This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

Robust Web Image Annotation via Exploring Multi-facet and Structural Knowledge

Mengqiu Hu, Yang Yang, Fumin Shen, Luming Zhang, Heng Tao Shen and Xuelong Li *Fellow, IEEE,*

***Abstract*— Driven by the rapid development of Internet and digital technologies, we have witnessed the explosive growth of Web images in recent years. Seeing that labels can reflect the semantic contents of the images, automatic image annotation, which can further facilitate the procedure of image semantic indexing, retrieval and other image management tasks, has become one of the most crucial research directions in multimedia. Most of the existing annotation methods heavily rely on well-labeled training data (expensive to collect) and/or single view of visual features (insufficient representative power). In this paper, inspired by the promising advance of feature engineering (e.g., CNN feature and SIFT feature) and inexhaustible image data (associated with noisy and incomplete labels) on the Web, we propose an effective and robust scheme, termed *Robust Multi-view Semi-supervised Learning* (RMSL), for facilitating image annotation task. Specifically, we exploit both labeled images and unlabeled images to uncover the intrinsic data structural information. Meanwhile, to comprehensively describe an individual datum, we take advantage of the correlated and complemental information derived from multiple facets of image data (i.e. multiple views or features). We devise a robust pair-wise constraint on outcomes of different views to achieve annotation consistency. Furthermore, we integrate a robust classifier learning component via $\ell_{2,p}$ loss, which can provide effective noise identification power during the learning process. Finally, we devise an efficient iterative algorithm to solve the optimization problem in RMSL. We conduct comprehensive experiments on three different datasets, and the results illustrate that our proposed approach is promising for automatic image annotation.**

***Index Terms*—Image annotation, multi-view learning, semi-supervised learning, $\ell_{2,p}$-norm.**

I. Introduction

NOWADAYS, because of the tremendous development of smart phones and wireless network communication technologies, it is very convenient to take a photo anytime, anywhere and share it on social network. As a result, recent years have witnessed the explosive growth of web images, which raise urgent demands in various multimedia applications, such as image semantic index, search, retrieve and other image management tasks. Despite much progress has been made in multimedia content analysis [1], [2], major commercial search engines are still powered by the textual index technologies. So, the improvement speed of performance still seriously lags behind the explosive increase of data and more promotion can be achieved by exploiting the big data of images.

In recent years, many researches have concentrated on automatic image annotation [3], [4], which utilizes a set of semantic concepts as the high level or abstract semantic descriptors and subsequently can be used to facilitate image semantic search, index and other various image related multimedia applications. Image annotation is the process of assigning labels or concepts to images, which can reflect the semantic information contained in the image visual features. Because it is usually a difficult, time-consuming, labor-intensive and costly process to manually label a large set of images, many researchers has poured considerable efforts into this area in recent years, emerging remarkable achievements [5], [6].

Image annotation essentially is equivalent to the problem of classification, where an image annotated a label can be treated as classified into the label class. Recently, in the field of multimedia and computer vision, a variety of machine learning and data mining algorithms [7]–[12] for automatic image annotation have been proposed in the literature by many researchers. In [11], a discriminatively nearest neighbor model named TagProp was proposed, where a model based on weighted nearest-neighbor was used to predict the labels of the new test images by exploiting labeled training images. For the problem of the well-known semantic gap, these works related to image annotation have shown promising performances powered by machine learning algorithms. In [13], image annotation approaches were roughly divided into two groups, i.e. tagging or retrieval-based paradigm and labeling- or learning-based algorithms. Typically, retrieval approaches encompass searching phase, where similar images are searched from web data sets, and mining-for-tags phase, where the labels for the test images correspondingly are mined from the textual information associated with the retrieved images.

In the era of big data, we can easily collect plenty of image data. Nonetheless, how to leverage these precious resources to train an annotation model for achieving better prediction performance has been attracting extensive attention. For the task of image tagging, users of many photo-sharing web sites can choose to take the opportunity to assign some labels when uploading photos or refuse. When developing automatic image annotation models, it would be beneficial to deliberate the intrinsic properties [14] of web image collection, including

- inter-class imbalance of labeled instances, i.e. there are only a very narrow number of images that are labeled

Mengqiu Hu, Yang Yang, Fumin Shen and Heng Tao Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (email: hmq_uestc@163.com; dlyyang@gmail.com; fumin.shen@gmail.com; shenhengtao@hotmail.com)

Luming Zhang is with the Department of CSIE, Hefei University of Technology, Hefei, Anhui, China. Email: zglumg@gmail.com

Xuelong Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China. Email: xuelong_li@opt.ac.cn

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

in many classes, in conjunction with a large number of additional images unlabeled;
- multiple facets, that is, multiple distinct features or views are usually needed to completely describe the contents of an image, such as hand-crafted features LLC [15] and FK [16] encoding of local features for bags of visual words models, and learned abstract features FC6 and FC7 derived directly from the output of the FC6 and FC7 layers of Alexnet [17];
- imbalance between classes, where the number of samples of some categories could be too large, while may be very few in the other classes; and
- noise issue, some labels of the images could be incomplete and noisy. Recently, many techniques, such as semi-supervised learning, multi-view learning and robust learning, have been proposed to address the various kind of problems above.

To efficiently retrieve, browse and manage the large number of web images, many approaches have been proposed. Most existing annotation methods are supervised learning [8], [18], designed for single view features that are often noisy and redundant. In [8] a probabilistic formulation was proposed for the tasks of semantic image annotation and retrieval, together with theoretical arguments and extensive experiments for illustration. Recently, many algorithms based on semi-supervised learning has also been proposed for image annotation. Considering the complexity of contents of images, the existing semi-supervised learning methods based on machine learning, which require a large number of training samples, are developed to achieve reasonable performance. A new semi-supervised annotation approach by optimal graph learning [12] was proposed, which can enclose the relevances of different data points more accurately. However, the real world images are always represented as multiple features with most unlabeled. In contrast to single-view algorithms, generally, more performance promotion can be achieved by utilizing multiple features properly. A multiview method based on Hessian regularization [19] was proposed for image annotation task.

Inspired by the above observations and analysis, in this paper, we propose a new model, termed as *robust multi-view semi-supervised learning* (RMSL), for image annotation task, which expands our previous work [20]. Our model illustrated as Figure 1 jointly explores intrinsic data structural knowledge as well as multi-faceted information embedded in the various features of data. Specifically, we employ semi-supervised learning based on graph to model local structure of image data, which can uncover the intrinsic data structural information by exploiting both labeled images and unlabeled images. Meanwhile, we devise a multi-view constraint to enforce the outcomes of different views to be as consistent as possible. By doing so, we are able to effectively explore the correlated and complemental information from different views to comprehensively describe data. Moreover, in order to further handle the problem of unreliable labels, we develop a robust learning component which leverages $\ell_{2,p}$ loss function to effectively suppress the negative influence of noisy samples in the training set and add flexibility and adaptability for various specific data. Finally, we devise an efficient iterative algorithm to solve the optimization problem in RMSL. We conduct extensive experiments on three different datasets, i.e. NUS-WIDE [21], MIRFLICKR-25000 [22] and IAPR TC-12 [23], [24], and the results illustrate that our proposed approach is superior for large scale web image automatic annotation task.

The remainder of the paper is organized as follows. Related work is briefly reviewed in Section 2. In section 3, we detail the proposed algorithm and its solution, followed by experimental results and analysis. Lastly, conclusions are drawn in Section 4.

II. Related Work

In this section, we briefly review the related research on semi-supervised learning, multiple feature learning and automatic image annotation.

A. Semi-Supervised Learning

In recent years, the power of narrowing the semantic gap has been demonstrated by many methods based on supervised learning [25], [26]. For example, Support Vector Machine (SVM) and its various variants, as a typical supervised method, have been widely investigated for image annotation in the literature [27], [28], [29], [30]. However, it is usually costly and not easy to manually label a large set of images, which is also a time-consuming, labor-intensive task, and there are many rich images with various contents that can be used to further advance the performance. Recently, many researchers poured much attention into semi-supervised learning and many successes have been achieved in the application of image annotation and classification [31], [32], [33], [34].

Considering the excessive cost of annotation of a large number of data manually, semi-supervised learning is committed to reducing this phenomenon. It is rather intuitive that there should be some common labels between two image samples having similar features, which is termed as manifold assumption [35]. The most common class of semi-supervised methods based on the manifold assumption leverages manifold regularization to seek the intrinsic geometry of the data distribution, which is achieved by penalizing the regularization term along the potential manifold. The most common way to characterize the underlying manifold geometry of a dataset is to choose the graph Laplacian [36], [37], whereas there are some other methods, which have been paid much attention and have brought about many achievements. It is flourishing in the literature that many semi-supervised methods based on graph Lapacian have been proposed, such as Laplacian Regularized Least Square Regression (LapRLS) [35], Laplacian Regularized Support Vecto Machine (LapSVM) [35] and Flexible Manifold Embedding [38].

