

# Cross-View Feature Learning for Scalable Social Image Analysis

Wenxuan Xie and Yuxin Peng\* and Jianguo Xiao

Institute of Computer Science and Technology, Peking University, Beijing 100871, China  
{xiewenxuan, pengyuxin, xiaojianguo}@pku.edu.cn

## Abstract

Nowadays images on social networking websites (e.g., Flickr) are mostly accompanied with user-contributed tags, which help cast a new light on the conventional content-based image analysis tasks such as image classification and retrieval. In order to establish a scalable social image analysis system, two issues need to be considered: 1) Supervised learning is a futile task in modeling the enormous number of concepts in the world, whereas unsupervised approaches overcome this hurdle; 2) Algorithms are required to be both spatially and temporally efficient to handle large-scale datasets. In this paper, we propose a cross-view feature learning (CVFL) framework to handle the problem of social image analysis effectively and efficiently. Through explicitly modeling the relevance between image content and tags (which is empirically shown to be visually and semantically meaningful), CVFL yields more promising results than existing methods in the experiments. More importantly, being general and descriptive, CVFL and its variants can be readily applied to other large-scale multi-view tasks in unsupervised setting.

## Introduction

Over the past years, content-based image analysis tasks such as image classification and retrieval (Smeulders et al. 2000; Torres et al. 2009) have always been plagued with the gap between low-level representation and high-level semantics, i.e., the semantic gap. How to construct visual representations that are able to properly reflect the underlying semantic meaning remains to be an open problem. However, fortunately, whilst it may be difficult to bridge the semantic gap by diving solely into the image content, the advent of social networking websites (e.g., Flickr) brings new opportunities to the content-based image analysis problem by providing user-contributed tags for social images. Although being inaccurate and incomplete sometimes, tags are easily available and beneficial for social image analysis.

To state conveniently, a social image in this paper consists of three components: image, tag and label. *Image* refers to the visual content, *tag* refers to the associated user-contributed tags, and *label* refers to the semantic concepts.

\*Corresponding author.

Through utilizing the above three components, many supervised and semi-supervised approaches to social image analysis have been proposed, e.g., multiple kernel learning (Lanckriet et al. 2004; Wang et al. 2010), feature selection (Guyon and Elisseeff 2003; Xu et al. 2010) and distance metric learning (Yang and Jin 2006; Bilenko, Basu, and Mooney 2004).

However, despite the effectiveness of the aforementioned approaches, we still need to consider two important issues in order to establish a scalable social image analysis system:

**Numerous concepts.** Large datasets are accompanied with many more concepts. For instance, ImageNet currently counts approximately 22K concepts (Deng et al. 2009), which makes it a futile task to model so many, and often visually similar, concepts. What is more, the number of concepts in the world can be far more than 22K. Instead of modeling so many concepts, unsupervised approaches<sup>1</sup> overcome this hurdle. As a consequence, we focus on unsupervised approaches in this paper, i.e., only image content and tags can be utilized.

**Scalability.** In the presence of enormous number of social images on the web, algorithms that lack spatial or temporal efficiency are prohibited. Therefore, in order to handle large datasets, we focus on algorithms whose time and space complexities are both limited to at most  $O(N)$  with respect to data size  $N$  in this paper.

Under the constraints of the above two issues, algorithms which are able to handle the problem of scalable social image analysis should be efficiently defined over feature representations. For example, feature combination (Atrey et al. 2010) is a straightforward method. Despite its simplicity, feature combination has been demonstrated empirically to be effective, which may be due to the fact that image content and tags are two different views and are complementary. Based on the combined feature representation, principal component analysis (PCA) (Hotelling 1933) extracts compressed representation by maximizing the variance in the principal subspace. What is more, instead of simply combining image and tags, algorithms such as canonical correlation analysis (CCA) (Hotelling 1936) have been proposed

<sup>1</sup>Image-tag pair is also a kind of supervision information. However, in this paper, we only refer supervision information to the *label* of a social image.

to model the relevance between these two views. CCA maximizes the correlation between visual and textual representations by learning two linear mappings. Similar to CCA, partial least squares regression (PLSR) (Wold 1985) pursues the direction which maximizes the covariance between these two views.

