datasets. The second one represents the regularization parameters, which are denoted as μ and γ in (6) in our framework. GLLR, FSNM and regularized least square regression also have such kind of parameters. We tune all these parameters from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ and report the best results of corresponding methods.

To evaluate the annotation performance, we compare the labels obtained through different classification approaches with the ground truth labels for the testing sets. Among various evaluation metrics, Mean Average Precision (MAP) is stable and has good discriminat-



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

Figure 3: The influence of unlabeled data on annotation. The blue bar corresponds to our semi-supervised algorithm, namely, SFSS as used in Figure 2. The yellow bar corresponds to the case when only labeled data are used (no unlabeled data). The largest relative improvement of 11%-13% is obtained when 10% training data are labeled. The large difference clearly shows that using unlabeled data improves annotation results.



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

Figure 4: Performance variance *w.r.t* μ and γ. The figure displays different annotation results when using different pairs of μ and γ. It can be seen that when μ and γ are close in value, the annotation performance is generally better than the case when they differ too much from each other.

ing capability. It is now widely used to evaluate accuracy so we use MAP in our experiments as well.

4.4 Image Annotation Results

Figure 2 shows the annotation results when different percentages of data are labeled. Table 2 to Table 4 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 number of labeled training samples increases, performance obviously increases. 2) Our method is the only one which has consistently high scores on all three datasets. All other methods have varying degrees of success on each dataset. 3) When 25% or more of the training data are labeled, our method is competitive with the best algorithms compared or better. However, our advantage over other supervised algorithms decreases in general. On the Corel-5K dataset GLLR [25] slightly outperforms our method; on the NUS-WIDE dataset our method is competitive with GLLR [25]; on the MSRA-MM dataset 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 datasets. This is especially visible on the Corel-5K and MSRA-MM datasets.

The good performance of our SFSS framework can first be attributed to the appealing property that it can select features jointly across the whole feature space. The incorporation of the sparse model facilitates the feature selection by finding the discriminative features, which can assist in learning a robust classifier for image annotation in return. Exploiting the correlation between the predicted labels, the ground truth labels and the data structure also contributes to the good performance. Moreover, besides labeled data, our method can leverage the unlabeled data in the training stage so that it can learn the manifold structure.

4.5 Influence of the Unlabeled Data

As our method uses both labeled and unlabeled data for feature selection, we conduct an experiment to discover the influence of the unlabeled data on the annotation results.

In this experiment, we leave out the unlabeled data in the training sets and only use labeled training data as the input of our framework. 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 done on all the three datasets with 2% (Corel-5K) or 1% (MSRA-MM&NUS-WIDE), 5%, 10%, 25% and 50% training data labeled. The results are displayed in Figure 3.

We observe that on all datasets using unlabeled data in addi-

tion to 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 dataset, 11% on the MSRA-MM dataset and 12% on the NUS-WIDE dataset. We conclude that using unlabeled data clearly improves annotation results.

4.6 Parameter Sensitivity Study

To understand how the two parameters μ and γ in our framework can affect the annotation performance, we perform an experiment on the parameter sensitivity in which 10% training data are labeled. Figure 4 demonstrates the MAP variance *w.r.t* μ and γ. It is interesting to notice that the performance is generally better when μ and γ are comparable in value.

The phenomenon is supposed to be related to the trait of the dataset. However, to discover the root cause is not simple and it is out of the scope of our current focus.

5. CONCLUSION

In this paper we have proposed a new classification framework for automatic image annotation. Our work integrates several state of the art innovations from semi-supervised learning and feature selection, leading to a framework with the following desirable properties. First, our method jointly selects the discriminative features across the entire feature space. In addition, our method can learn the manifold structure from both labeled and unlabeled data. We performed experiments on three real-world image datasets. Results showed that if most of the training data are labeled, our method consistently yields competitive or better accuracy than other methods. When the majority of the data are unlabeled, our method clearly outperforms other methods on all three datasets. Using unlabeled data to boost classification performance is particularly important for image data as high quality manual labeling is expensive. We thus conclude that our method is suitable for large-scale image annotation.

6. ACKNOWLEDGMENTS

The work of Z. Ma, J. Uijlings, and N. Sebe was supported by the Glocal FP7 IP and the S-PATTERNS FIRB projects. 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.

7. REFERENCES

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[2] 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.

[3] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the ACM international conference on Knowledge discovery and data mining (ACM SIGKDD)*, 2010.

[4] G. Carneiro, A. B. Chan, P. J. 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, March 2007.

[5] G. Cawley, N. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[6] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 8–10, July 2009.

[7] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1567, December 2004.

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

[9] Y. Gao, J. Fan, X. Xue, and R. Jain. Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2006.

[10] X. He. Incremental semi-supervised subspace learning for image retrieval. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2004.

[11] S. C. H. Hoi, W. Liu, and S.-F. 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.

[12] S. C. H. Hoi, M. R. 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.

[13] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2004.

[14] H. Li, M. Wang, and X.-S. Hua. MSRA-MM 2.0: A large-scale web multimedia dataset. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 164–169, December 2006.

[15] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic manifold learning for image retrieval. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2005.

[16] J. Liu, M. Li, W.-Y. Ma, Q. Liu, and H. Lu. An adaptive graph model for automatic image annotation. In *Proceeding of the 8th ACM international workshop on Multimedia information retrieval*, 2006.

[17] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l21-norms minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[18] F. Nie, D. Xu, T. W. 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.

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[20] 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.

[21] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5):733–746, 2009.

[22] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai. Optimizing multi-graph learning: towards a unified video annotation scheme. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2008.

[23] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1483–1490, June 2006.

[24] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2010.

[25] F. Wu, Y. Yuan, and Y. Zhuang. Heterogeneous feature selection by group lasso with logistic regression. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2010.

[26] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-norm regularized discriminative feature selection for unsupervised learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2011.

[27] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *Proceeding of the ACM International Conference on Multimedia (ACM MM)*, 2009.

[28] 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.

[29] Z. Zha, X. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[30] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

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

[32] Z. Zhao, L. Wang, and H. Liu. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010.

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

[34] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.