During recent years, graph-based semi-supervised learning, as one of the important branches of semi-supervised learning, has been developed by many researchers [39] and many advancements in this direction have been achieved in the literature [19], [35], [38], [40], [41]. In [40], a semi-supervised

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

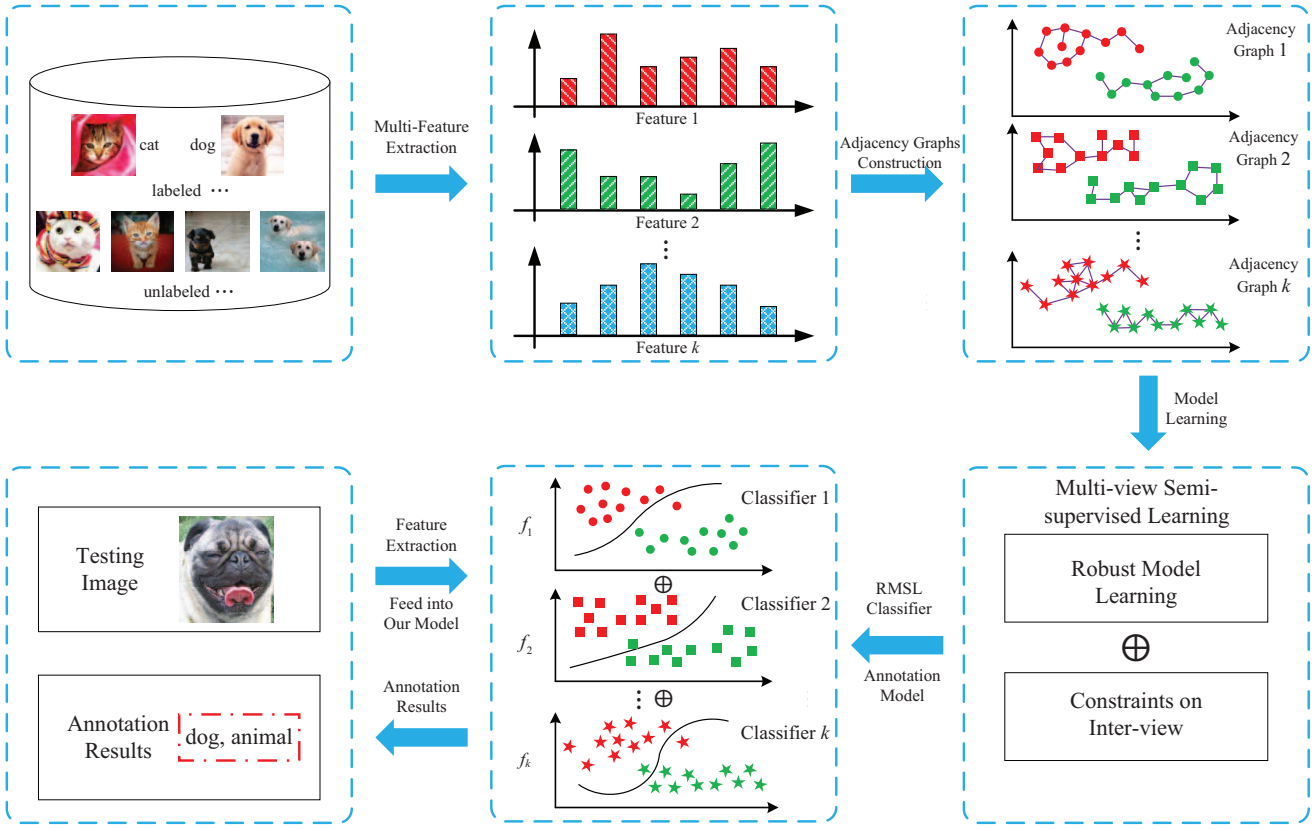


Fig. 1: *The illustration of our RMSL framework for automatic image annotation.*

method based on graph was proposed, which consider to learn from both local consistency of similarity and global manifold consistency. In the method, the label information of each point is iteratively spread to its neighbors before satisfying a global stable state, so the prediction of label information for unlabeled samples was finished during the iterative procedure. However, a new testing image which is not in the training set can not get the label information by LGC, because it belongs to the class of transductive method. In contrast to transductive algorithm, some other inductive method of semi-supervised learning which can predict labels for new testing set, such as Laplacian Regularized Least Square Regression [35] and Flexible Manifold Embedding [38], have also been proposed. Typically, the methods based on manifold framework encompass two terms, i.e. a loss function and a regularizer. So, by substituting different kind of loss functions and regularizers, which affect classification performance significantly, we can get a variety of models. In [19], a semi-supervised method based on Hessian regularization was proposed for image annotation, instead of Laplacian regularization.

B. Multiple Feature Learning

A variety of methods which can perform well in single view has been proposed in the literature, whereas in multiple features, they can not exhibit the superiority of multi-view. For an image, many different features can be extracted using different methods, such as hand-crafted features LLC and FK, and learned abstract features FC6 and FC7. Each view contains different discriminative information, which characterises specific contents of the image from one aspect, and different views usually are complementary to each other. Even if there are not natural multiple features, multiple view features can be achieved by splitting one feature.

Recently, multiple feature learning has been investigated a lot in the literature [19], [42], [43]. In general, there are three feature fusion strategies, including early fusion, late fusion and multi-stage fusion.

In early fusion strategy, faced with multiple view features, it is simple to directly concatenate each feature, thus resulting in a long feature vector. Although a good performance may be achieved by this simple direct fusion scheme, such as in [44], there are also accompanied with some other problems, such as the burden of more computational resources and over-fitting problem especially in case of small training set. However, more computational resources required to process the long feature vectors formed by concatenating directly can not guarantee improved performance, that is, worse performance may be obtained especially when faced heterogenous features

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

[45]. It may be caused by the lost of individual structural information of each feature when concatenating.

Different with early fusion methods, late fusion strategy firstly learn multiple models by efficiently leveraging each kind of the features, then the multiple models are fused into a unified objective framework by certain criterion. As shown in the name, i.e. late fusion, the fusion was turn up after separate learning of each feature in stead of the fusion of multiple features firstly. In [46], a late fusion algorithm based on SVM, which combines KCCA and SVM two-stage learning into a single optimisation termed SVM-2K, has been proposed to process the situation of two types of features. Another typical statistical approach, named Canonical Correlation Analysis (CCA), try to seek the projection directions which can maximizes the correlations between two sets of multidimensional variables. [47] established an equivalence least-squares formulation for CCA, which has been successfully applied for multi-label classification. However, late fusion also has the same drawback as early fusion, that is the high expense of learning. Moreover, correlations among multiple features have not been taken into consideration by most late fusion approaches.

In addition to early and late fusion strategies, the multi-stage fusion scheme, i.e. combining early and fusion scheme into a unified framework, has also been explored recently. For instance, in [48], a two-step feature fusion method was proposed, where multiple kernel learning is firstly used to integrate various visual and audio multi-modal features as the early fusion scheme, followed by late score fusion strategies. It was justified that additional performance improvements can be gained by multi-stage fusion. It is generally beneficial to combine multiple features for image content analysis. However, improvement can't always be guaranteed and even worse performance may be obtained by feature fusion if the different features are paradoxical or a feature is too strong together with a very weak feature. There are still many things needed to be investigated to evaluate the appropriateness of combining multiple features [49].

C. Automatic Image Annotation

Nowadays, more and more images are available and most of content-based image retrieval methods having been used still far can't get satisfactory experience. Therefore, automatic image annotation, which can assign correlated labels to images thus transform content-based image retrieval to text-based image retrieval, is particularly important. Assuming images represented with features and corresponding semantic labels are collected, various machine learning algorithms can be chosen to learning a family of models to fit the matching relationships between image features and semantic labels. Once a model was trained, a new image can be annotated using the algorithm. In [50] automatic image annotation methods were classified into three types, that is, single labeling annotation using conventional classification methods, multi-labeling annotation and the web based image annotation using metadata.

The first annotation approach, i.e. single labeling annotation using binary classification, typically encompasses SVM and SVM-based methods, decision tree (DT) and decision-tree-like algorithms, etc. SVM has been shown high effectiveness in many applications [27]–[30], [46], especially when the size of training data set is small and the dimensionality of the feature vector is high. Originally, SVM is a classifier designed to address this situation where there are only two categories, working by finding the optimal hyperplane from the training samples to separate them maximally. However, image annotation is usually a multiple classes problem where images are always correlated with several labels. Faced with multiple classes, it is always common and efficient to choose the OVA scheme, that is training a separate SVM for each label class against all the rest of the class. In [27] image segmentation and classification were conducted simultaneously using multiple SVMs, where the images were segmented into regions and annotated for each segmented regions.