However, despite the advantages, CCA involves a generalized eigenvalue problem, which is computationally expensive to solve. Furthermore, it is challenging to derive variants of CCA, e.g., sparse CCA. As sparse CCA involves a difficult sparse generalized eigenvalue problem, convex relaxation of sparse CCA has been studied in (Sriperumbudur, Torres, and Lanckriet 2007), where the exact formulation has been relaxed in several steps. Therefore, although CCA can be regularized to prevent the overfitting and avoid the singularity (Bach and Jordan 2003), it is generally difficult for CCA to be compatible with many other kinds of regularizers, e.g., sparsity and group sparsity.

In this paper, we propose a cross-view feature learning (CVFL) framework to explicitly model the relevance between image and tags by learning a linear mapping from textual representation to visual representation. In contrast to CCA, the relevance learned by CVFL is shown to be visually and semantically meaningful. More notably, CVFL is a general framework and can be readily applied to other large-scale multi-view tasks by imposing other kinds of regularizers. Moreover, besides its adaptability, CVFL is shown to be more effective and efficient empirically.

The rest of this paper is organized as follows. Section 2 presents a brief overview of related studies. The CVFL framework is introduced in Section 3. Section 4 justifies CVFL by showing its technical soundness. In order to dive deeper into the problem, more details on the relation between CVFL and CCA are shown in Section 5. In Section 6, the performance of our proposed method is evaluated on two real-world datasets in image classification. Finally, Section 7 concludes our paper.

## Related Work

In this section, we present a brief overview of related studies and discuss their similarities and differences with the proposed CVFL. These studies include multiple kernel learning, feature selection, distance metric learning and latent semantic learning.

### Multiple Kernel Learning

Given multiple kernels for feature combination, it is non-trivial to determine weights of each kernel. To tackle this issue, supervised multiple kernel learning (MKL) (Lanckriet et al. 2004) algorithms have been proposed to determine these weights based on the max-margin criterion, and have achieved state-of-the-art performance on some image recognition applications (Vedaldi et al. 2009). Moreover, semi-supervised MKL (Wang et al. 2010) has also been proposed to utilize unlabeled data. However, since these approaches require labels, it may be challenging to apply these approaches to problems with huge numbers of concepts.

### Feature Selection

Feature selection is aiming at selecting a subset of features satisfying some predefined criteria, which is essentially a computationally expensive combinatorial optimizing problem which is NP-hard (Amaldi and Kann 1998). The criteria of most feature selection algorithms are related to labels (Guyon and Elisseeff 2003; Xu et al. 2010). Besides, in cases where no labels are provided, unsupervised feature selection (Xing and Karp 2001; Cai, Zhang, and He 2010) is applicable by using criteria such as saliency, entropy, etc. However, as stated above, feature selection involves a combinatorial optimizing problem which is time-consuming to solve. As an example, (Cai, Zhang, and He 2010) requires a time complexity of  $O(N^2)$  with respect to data size  $N$ , and thus may face challenges in dealing with large-scale tasks. What is more, feature selection does not take into account the relevance between multiple views.

### Distance Metric Learning

Most distance metric learning (DML) algorithms (Yang and Jin 2006; Bilenko, Basu, and Mooney 2004) learn a Mahalanobis distance matrix. Due to the positive semi-definiteness of the matrix, the learned result can be derived as a linear mapping of the original feature, which is similar to the proposed CVFL at this point. However, most DML methods require triplets  $(\mathbf{x}, \mathbf{x}_+, \mathbf{x}_-)$  as inputs, which indicate that image  $\mathbf{x}$  and  $\mathbf{x}_+$  are similar/relevant, while image  $\mathbf{x}$  and  $\mathbf{x}_-$  are dissimilar/irrelevant. In a social image analysis system, similar pairs can be easily obtained (i.e., image-tag pairs), but dissimilar pairs are generally not directly available due to the fact that the dissimilar information is often related to labels (Wu et al. 2013) or user feedback (Xia, Wu, and Hoi 2013). Besides, (Li et al. 2012) only utilizes image-tag pairs to derive a distance metric; however, the spatial and temporal complexities are both  $O(N^2)$ .

### Latent Semantic Learning

Many algorithms have been proposed to model the correspondence of different views in automatic image annotation. Probabilistic graphical models (Putthividhy, Attias, and Nagarajan 2010; Jia, Salzmann, and Darrell 2011) aim to describe the correlations between images and tags by learning a correspondence between visual representation and textual representation in order to predict annotations of a new test image. Different from the aim of the above methods (which is to predict tags given a new image), the objective of CVFL is to derive a descriptive representation for social image analysis *given both image and tags*. More notably, CVFL deals with a convex optimization problem and the solution does not rely on a time-consuming iterative optimization procedure.