Compared to binary classification approaches, multiple labeling methods annotate an image with multiple labels simultaneously. Generally, the content of an image sample is associated with multiple objects [51], so it's commonly an image correlated with multiple semantic concepts. One of typical multi-labeling methods is probabilistic based algorithms such as the Bayesian methods [7], [52]. In [52], a relevance model was proposed assuming that the regions in an image can be represented using a small vocabulary of blobs, which were generated from image features using clustering. For a new test image, the probabilistic of generating a word associated with it can be derived from the relevance model trained from the given training samples. In contrast to algorithms based on probabilistic, there are also many non-probabilistic based methods proposed in the literature [13], [41], [49]. A new algorithm framework for image annotation by simultaneously considering label correlation and visual similarity was proposed [13].

The images in the web are usually accompanied with metadata, such as text descriptions, URL, HTML code, GPS data and timestamps, etc. In the literature, many image annotation methods incorporating metadata have been proposed [53], [54]. In [54], a non-parametrical model was proposed using image metadata. In this method, firstly, Jaccard similarities was used to generate related neighborhoods for an image, then the visual information of an image and its neighbors was combined together by a deep neural network. An approach called weakly semi-supervised deep learning for multi-label image annotation (WeSed) [55] was proposed recently. In WeSed, a novel weakly weighted pairwise ranking loss was effectively utilized to handle weakly labeled images, while a triplet similarity loss was employed to harness unlabeled images.

In the next section, we propose a new model for image annotation task, termed as Robust Multi-view Semi-supervised Learning (RMSL). In our method, we apply graph Laplacian based semi-supervised learning to explore underlying data structural knowledge. Different from the single view semi-supervised method SFSS [41] based on feature selection for automatic image annotation, which aims to jointly select the most relevant features from all the data points by using a sparsity-based model, we propose to incorporate $l_{2,p}$-norm into

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

our objective function to add flexibility and adaptability from our data, which can show effective noise identification power during the training process. In the method LRGA proposed by Yang et al [56], for each data point, a local linear regression model, where ridge regression was adopted to weakly handle noisy samples, was used to predict the ranking values of its neighbouring points, while our robust $l_{2,p}$ loss function was constructed using the whole dataset directly for the task of automatic image annotation. In order to maximize the efficacy of multiple views (e.g., LLC and FC7) and to achieve the semantic consistency between different views, we deliberately incorporate the consistent term via robust $\ell_{2,q}$ loss (resistant to noisy samples), which is mostly different from many existed multiple views learning methods [47], [57].

III. The Proposed Approach

In this section, we present the proposed Robust Multi-view Semi-supervise Learning (RMSL) approach. Firstly, we start with a recap of graph-based semi-supervised learning, and then elaborate the formulation of our RMSL, followed by an efficient algorithm for optimizing the model. Lastly, we theoretically analysis the convergence of our algorithm.

A. Notations

Suppose we have a training dataset of n observations from m views. Denote $X_t = [x_1^{(t)}, x_2^{(t)}, \ldots, x_l^{(t)}, x_{l+1}^{(t)}, \ldots, x_n^{(t)}]$ as the t-th features of these samples, where $x_i^{(t)} \in \mathbb{R}^{d_t \times 1} (1 \leq i \leq n)$ is the t-th view feature of the i-th observation, where d_t is the dimensionality of the t-th feature space. Note that the first l samples in the datasets are associated with labels while the rest $n-l$ samples are unlabeled. Given the label matrix of the training dataset corresponding to the t-th view $Y_t = [y_1^{(t)}, y_2^{(t)}, \ldots, y_l^{(t)}, y_{l+1}^{(t)}, \ldots, y_n^{(t)}]^T \in \{-1, 0, 1\}^{n \times c}$, where c is the number of labels, $y_i^{(t)} \in \{-1, 1\}^c$ if $(1 \leq i \leq l)$ (i.e., labeled sample) and $y_i^{(t)}$ is all-zero vector if $(l+1 \leq i \leq n)$ (i.e., unlabeled sample). Let $y_{ij}^{(t)}$ denote the j-th class of the i-th datum corresponding to the t-th view, then $y_{ij}^{(t)} = 1$ if the i-th sample is in the j-th class, and $y_{ij}^{(t)} = -1$ otherwise. If the sample is unlabeled, $y_{ij}^{(t)}$ is set to zero. The objective of this work is to utilize multiple views of both the labeled and unlabeled data to learn robust classifiers for image annotation.

B. Graph-based Semi-Supervised Learning

Given a set of data samples, we can use the visual features to construct a graph model S, whose element S_{ij} reflects the visual similarity between the two image samples x_i and x_j on the graph. In order to reduce the number of parameters, we simply define S as below:

$$S_{ij} = \begin{cases} 1, & x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i); \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathcal{N}_k(\cdot)$ denotes the set of k nearest neighbors of a datum. By defining a diagonal matrix D, whose i-th diagonal element is computed as $D_{ii} = \sum_{j=1}^{n} S_{ij}$, we have the graph Laplacian matrix $L = D - S$.

In order to exploit labeled and unlabeled data simultaneously, we define $F = [f_1, f_2, \cdots, f_n]^T \in \mathbb{R}^{n \times c}$ as a predicted label matrix for all the training data, where $f_i \in \mathbb{R}^c (1 \leq i \leq n)$ is the predicted label of the i-th sample. According to the idea of semi-supervised learning [39], F should be simultaneously consistent with the ground truth labels and the visual graph model S. Therefore, F can be obtained by solving the minimization optimization problem of the following objective function:

$$\min_F Tr(F^T L F) + Tr((F-Y)^T U (F-Y)), \tag{2}$$

where $U \in \mathbb{R}^{n \times n}$ is a diagonal matrix and named as a decision rule matrix, whose diagonal elements U_{ii} is a large number(set as 10^{10} in our experiment) if the i-th data point is labeled and $U_{ii} = 1$ otherwise. This setting of decision rule matrix U can keep the solved labels F in line with the ground truth label matrix Y where the images are labeled.

In order to learn a robust classifier, which should be tolerant to samples with noisy labels, we propose to integrate a robust loss function with adaptive power to different noise levels. To this end, we employ the generalized $\ell_{2,p}$ loss, then the graph-based semi-supervised classification learning framework can be rewritten as follows:

$$\begin{aligned} \min_{F,W,b} \; & Tr(F^T L F) + Tr((F-Y)^T U (F-Y)) \\ & + \mu(\|X^T W + \mathbf{1}_n b^T - F\|_{2,p} + \gamma \|W\|_F^2), \end{aligned} \tag{3}$$

where μ and γ are balance parameters. $W \in \mathbb{R}^{d \times c}$ is the mapping matrix and $b \in \mathbb{R}^c$ is the bias term. $\|W\|_F^2$ is the regularization term, $\mathbf{1}_n$ is an all-one vector. The $\ell_{2,p}$ norm of a matrix M is defined as

$$\|M\|_{2,p} = \sum_{i=1}^{n} \|M_i\|_2^p, \tag{4}$$

where M_i is the i-th row of M.

C. Multi-view Semi-supervised Learning

For the t-th view of the data, we can compute a view-depended Laplacian matrix L_t from the view feature X_t. Then we can calculate a view-depended predicted label matrix F_t from Eq. (3) accordingly. Thus we can introduce the idea of multi-view learning into the aforementioned graph-based semi-supervised classification learning to leverage the correlated and complemental information between different views for better performance. To this end, we propose to jointly minimize all the view-specific objective functions and restricting all view-specific F_t to be as closely as possible.

Therefore, we can express our final objective function of the robust multi-view semi-supervised learning (RMSL) as following:

$$\begin{aligned} \min_{\{F_t, W_t, b_t\}|_{t=1}^m} & \sum_{t=1}^{m} \big(Tr(F_t^T L_t F_t) + Tr((F_t - Y)^T U (F_t - Y)) \\ & + \mu(\|X_t^T W_t + \mathbf{1}_n b_t^T - F_t\|_{2,p} + \gamma \|W_t\|_F^2) \big) + \lambda \sum_{t,s} \|F_t - F_s\|_{2,q}, \end{aligned} \tag{5}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

where λ is a balance parameter. The above formulation benefits from multi-view learning together with graph-based semi-supervised learning. This model effectively utilizes the large amount of unlabeled data and complemental information from different views. The last term $\lambda \sum_{t,s} \|F_t - F_s\|_{2,q}$ is able to enforce the outcomes of all pairs of views to be as consistent as possible, thereby leading to better performance.

D. Solution

In this section, we propose an efficient iterative algorithm to solve our model. Note that it is non-trivial to directly solve the problem in Eq. (5) due to the non-convexity of the $\ell_{2,p}$ loss and the $\ell_{2,q}$ regularizer. To overcome this problem, we first transform the original formulation to the following alternative one:

$$\begin{aligned}\min_{\{F_t,W_t,b_t\}|_{t=1}^m} \sum_{t=1}^{m} \big(&Tr(F_t^T L_t F_t) + Tr((F_t - Y)^T U (F_t - Y)) \\ &+\mu Tr(X_t^T W_t + 1 b_t^T - F_t)^T D_t^{(l)} (X_t^T W_t + 1 b_t^T - F_t) \\ &+\mu\gamma \|W_t\|_F^2\big) + \lambda \sum_{t,s} Tr(F_t - F_s)^T D_{ts}^{(r)} (F_t - F_s),\end{aligned} \tag{6}$$

where $D_t^{(l)}$ a diagonal matrix with its i-th diagonal element computed as

$$(D_t^{(l)})_{ii} = \frac{1}{\frac{2}{p}\|r_t^i\|_2^{2-p}}, \tag{7}$$

where r_t^i is the i-th row of the matrix $X_t^T W_t + \mathbf{1}_n b_t^T - F_t$. Similarly, $D_{ts}^{(r)}$ is diagonal and its diagonal element is calculated as

$$(D_{ts}^{(r)})_{ii} = \frac{1}{\frac{2}{q}\|r_{ts}^i\|_2^{2-q}}, \tag{8}$$

where r_{ts}^i is the i-the row of the matrix $F_t - F_s$.

Note that both $D_t^{(l)}$ and $D_{ts}^{(r)}$ are related to F_t, W_t and/or b_t, which makes the problem in Eq.(6) hard to solve. Thus, we design an iterative method which uses fixed $D_t^{(l)}$ and $D_{ts}^{(r)}$ obtained in the previous iteration to bypass the obstacle. In this way, we can see that F_t, W_t and b_t can be solved from Eq. (6).

By setting the derivative of (6) w.r.t. b_t to be zero, we have

$$b_t^T = \frac{\mathbf{1}^T D_t^{(l)} (F_t - X_t^T W_t)}{\mathbf{1}^T D_t^{(l)} \mathbf{1}}. \tag{9}$$

Substituting b_t in (6) by (9) and setting the derivative of (6) w.r.t. W_t to be zero again, we get

$$W_t = A_t F_t, \tag{10}$$

where

$$H_t = I - \frac{\mathbf{1}_n \mathbf{1}_n^T D_t^l}{\mathbf{1}_n^T D_t^l \mathbf{1}_n}, \tag{11}$$

$$A_t = (X_t H_t^T D_t^l H_t X_t^T + \gamma I)^{-1} X_t H_t^T D_t^l H_t. \tag{12}$$

Substituting b_t, W_t in (6) by (9), (10) respectively, we arrive at

$$\begin{aligned}\min_{F_t|_{t=1}^m} \sum_{t=1}^{m} \Big(&Tr(F_t^T (L_t + \mu(I - A_t^T X_t) H_t^T D_t^{(l)} H_t) F_t) \\ &+Tr(F_t - Y)^T U (F_t - Y)\Big) + \lambda \sum_{t,s} Tr(F_t - F_s)^T D_{ts}^{(r)} (F_t - F_s)\end{aligned} \tag{13}$$

By setting the derivative of the above objective function w.r.t. F_t to be zero, we have

$$F_t = M_t Q_t, \tag{14}$$

where

$$M_t = (L_t + \mu(I - A_t^T X_t) H_t^T D_t^l H_t + U + \lambda \sum_{s=1}^{m} D_{ts}^{(r)})^{-1}, \tag{15}$$

$$Q_t = (UY + \lambda \sum_{s=1}^{m} D_{ts}^{(r)} F_t), \tag{16}$$

and we set $D_{ts}^{(r)} = 0$ when $t = s$, $t = 1, 2, \ldots, m$.

Algorithm 1: The algorithm for optimizing RMSL model.

Input : The tth view feature matrix of training set $X_t \in \mathbf{R}^{d_t \times n}$, and corresponding ground-truth label matrix of the training set $Y \in \mathbf{R}^{n \times c}$;

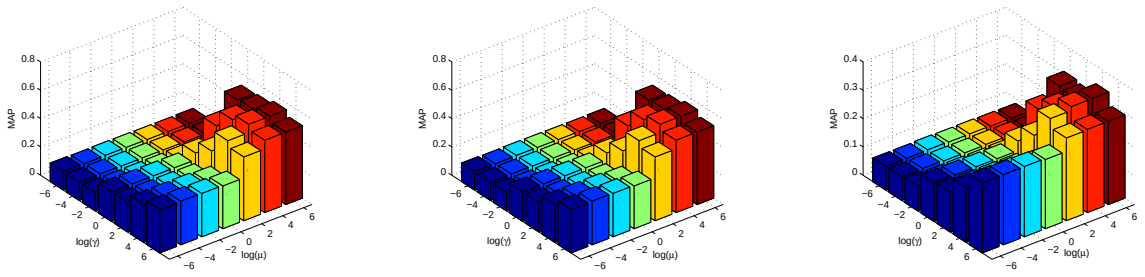
Output: Optimized classification parameters matrix $W_t \in \mathbf{R}^{d_t \times c}$, and bias vector b_t;

```
Randomly initialize F_t, W_t and b_t, t = 1, 2, ..., m;
Compute Laplacian matrix L_t of the t-th view according to Eq. (1);
repeat
    Compute D_t^(l) and D_ts^(r) according to Eq. (7) and Eq. (8), respectively;
    for t = 1, 2, ..., m do
        Compute H_t according to Eq. (11);
        Compute A_t according to Eq. (12);
        M_t according to Eq. (15);
        Q_t according to Eq. (16);
        Update F_t according to Eq. (14);
        Update W_t according to Eq. (10);
        Update b_t^T according to Eq. (9);
    end
until there is no change to F_t, W_t and b_t, t = 1, 2, ..., m;
return F_t, W_t and b_t, t = 1, 2, ..., m;
```

In this way, we can solve the objective function to obtain the optimal solutions of F_t, W_t, b_t by the proposed RMSL algorithm. The detailed approach is summarized in **Algorithm 1**. After getting the values of W_t, b_t, we use the multi-view relations to calculate the predicted values for the new testing set, which is formulated as the following equation:

$$\widehat{F} = \frac{1}{m}\left(\sum_{t=1}^{m} (X_t^T W_t + 1 b_t^T)\right). \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 2: *The effects of the parameters μ and γ in terms of MAP performance on three datasets.*

E. Convergence Study

In this section, we firstly present two lemmas together with detailed proof, which are helpful for proving **Theorem 1**. Then we give and prove the theorem for the convergence of our proposed **Algorithm 1**.

Lemma 1: Let r_t^i be the i-th row of the residual $R_t = X_t^T W_t + \mathbf{1}_n b_t^T - F_t$ in previous iteration and $\tilde{r}_t^i$ be the i-th row of the residual $\tilde{R}_t = X_t^T \tilde{W}_t + \mathbf{1}_n \tilde{b}_t^T - \tilde{F}_t$ in current iteration, then the following inequality holds:

$$\|\tilde{r}_t^i\|^p - \frac{p\|\tilde{r}_t^i\|^2}{2\|r_t^i\|^{2-p}} \leq \|r_t^i\|^p - \frac{p\|r_t^i\|^2}{2\|r_t^i\|^{2-p}}. \tag{18}$$

Proof. Detailed proof is in Appendix.

Lemma 2: Given $R_t = [r_t^1, r_t^2, \ldots, r_t^n]^T$, where r_t^i is the i-th row of R_t, then we have the following conclusion:

$$\sum_{i=1}^{n} \|\tilde{r}_t^i\|^p - \sum_{i=1}^{n} \frac{p\|\tilde{r}_t^i\|^2}{2\|r_t^i\|^{2-p}} \leq \sum_{i=1}^{n} \|r_t^i\|^p - \sum_{i=1}^{n} \frac{p\|r_t^i\|^2}{2\|r_t^i\|^{2-p}}. \tag{19}$$

Proof. We sum up inequalities corresponding to all rows of R_t in Lemma 1 we have the conclusion in Lemma 2.

Similarly, we can get the two lemmas for r_{ts}^i and $\tilde{r}_{ts}^i$, which are the i-th row of the matrix $F_t - F_s$ in previous iteration and the matrix $\tilde{F}_t - \tilde{F}_s$ in current iteration, respectively.

Theorem 1: At the iteration (line 4 to line 9) of Algorithm 1, the value of the objective function in Eq. (5) monotonically decreases.

Proof. Detailed proof is in Appendix.

IV. Experiments

In this section, in order to validate the effectiveness of our proposed approach on image annotation task, we conduct comprehensive experiments on three real web image datasets, i.e., NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12.