### The CVFL Framework

In this section, we introduce the CVFL framework by giving some preliminaries first. After that, we present the formulation and solution to CVFL, respectively. Finally, the complexity issue is discussed.

## Preliminaries

Unless otherwise specified,  $X \in \mathbb{R}^{N \times M_1}$  and  $V \in \mathbb{R}^{N \times M_2}$  denote feature representations in two heterogeneous views, respectively (e.g., image and tag representation in this paper).  $U \in \mathbb{R}^{M_2 \times M_1}$  denotes a linear mapping. In this case, the number of samples is  $N$ , and the feature dimensionalities in the two views are  $M_1$  and  $M_2$ . Moreover,  $X_{i \cdot}$ ,  $X_{\cdot j}$  and  $X_{ij}$  respectively denote the  $i$ -th row vector,  $j$ -th column vector and  $(i, j)$ -th element of the matrix  $X$ . In other words,  $X_{i \cdot}$  is the  $i$ -th sample, and  $X_{\cdot j}$  represents the  $j$ -th feature.

## Formulation

In order to effectively handle social image analysis, it is important to derive a descriptive feature representation. In this paper, we propose to embed the textual semantics into the visual representation through explicitly modeling the relevance between image and tags. To ensure its descriptive power, the learned representation (denoted as  $\hat{X}$ ) should encode the information from both  $X$  and  $V$ . With these considerations in mind, we have the following equation.

$$\min_{\hat{X}, U} \|\hat{X} - X\|_F^2 \quad \text{s. t.} \quad \hat{X} = VU \quad (1)$$

As shown in Eq. 1, by constraining the resultant representation  $\hat{X}$  to be close to  $X$  and to be a linear mapping of  $V$ ,  $\hat{X}$  obtains information conveyed by both  $X$  and  $V$  (which will be justified later in Section 4). Being a reconstruction formulation, Eq. 1 is compatible with many regularizers such as graph Laplacian and sparsity. As a first step, we pose an  $L_2$  regularizer to avoid overfitting and ensure numerical stability.

$$\min_{\hat{X}, U} \|\hat{X} - X\|_F^2 + \lambda \|U\|_F^2 \quad \text{s. t.} \quad \hat{X} = VU \quad (2)$$

where  $\lambda$  is a regularization parameter. Furthermore, based on the fact that visual features (i.e.,  $X_{\cdot j}$ ) are linearly reconstructed by textual features, we may derive a constraint on the reconstruction weights by modeling the affinity of visual features. As long as we constrain the difference between the reconstruction weights of similar visual features to be small, we may arrive at the following Laplacian regularizer (Belkin and Niyogi 2001).

$$\sum_{i,j} \|U_{\cdot i} - U_{\cdot j}\|_2^2 A_{ij} = \text{tr}(ULU^\top) \quad (3)$$

where  $A_{ij}$  denotes the affinity of the  $i$ -th and  $j$ -th visual features, and  $L \in \mathbb{R}^{M_1 \times M_1}$  is a Laplacian matrix. The normalized Laplacian is defined as  $L = I - G^{-1/2}AG^{-1/2}$ , where  $G$  is a diagonal matrix with its  $(i, i)$ -th element equal to the sum of the  $i$ -th column vector of  $A$ . For conciseness, we obtain the pairwise affinity graph  $A$  by computing the inner products of all observations shown as follows, although there are a few studies on constructing better graphs (Yan and Wang 2009; Lu and Peng 2013).

$$A = X^\top X \quad (4)$$

It should be noted that, many graph-based approaches model the affinity of samples (Li et al. 2012), and thus involves an  $N \times N$  graph which is difficult to compute and

store in large-scale datasets. In contrast, CVFL models the affinity of visual features, which is irrelevant to the data size  $N$ . As a consequence, by imposing the above Laplacian regularizer onto Eq. 2, we obtain the following objective function, which is the formulation of CVFL.

$$\min_{\hat{X}, U} \|\hat{X} - X\|_F^2 + \lambda \|U\|_F^2 + \gamma \text{tr}(ULU^\top) \quad \text{s. t.} \quad \hat{X} = VU \quad (5)$$

where  $\gamma$  is also a regularization parameter.