A. Datasets and Features

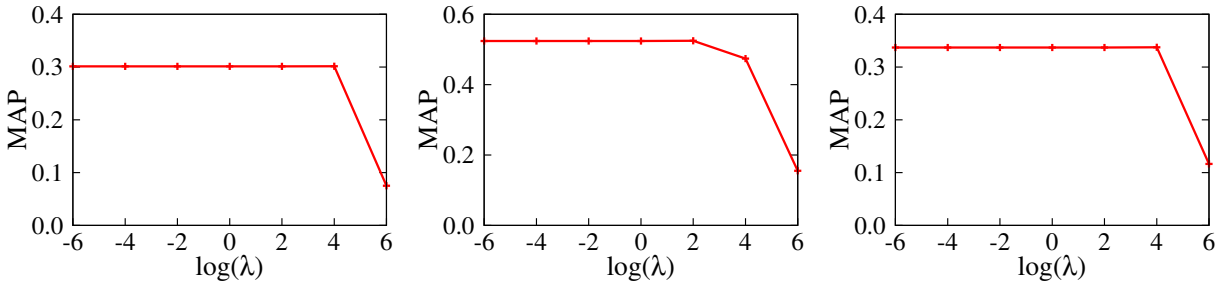
The NUS-WIDE dataset totally includes 269,648 real-world images which are labeled with 81 concepts. The dataset is separated into two parts, i.e., the training set containing 161,789 images and the testing set consisting of 107,859 images. The MIRFLICKR-25000 dataset comprises 25,000 images with 24 concepts. In this paper, we use the potential labels in a very wide sense as our ground-truth, which is included in the newest version files provided by the authors. Following the setting of Standardized Challenge # 1 suggesting in [22], the total image set consisting of 25,000 images is divided into two parts, i.e., the training set and test set including 15,000 and 10,000 images respectively. As suggested in the paper [22], we partition every five images to reduce bias, that is, the first three are assigned as training images, the last two as test images. The IAPR TC-12 dataset consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. While this dataset was created for the CLEF cross-language image retrieval track (ImageCLEF), it has also been used as a benchmark for the task of automatic image annotation in [11], [58]. In [24], the 20,000 images in the collection have been annotated using 255 labels. For our experiments, we collect the 40 most frequent labels for evaluation, which gives us 19,296 images, and we spilt it into two subsets equally, i.e., the training set and test set both including 9,648 images.

In our experiments, we first extracted two types of hand-crafted visual features based on two novel encodings for bag of visual words models using SIFT local descriptor [59], which are locality-constrained linear encoding (LLC) [15] and improved Fisher encoding (FK) [16], using the code [60] and [61]. The final dimensionality of the LLC feature vector equals to k (e.g., the vocabulary size) and we set $k = 4096$. For the improved Fisher encoding, the dimensionality of the FK feature vector equals to $2d * k$, where d is the descriptor (SIFT) dimensionality and k is still the vocabulary size, and in our experiments we reduce the dimension of SIFT descriptor from 128 to 50 by PCA. Now our FK feature vector is 25,600 dimensions, which is further reduced to 4096 by PCA to save computational cost. Consequently, the LLC and FK feature vectors are both 4096-d. In addition, we extracted two new sets of deep learning features, i.e., FC6 and FC7, which are also both 4096-dimension, based on the outputs of the 6th and the 7th fully connected layers by Caffe [62].

B. Compared Methods and Experimental Setup

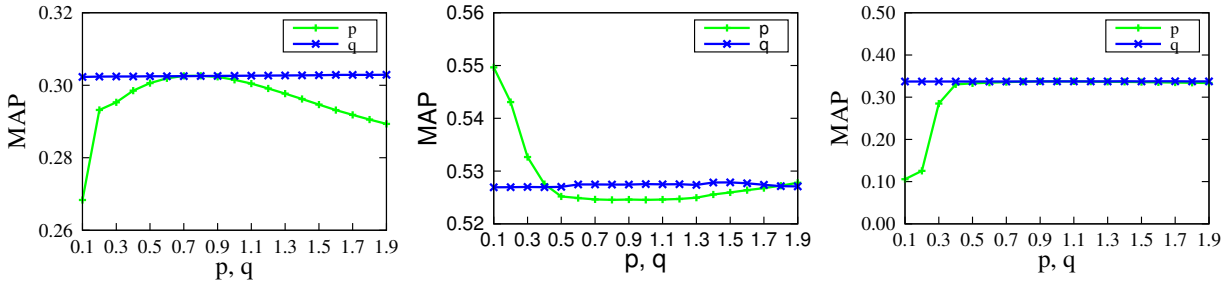
For the comparison of the proposed RMSL with the existing algorithms in the case of multiple features, we present the comparison results with the representative multiple feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 3: *The effects of the parameter λ in terms of MAP performance on three datasets.*



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 4: *The effects of the parameters p and q in terms of MAP performance on three datasets.*

learning algorithm CCA [47] (followed by Least Square regression and SVM, which are denoted as CCA-LS and CCA-SVM, resprestively) and a new method for multiple view semi-supervised dimensionality reduction [57] (followed by Least Square regression, which are denoted as MVSSDR-LS). Also we report the results between our RMSL with two semi-supervised algorithms, i.e., Structural Feature Selection with Sparsity (SFSS) [41] and Flexible Manifold Embedding (FME) [38]. Additionally, we also show the results between our RMSL and a new family of boosting algorithms, denoted TaylorBoost, and in our experiments we choose the Laplace loss which is usually the best of losses pointed in [63].

In our experiment, we follow the convention of semi-supervised learning approaches setting for comparison. To simulate a semi-supervised learning scenario, we randomly divided the training data into two subsets: one set was called labeled set whose labels are known, the rest was called unlabeled set whose labels were hidden. Specifically, the training set encompassing both labeled and unlabeled data is used to train the model, and the testing set is only available for predicting. Denote c as the number of classes in each dataset (i.e., $c = 81$ for NUS-WIDE, $c = 24$ for MIRFLICKR-25000 and $c = 40$ for IAPR TC-12). For all the subsequent experiments, there are a total of 2000 training images, which include l labeled images ($l = 1$, 3, 5, 10, and 15) per category randomly sampled from the training set, together with the remaining selected images unlabeled.

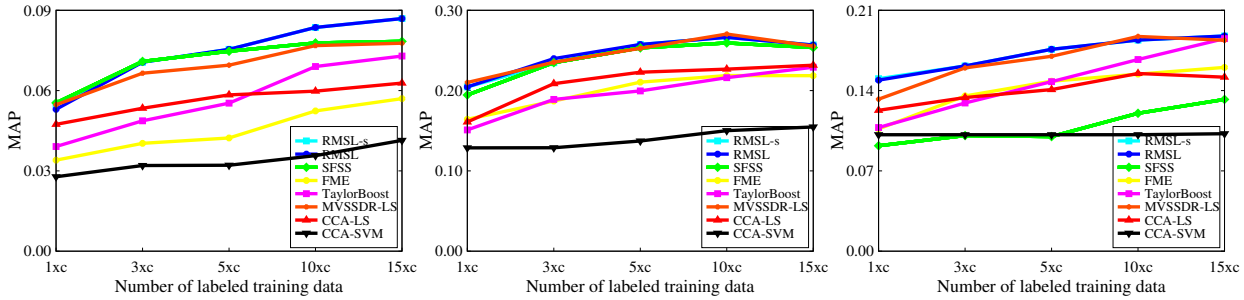
The number k of nearest neighbors for computing Laplacian matrix is set to 15. In our proposed method, there are totally five parameters, i.e., μ, γ, λ, p and q. We conducted a range of experiments to tune the parameters and selected the parameters around the best performance in our experiments. Specifically, we tuned the parameters μ, γ and λ in the range of $\{10^{-6}, 10^{-4}, \ldots, 10^{4}, 10^{6}\}$ and chose p, q from $\{0.1, 0.2, \ldots, 1.8, 1.9\}$. For CCA-LS, CCA-SVM, MVSSDR-LS, FME and SFSS, we also tune their parameters from the same range of $\{10^{-6}, 10^{-4}, \ldots, , 10^{4}, 10^{6}\}$. For CCA-SVM, we used linear kernel for SVM and there was no parameter to tune for TaylorBoost. The parameters corresponding to the best results were used in our experiments. In our experiments, we used the metric of Mean Average Precision (MAP) to evaluate performance in terms of the task of image annotation.

C. Parameters Setting

In this part, we evaluate the effects of different parameters on our methods and we don't tune all the 5 parameters together because of different sensitivities and saving the computing time. Specifically, we first tune μ and γ together while keeping other parameters, i.e. λ, p and q fixed at 1. The results are showed in Figure 2 for NUSWIDE, MIRFLICKR-25000 and IAPR TC-12 datasets respectively. Since our experimental performances are not sensitive to parameters λ, p and q compared with μ and γ, in order to simplify the tuning process and reduce the running time, we report the effects of each individual parameter while keeping other parameters fixed at the best if tuned or default 1 if not. While Figure 3 and 4 shows the performance variance w.r.t. λ and p, q respectively.

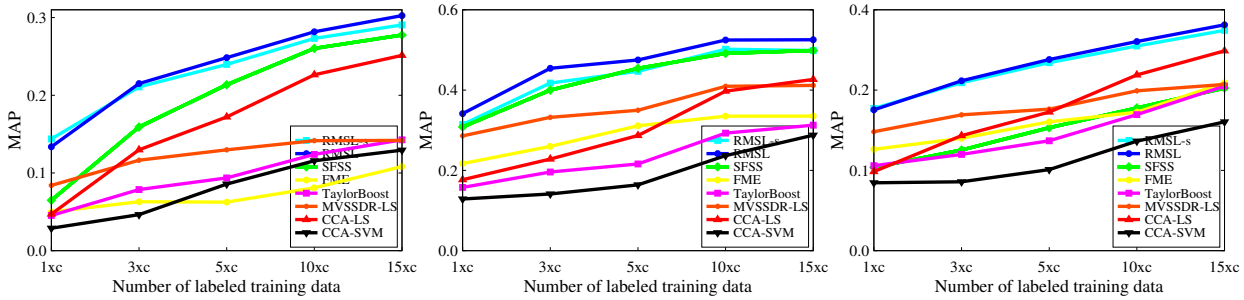
From Figure 2, we can see that an approximately analogous pattern for performance, i.e. three similar variance tendencies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 5: *Performance comparisons w.r.t. the number of labeled training data on three datasets when using LLC and FK features.*



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 6: *Performance comparisons w.r.t. the number of labeled training data on three datasets when using FC6 and FC7 features.*

Both Figure 2.(a), (b) and (c) show that as μ and γ increase from 10^{-6} to 10^{6}, the performance keeps going up; best μ and γ all appear in the right top areas, i.e. μ and γ both take a big value. Such phenomenon implies that both classifier error term and regularization term play important roles in finding the optimal classifier coefficients, that is both large parameters help to achieve the best results. For parameter λ, which controls the similarity among the outputs of different classifiers, the performance keeps little changed when λ varies in little values domain; as λ take a big value, the performance becomes worse sharply, for example, the MAP value becomes very low when λ takes 10^{6} for NUS-WIDE dataset, MIRFLICKR-25000 dataset and IAPR TC-12 dataset. Such phenomenon implies that too large value of parameter λ helps to achieve very poor results. As balance parameter λ controls the level of consistency between different views, the correlated and complemental information may become the same and disappear if the value of λ is too large. For parameters p and q, which can provide effective noise identification power during the training process, as showed in Figure 4 we can see there are three totally inverse variance trends for p. This implies NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12 datasets have different noise distributions and, thanks for the adaptability of $l_{2,p}$-norm, we can flexibly select appropriate p according to the data and noise distributions. From Figure 4, we also can find that the performance varies a little when parameter q changed from $\{0.1, 0.2, \ldots, 1.8, 1.9\}$. For this phenomena, because of the same ability of noise identification, q varies less sharply than p and maybe there will not be more gains if more $l_{2,p}$-norm are adopted.

Finally, we set all parameters to the values corresponding to the best performance for our subsequent experiments, i.e. $\mu = 10^{6}$, $\gamma = 10^{4}$, $\lambda = 10^{4}$, $p = 0.8$, and $q = 1.9$ for NUS-WIDE dataset, $\mu = 10^{6}$, $\gamma = 10^{2}$, $\lambda = 10^{2}$, $p = 0.4$, and $q = 1.5$ for MIRFLICKR-25000 dataset and $\mu = 10^{6}$, $\gamma = 10^{4}$, $\lambda = 10^{4}$, $p = 1.0$, and $q = 1.6$ for IAPR TC-12 dataset.

D. Comparison and Analysis

In this part, we evaluate the effectiveness of our proposed approach by comparing to other methods.

1) Comparison with Different Algorithms: First, we compare the proposed RMSL algorithm with other state-of-art algorithms using the fusion of two different features, for the CCA based algorithms can only address the situation of two views. As the algorithms SFSS and TaylorBoost are single-view methods, we directly concatenate different feature vectors to form a long feature vector for the subsequent experiments. Also, we add one more baseline - Laplacian Regularization + Robust loss on concatenated features (denoted as RMSL-s) to justify the need for multi-view term. The results of using LLC and FK features are shown in Figure 5 for NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12 datasets respectively. The results of using FC6 and FC7 are shown in Figure 6 for NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12 datasets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

TABLE I

Comparison of our approach and SFSS on single-view and multi-view for NUS-WIDE.

		LLC	FK	LLC+FK		FC6	FC7	FC6+FC7	
RMSL	$1 \times c$	5.94%	4.02%	5.30%		13.35%	13.48%	13.37%	
	$3 \times c$	6.64%	5.94%	7.06%		19.35%	19.21%	21.51%	
	$5 \times c$	7.00%	6.34%	7.54%		22.36%	21.81%	24.84%	
	$10 \times c$	7.54%	6.95%	8.36%		25.16%	25.40%	28.17%	
	$15 \times c$	7.44%	7.49%	8.69%		26.82%	27.28%	30.25%	
		LLC	FK	LLC+FK (ef)	LLC+FK (lf)	FC6	FC7	FC6+FC7 (ef)	FC6+FC7 (lf)
SFSS	$1 \times c$	5.93%	4.02%	5.49%	5.54%	8.19%	8.79%	9.34%	6.49%
	$3 \times c$	6.56%	5.52%	6.65%	7.09%	15.18%	15.16%	16.81%	15.92%
	$5 \times c$	7.06%	2.98%	2.98%	7.47%	18.41%	17.75%	20.47%	21.36%
	$10 \times c$	7.43%	2.97%	2.97%	7.78%	21.26%	21.48%	24.55%	26.03%
	$15 \times c$	7.32%	3.48%	3.49%	7.84%	22.16%	22.28%	26.19%	27.76%

TABLE II

Comparison of our approach and SFSS on single-view and multi-view for MIRFLICKR-25000.

		LLC	FK	LLC+FK		FC6	FC7	FC6+FC7	
RMSL	$1 \times c$	20.59%	16.53%	20.46%		30.77%	34.23%	34.14%	
	$3 \times c$	24.23%	18.76%	23.98%		39.45%	46.01%	45.43%	
	$5 \times c$	24.55%	22.47%	25.74%		43.43%	47.47%	47.50%	
	$10 \times c$	26.04%	23.40%	26.63%		48.82%	51.76%	52.47%	
	$15 \times c$	25.55%	22.13%	25.67%		48.85%	51.76%	52.54%	
		LLC	FK	LLC+FK (ef)	LLC+FK (lf)	FC6	FC7	FC6+FC7 (ef)	FC6+FC7 (lf)
SFSS	$1 \times c$	19.92%	15.73%	19.47%	19.45%	30.34%	34.41%	31.00%	30.81%
	$3 \times c$	24.27%	18.56%	23.53%	23.48%	38.99%	46.01%	40.19%	40.00%
	$5 \times c$	24.36%	22.35%	25.38%	25.32%	44.32%	49.04%	45.37%	45.42%
	$10 \times c$	25.36%	22.26%	25.94%	25.93%	48.11%	52.72%	49.24%	49.14%
	$15 \times c$	25.28%	21.59%	25.33%	25.36%	48.86%	53.70%	50.00%	49.87%

TABLE III

Comparison of our approach and SFSS on single-view and multi-view for IAPR TC-12.

		LLC	FK	LLC+FK		FC6	FC7	FC6+FC7	
RMSL	$1 \times c$	15.39%	12.89%	14.91%		20.95%	22.51%	21.05%	
	$3 \times c$	16.54%	14.83%	16.15%		24.75%	25.91%	25.39%	
	$5 \times c$	17.19%	15.87%	17.60%		27.86%	28.27%	28.57%	
	$10 \times c$	17.28%	17.12%	18.41%		30.36%	30.17%	31.26%	
	$15 \times c$	17.29%	18.15%	18.76%		32.65%	32.57%	33.75%	
		LLC	FK	LLC+FK (ef)	LLC+FK (lf)	FC6	FC7	FC6+FC7 (ef)	FC6+FC7 (lf)
SFSS	$1 \times c$	11.36%	9.17%	9.19%	9.16%	12.09%	16.99%	12.44%	14.41%
	$3 \times c$	12.99%	10.49%	10.08%	9.87%	14.87%	20.73%	15.05%	18.68%
	$5 \times c$	13.82%	9.79%	9.99%	9.79%	17.75%	25.81%	18.38%	24.18%
	$10 \times c$	14.08%	11.69%	12.02%	9.98%	20.90%	28.56%	21.33%	27.35%
	$15 \times c$	14.00%	12.93%	13.23%	10.72%	23.87%	31.97%	24.30%	30.79%

respectively. From Figure 5 and 6, we have the following observations.

- First, it can be seen from Figure 5 that RMSL gains the highest MAP value over other algorithms in almost all situations. Figure 6 has similar results.
- The performance can be gained when multiple features are used. The proposed multi-view RMSL algorithm outperform over the single-view SFSS algorithm which simply concatenates multiple features.
- Our late fusion of multiple features is effective. Although RMSL outperforms early fusion based algorithm, for instance SFSS by simply concatenating features, CCA by looking for maximally correlated directions in the two features spaces and MVSSDR-LS by learning the consensus pattern, it remains unclear between early and late fusion. Additionally, although our single-view version on concatenated features, i.e. RMSL-s, gains comparable performance compared with RMSL, it usually suffers from the burden of more computational resources and over-fitting problem especially in case of small training set. Also, it emphasizes the importance of robust loss compared with FME, and the need of multi-view term will be further justified in the next subsection.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

TABLE IV

Comparison of our approach on single-view and multi-view.