## Solution

The formulation of CVFL shown in Eq. 5 is a constrained optimization problem. By substituting the constraint  $\hat{X} = VU$  into Eq. 5, we may arrive at the following unconstrained optimization function

$$\min_U \|VU - X\|_F^2 + \lambda \|U\|_F^2 + \gamma \text{tr}(ULU^\top) \quad (6)$$

which is a convex optimization problem. By computing the derivative of Eq. 6 with respect to  $U$  and set it to 0, we can obtain the following equation after some algebra

$$(V^\top V + \lambda I_{M_2})U + \gamma UL - V^\top X = 0 \quad (7)$$

where  $I_{M_2}$  stands for an  $M_2 \times M_2$  identity matrix. Eq. 7 is exactly the form of a Lyapunov-like equation in control theory, i.e., the Sylvester equation, whose standard form is

$$S_1 U + U S_2 + S_3 = 0 \quad (8)$$

where  $S_1 = V^\top V + \lambda I_{M_2}$ ,  $S_2 = \gamma L$ , and  $S_3 = -V^\top X$  here. Eq. 8 can be rewritten as follows

$$(I_{M_1} \otimes S_1 + S_2^\top \otimes I_{M_2}) \cdot \text{vec}(U) = -\text{vec}(S_3) \quad (9)$$

where  $\otimes$  is the Kronecker product defined as  $U \otimes V = [U_{ij} \cdot V]$  for any two matrices  $U$  and  $V$ , and  $\text{vec}(S_3)$  is the unfolded vector of matrix  $S_3$ . Then, Eq. 9 can be solved by a linear equation shown as follows.

$$\text{vec}(U) = -(I_{M_1} \otimes S_1 + S_2^\top \otimes I_{M_2})^{-1} \cdot \text{vec}(S_3) \quad (10)$$

After obtaining  $U$ , we can derive  $\hat{X} \in \mathbb{R}^{N \times M_1}$  by computing  $\hat{X} = VU$  as the final representation for the social image analysis task.

## Computational Complexity

Solving Eq. 8 requires  $O(\max(M_1, M_2)^3)$ . Moreover, computing  $V^\top V$  and  $V^\top X$  in Eq. 7 may both incur a time complexity of  $O(NM_1M_2)$ , and the complexity of constructing a Laplacian  $L$  is  $O(NM_1^2)$ . In summary, the total complexity is  $O(NM_1M_2 + NM_1^2 + \max(M_1, M_2)^3)$ , which is linear with respect to the data size  $N$ . Therefore, CVFL can perform efficiently as the data size increases. Notably, however, feature dimensionalities (i.e.,  $M_1$  and  $M_2$ ) generally remain to be moderate and stable in practice.

## Justification of CVFL

The underlying rationale of CVFL is presented in this section. We begin by justifying the effectiveness, and then illustrate the learned relevance between image content and tags with a toy example.

## Justification

Since Eq. 5 is a regularized form of Eq. 1, for simplicity, we demonstrate that the learned representation  $\hat{X}$  encodes the information from both  $X$  and  $V$  according to Eq. 1. To begin with, since the squared error between  $\hat{X}$  and  $X$  is minimized,  $\hat{X}$  is constrained to be close to  $X$  and thus obtains the information conveyed by  $X$ . However, it is nontrivial to demonstrate the relation between  $\hat{X}$  and  $V$ .

Inspired by locality-sensitive hashing (Gionis, Indyk, and Motwani 1999) which constructs randomized hash functions by using random projections, we can establish distance-preserving mappings through using random projections. Random projection is based on the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984) and can be extended to the following theorem (Vempala 2004), where inner products are preserved under random projection.

**Theorem 1** *Let  $p, q \in \mathbb{R}^{M_2}$  and that  $\|p\| \leq 1$  and  $\|q\| \leq 1$ . Let  $f(x) = \frac{1}{\sqrt{M_1}}Ux$  where  $U$  is an  $M_1 \times M_2$  matrix, where each entry is sampled i.i.d from a Gaussian  $\mathcal{N}(0, 1)$ . Then,  $\Pr(|p \cdot q - f(p) \cdot f(q)| \geq \epsilon) \leq 4e^{-(\epsilon^2 - \epsilon^3)k/4}$ .*

where  $\Pr(\cdot)$  denotes a probability. Besides sampling i.i.d from a normalized Gaussian, sparse random projections (Achlioptas 2001) are also applicable. What is more, the authors in (Bingham and Mannila 2001) have stated that:

*In random projection, the original  $d$ -dimensional data is projected to a  $k$ -dimensional ( $k \ll d$ ) subspace base through the origin, using a random  $k \times d$  matrix whose columns have unit lengths.*