(a) *Single-view* ($15 \times c$)

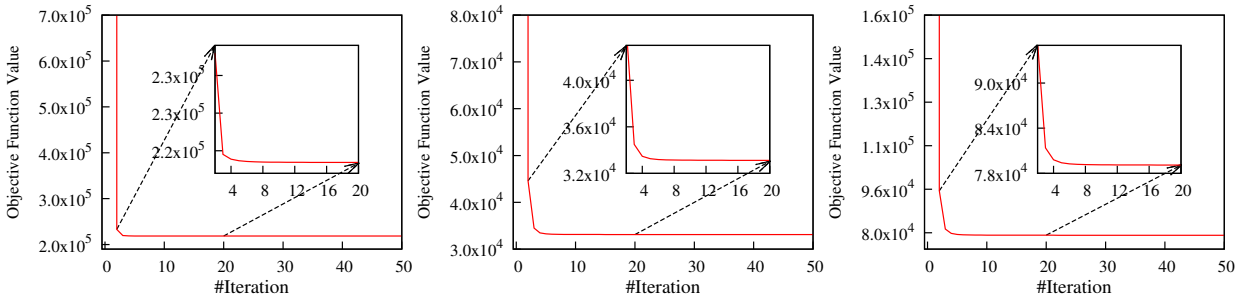
	LLC	FK	FC6	FC7
NUS-WIDE	7.44%	7.49%	26.82%	27.28%
MIRFLICKR-25000	25.55%	22.13%	48.85%	51.76%
IAPR TC-12	17.29%	18.15%	32.65%	32.57%

(b) *Two-view* ($15 \times c$)

	LLC+FK	FC6+FC7	LLC+FC6	LLC+FC7	FK+FC6	FK+FC7
NUS-WIDE	8.69%	30.25%	26.84%	27.30%	26.84%	27.28%
MIRFLICKR-25000	25.67%	52.54%	48.20%	51.73%	48.19%	51.74 %
IAPR TC-12	18.76%	33.75%	32.61%	32.73%	32.60%	32.73 %

(c) *Multi-view* ($15 \times c$)

	LLC+FK+FC6	LLC+FK+FC7	LLC+FC6+FC7	FK+FC6+FC7	LLC+FK+FC6+FC7
NUS-WIDE	26.83%	27.31%	30.28%	30.27%	30.29%
MIRFLICKR-25000	47.96%	51.74%	52.60%	52.60%	52.60%
IAPR TC-12	32.50%	32.88%	33.77%	33.76%	33.76%



(a) *NUS-WIDE* **(b)** *MIRFLICKR-25000* **(c)** *IAPR TC-12*

Fig. 7: *Convergence study.*

- The semi-supervised algorithms, i.e. our proposed RMSL and SFSS, gain higher MAP values than the supervised ones (TaylorBoost, CCA-LS and CCA-SVM), suggesting that additional performance improvement can be gained leveraging the unlabeled data for image annotation in these two datasets, especially when the number of labeled data is small.
- The FC6 and FC7 features based on deep learning are obviously more informative and discriminative than hand-crafted features, even if incorporating recent advance in encoding for bags of word model, which reveal the strong representative power of deep learning based features.

2) Performance Comparison between multi-view and single-view: Next, we conduct extensive experiments to validate the effectiveness of multi-view learning of the proposed method. First, we compare RMSL with SFSS by forming two two-view combinations (i.e., LLC+FK and FC6+FC7) from the four evaluated visual features, where two different features are directly concatenated (denote as 'ef' for early fusion) or averaged (denote as 'lf' for late fusion). The corresponding results are shown in Table I, II and III. We also test and evaluate our proposed approach in more views setting (i.e., three and four views). The experimental results are shown in Table IV. From Table I, II, III and IV, we have the following observations.

- It can be seen from Table I, II and III that RMSL always gains better performance on two two-view settings than single-view on all NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12 datasets.
- The proposed RMSL can gain more promotions than single-view, although we can gain improvements by simply employing the way of concatenating or averaging multiple features.
- Even we employ three features or four features, the performance may still increases a little, but not always. We can get more or less promotions when adding more views which may depend on the relations among the views and dataset distribution. Still, we can see the strong representative power of deep learning based features.

3) Convergence study: In this part, we conduct experiments to study the convergence of our iterative algorithm in NUS-WIDE, MIRFLICKR-25000 and IAPR TC-12 datasets respectively, as illustrated in Figure 7. Here, We fix all the five parameters, i.e. μ, γ, λ, p and q to 1. As we can see from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2717185, IEEE Transactions on Image Processing

Figure 7, our approach can converge very fast within only a few iterations, which indicates its efficiency for practical use.

V. Conclusions

In this paper, we proposed a new model for image annotation task, termed as Robust Multi-view Semi-supervised Learning (RMSL). In our proposed method, we apply graph Laplacian based semi-supervised learning to explore underlying data structural knowledge. In order to utilize the correlated and complemental information from different views to comprehensively depict data, pair-wise constraints were imposed on our model to guarantee consistency among different views. We then incorporate $l_{2,p}$-norm into our objective function to add flexibility and adaptability from our data, which can show effective noise identification power during the training process. Finally, thus we can boost image annotation performance in our model by integrating them jointly. We further devised an effective iterative algorithm to optimize the model. Extensive experiments on three real-world image data sets showed the effectiveness of the proposed RMSL method as compared to the state-of-the-art approaches. In future, we intend to gain more promotions by exploiting the correlations between multiple different labels, since image annotation can be regarded as a multi-label problem.

Appendix A
Proof of Lemma 1

Proof. We consider the following function

$$g(a) = pa^2 - 2a^p + (2-p), \tag{20}$$

where $p \in (0, 2)$. We expect to show that when $a > 0$, $g(a) \geq 0$. The first and second order derivatives of the function in Eq. (20) are $g'(a) = 2pa - 2pa^{p-1}$ and $g''(a) = 2p - 2p(p-1)a^{p-2}$, respectively. We can see that $a = 1$ is the only point that satisfies $g'(a) = 0$. Also, when $0 < a < 1$, $g'(a) < 0$ and when $a > 1$, $g'(a) > 0$. This means that $g(a)$ is monotonically decreasing when $0 < a < 1$ and monotonically increasing when $a > 1$. Moreover, we have $g''(1) = 2p(2-p) > 0$. Therefore, for $\forall a > 0$, $g(a) \geq g(1) = 0$.

Then, by substituting $a = \frac{\|\tilde{r}_t^i\|}{\|r_t^i\|}$ into Eq. (20), we obtain the conclusion

$$\begin{aligned}
& p\frac{\|\tilde{r}_t^i\|^2}{\|r_t^i\|^2} - 2\frac{\|\tilde{r}_t^i\|^p}{\|r_t^i\|^p} + (2-p) \geq 0, \\
\Leftrightarrow \quad & p\|\tilde{r}_t^i\|^2 - 2\|\tilde{r}_t^i\|^p\|r_t^i\|^{2-p} + (2-p)\|r_t^i\|^2 \geq 0, \\
\Leftrightarrow \quad & p\|\tilde{r}_t^i\|^2\|r_t^i\|^{p-2} - 2\|\tilde{r}_t^i\|^p + (2-p)\|r_t^i\|^p \geq 0, \\
\Leftrightarrow \quad & 2\|\tilde{r}_t^i\|^p - p\|\tilde{r}_t^i\|^2\|r_t^i\|^{p-2} \leq (2-p)\|r_t^i\|^p, \\
\Leftrightarrow \quad & \|\tilde{r}_t^i\|^p - \frac{p\|\tilde{r}_t^i\|^2}{2\|r_t^i\|^{2-p}} \leq \|r_t^i\|^p - \frac{p\|r_t^i\|^2}{2\|r_t^i\|^{2-p}}.
\end{aligned}$$

Appendix B
Proof of Theorem 1

Proof. Denote $\tilde{R}_t = X_t^T\tilde{W}_t + \mathbf{1}_n\tilde{b}_t^T - \tilde{F}_t$, $\tilde{R}_{ts} = \tilde{F}_t - \tilde{F}_s$ and $\mathcal{R}(F_t, W_t) = Tr(F_t^T L_t F_t) + Tr((F_t - Y)^T U(F_t - Y)) + \mu\gamma\|W_t\|_F^2)$. Suppose $\tilde{F}_t, \tilde{W}_t, \tilde{b}_t$ are the optimized solution of the alternative problem (6), then we obtain the following conclusion:

$$\begin{aligned}
& \sum_t \Big(\mathcal{R}(\tilde{F}_t, \tilde{W}_t) + \mu Tr(\tilde{R}_t^T D_t^l \tilde{R}_t)\Big) + \sum_{t,s} \lambda Tr(\tilde{R}_{ts}^T D_{ts}^r \tilde{R}_{ts}) \leq \\
& \sum_t \Big(\mathcal{R}(F_t, W_t) + \mu Tr(R_t^T D_t^l R_t)\Big) + \sum_{t,s} \lambda Tr(R_{ts}^T D_{ts}^r R_{ts}) \\
& \Rightarrow \sum_t \left(\mathcal{R}(\tilde{F}_t, \tilde{W}_t) + \mu \sum_{i=1}^n \frac{p\|\tilde{r}_t^i\|^2}{2\|r_t^i\|^{2-p}}\right) + \lambda \sum_{t,s}\sum_{i=1}^n \frac{p\|\tilde{r}_{ts}^i\|^2}{2\|r_{ts}^i\|^{2-p}} \\
& \leq \sum_t \left(\mathcal{R}(F_t, W_t) + \mu \sum_{i=1}^n \frac{p\|r_t^i\|^2}{2\|r_t^i\|^{2-p}}\right) + \lambda \sum_{t,s}\sum_{i=1}^n \frac{p\|r_{ts}^i\|^2}{2\|r_{ts}^i\|^{2-p}} \\
\Rightarrow & \sum_t \left(\mathcal{R}(\tilde{F}_t, \tilde{W}_t) + \mu \sum_{i=1}^n \|\tilde{r}_t^i\|^p - \mu\left(\sum_{i=1}^n \|\tilde{r}_t^i\|^p - \sum_{i=1}^n \frac{p\|\tilde{r}_t^i\|^2}{2\|r_t^i\|^{2-p}}\right)\right) \\
& + \lambda \sum_{t,s}\left(\sum_{i=1}^n \|\tilde{r}_{ts}^i\|^p - \left(\sum_{i=1}^n \|\tilde{r}_{ts}^i\|^p - \sum_{i=1}^n \frac{p\|\tilde{r}_{ts}^i\|^2}{2\|r_{ts}^i\|^{2-p}}\right)\right) \\
\leq & \sum_t \left(\mathcal{R}(F_t, W_t) + \mu \sum_{i=1}^n \|r_t^i\|^p - \mu\left(\sum_{i=1}^n \|r_t^i\|^p - \sum_{i=1}^n \frac{p\|r_t^i\|^2}{2\|r_t^i\|^{2-p}}\right)\right) \\
& + \lambda \sum_{t,s}\left(\sum_{i=1}^n \|r_{ts}^i\|^p - \left(\sum_{i=1}^n \|r_{ts}^i\|^p - \sum_{i=1}^n \frac{p\|r_{ts}^i\|^2}{2\|r_{ts}^i\|^{2-p}}\right)\right).
\end{aligned}$$

Given the conclusion of Lemma 2, we finally arrive at

$$\begin{aligned}
& \Rightarrow \sum_t \left(\mathcal{R}(\tilde{F}_t, \tilde{W}_t) + \mu \sum_{i=1}^n \|\tilde{r}_t^i\|^p\right) + \lambda \sum_{t,s}\left(\sum_{i=1}^n \|\tilde{r}_{ts}^i\|^p\right) \\
& \leq \sum_t \left(\mathcal{R}(F_t, W_t) + \mu \sum_{i=1}^n \|r_t^i\|^p\right) + \lambda \sum_{t,s}\left(\sum_{i=1}^n \|r_{ts}^i\|^p\right).
\end{aligned}$$

Hence, the value of the objective function in Eq. (5) monotonically decreases in each iteration.

References

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] W. Liu, H. Liu, D. Tao, Y. Wang, and K. Lu, "Manifold regularized kernel logistic regression for web image annotation," *Neurocomputing*, vol. 172, pp. 3–8, 2016.

[3] X. Li, B. Shen, B. Liu, and Y. Zhang, "A locality sensitive low-rank model for image tag completion," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 474–483, 2016.

[4] C. Luo, B. Ni, S. Yan, and M. Wang, "Image classification by selective regularized subspace learning," *IEEE Trans. Multimedia*, vol. 18, no. 1, pp. 40–50, 2016.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, 2008.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Computing, Comm. and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[7] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," in *Proc. ACM Multimedia*, 2007.

[8] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, 2007.

[9] X. Wang, L. Zhang, X. Li, and W. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, 2008.

[10] L. Cao, J. Luo, H. A. Kautz, and T. S. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 208–219, 2009.

[11] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. ICCV*, 2009.

[12] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. CVPR*, 2015.

[13] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1339–1351, 2012.

[14] A. Fakeri-Tabrizi, M. Amini, and P. Gallinari, "Multiview semi-supervised ranking for automatic image annotation," in *Proc. ACM Multimedia*, 2013.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. CVPR*, 2010.

[16] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. ECCV*, 2010.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS 25*, 2012.

[18] G. Carneiro and N. Vasconcelos, "Formulating semantic image annotation as a supervised learning problem," in *Proc. CVPR*, 2005.

[19] W. Liu and D. Tao, "Multiview hessian regularization for image annotation," *IEEE Trans. Image Processing*, vol. 22, no. 7, pp. 2676–2687, 2013.

[20] M. Hu, Y. Yang, H. Zhang, F. Shen, J. Shao, and F. Zou, "Multi-view semi-supervised learning for web image annotation," in *Proc. ACM Multimedia*, 2015.

[21] 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 *Proc. CIVR*, 2009, pp. 48:1–48:9.

[22] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. MIR*, 2008.

[23] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in *International Workshop OntoImage*, vol. 5, 2006, p. 10.

[24] H. J. Escalante, C. A. Hernández, J. A. González, A. López-López, M. Montes-y-Gómez, E. F. Morales, L. E. Sucar, L. V. Pineda, and M. Grubinger, "The segmented and annotated IAPR TC-12 benchmark," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 419–428, 2010.

[25] H. Ma, J. Zhu, M. R. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, 2010.

[26] H. Zhang, Z. Zha, Y. Yang, S. Yan, Y. Gao, and T. Chua, "Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval," in *Proc. ACM Multimedia*, 2013.

[27] C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using svm," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2003, pp. 330–338.

[28] K. Goh, E. Y. Chang, and B. Li, "Using one-class and two-class svms for multiclass image annotation," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 10, pp. 1333–1346, 2005.

[29] X. Qi and Y. Han, "Incorporating multiple svms for automatic image annotation," *Pattern Recognition*, vol. 40, no. 2, pp. 728–741, 2007.

[30] Y. Gao, J. Fan, X. Xue, and R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers," in *Proc. ACM Multimedia*, 2006.

[31] R. Socher and F. Li, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. CVPR*, 2010.

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

[33] W. Li and M. Sun, "Semi-supervised learning for image annotation based on conditional random fields," in *Proc. CIVR*, 2006.

[34] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Processing*, vol. 22, no. 2, pp. 523–536, 2013.

[35] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[36] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM Multimedia*, 2010.

[37] H. Yang, J. T. Zhou, and J. Cai, "Improving multi-label learning with missing labels by structured semantic correlations," in *Proc. ECCV*, 2016.

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

[39] X. Zhu, "Semi-supervised learning literature survey," 2005.

[40] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *NIPS 16*, 2003.

[41] Z. Ma, Y. Yang, F. Nie, J. R. R. Uijlings, and N. Sebe, "Exploiting the entire feature space with sparsity for automatic image annotation," in *Proc. ACM Multimedia*, 2011.

[42] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.

[43] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[44] X. Sun, M. Chen, and A. G. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. CVPR*, 2009.

[45] C. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. ACM Multimedia*, 2005.

[46] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmák, "Two view learning: Svm-2k, theory and practice," in *NIPS 18*, 2005, pp. 355–362.

[47] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, 2011.

[48] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan, "Multimodal feature fusion for robust event detection in web videos," in *Proc. CVPR*, 2012.

[49] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 572–581, 2013.

[50] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.

[51] X. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang, "Multi-label dictionary learning for image annotation," *IEEE Trans. Image Processing*, vol. 25, no. 6, pp. 2712–2725, 2016.

[52] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. SIGIR*, 2003.

[53] D. Cai, X. He, Z. Li, W. Ma, and J. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in *Proc. ACM Multimedia*, 2004.

[54] J. Johnson, L. Ballan, and F. Li, "Love thy neighbors: Image annotation by exploiting image metadata," in *Proc. ICCV*, 2015.

[55] F. Wu, Z. Wang, Z. Zhang, Y. Yang, J. Luo, W. Zhu, and Y. Zhuang, "Weakly semi-supervised deep learning for multi-label image annotation," *IEEE Trans. Big Data*, vol. 1, no. 3, pp. 109–122, 2015.

[56] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. ACM Multimedia*, 2009.

[57] C. Hou, C. Zhang, Y. Wu, and F. Nie, "Multiple view semi-supervised dimensionality reduction," *Pattern Recognition*, vol. 43, no. 3, pp. 720–730, 2010.

[58] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. ECCV*, 2008.

[59] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, pp. 1150–1157.

[60] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. BMVC*, 2011.

[61] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[62] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014.

[63] M. J. Saberian, H. Masnadi-Shirazi, and N. Vasconcelos, "Taylorboost: First and second-order boosting algorithms with explicit margin control," in *Proc. CVPR*, 2011.