Based on the above statement, we now demonstrate that the columns of the linear mapping matrix  $U$  have unit lengths. Given that all the features in  $X$  and  $V$  have been normalized to zero mean and unit variance, column vectors  $X_{\cdot i}$  and  $V_{\cdot j}$  can be viewed as normalized Gaussian  $\mathcal{N}(0, 1)$ . Furthermore, being constrained to be close to  $X_{\cdot i}$ ,  $\hat{X}_{\cdot i}$  can also be viewed as a normalized Gaussian. Therefore, we have

$$\hat{X} = VU \implies \hat{X}_{\cdot i} = VU_{\cdot i}$$

and

$$\begin{aligned} VU_{\cdot i} &= \sum_{j=1}^{M_2} U_{ji} \cdot V_{\cdot j} \implies VU_{\cdot i} \sim \sum_{j=1}^{M_2} U_{ji} \cdot \mathcal{N}(0, 1) \\ \implies VU_{\cdot i} &\sim \sum_{j=1}^{M_2} \mathcal{N}(0, U_{ji}^2) \implies VU_{\cdot i} \sim \mathcal{N}(0, \sum_{j=1}^{M_2} U_{ji}^2) \end{aligned}$$

Since  $\hat{X}_{\cdot i} \sim \mathcal{N}(0, 1)$ , the equation  $\sum_{j=1}^{M_2} U_{ji}^2 = 1$  holds, which demonstrates that the columns of the linear mapping matrix  $U$  have unit lengths, and that the conditions in Theorem 1 are satisfied.

If we let  $p$  and  $q$  be the  $i$ -th and  $j$ -th textual samples (i.e.,  $V_{\cdot i}$  and  $V_{\cdot j}$ ) respectively,  $f(p)$  and  $f(q)$  respectively denote  $\hat{X}_{\cdot i}$  and  $\hat{X}_{\cdot j}$ .  $k$  is the dimensionality of visual representation (say 1,000 in practice) and  $\epsilon$  is a parameter (say 0.1 in the current case). As a consequence, the following inequality holds according to Theorem 1.

$$\Pr(|V_{\cdot i} \cdot V_{\cdot j} - \hat{X}_{\cdot i} \cdot \hat{X}_{\cdot j}| \geq 0.1) \leq 0.105$$

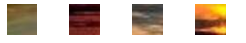




Sample patches of a visual feature	Top two words
	clouds bay
	runway bowing
	city mountain
	flag helicopter
	branch leaf

Figure 1: A toy example of some visual features and their top two relevant textual features. Visual features are represented by their corresponding unquantized image patches.

It can be observed that the inner products are preserved with a high probability, and thus we may conclude that  $\hat{X}$  encodes the information conveyed by  $V$ .

## Illustration of the Learned Relevance

CVFL explicitly models the relevance between image and tags by learning a linear mapping from textual representation to visual representation. To better illustrate the learned relevance, we interpret the model from the perspective of weighted summation. Based on Eq. 1, we have

$$\min_U \|VU - X\|_F^2 \quad (11)$$

Since the squared Frobenius norm is always nonnegative, the minimum of Eq. 11 is 0 if  $X = VU$  holds. Hence, we may arrive at

$$X_{\cdot i} = \sum_{j=1}^{M_2} V_{\cdot j} U_{ji} \quad (12)$$

It means that the  $i$ -th visual feature equals to a weighted sum of all the  $M_2$  textual features. The higher the weight  $U_{ji}$  (which can be interpreted as a relative importance), the more relevant the visual feature  $X_{\cdot i}$  and the textual feature  $V_{\cdot j}$  are. It should be noted that, a visual feature is a cluster center derived by image patches (and thus can be represented by these patches), and a textual feature is a word. To illustrate the learned relevance, we pick out the textual features which are most relevant to a given visual feature by sorting  $U_{ji}$  in descending order, where  $j \in \{1, \dots, M_2\}$ . Fig. 1 illustrates a toy example of some visual features (i.e., image patches) and their corresponding top two relevant textual features (i.e., words), which is derived from a subset of the Corel-5K dataset (Duygulu et al. 2002). It can be observed that the learned relevance between visual features and textual features is both visually and semantically meaningful.

## Relation to CCA

In this section, we discuss the relation between CCA and the proposed CVFL, both of which model the relevance between two views. CCA maximizes the correlation coefficient, whereas CVFL learns a linear mapping from one view

to another. In (Sun, Ji, and Ye 2011), the authors demonstrate that CCA in the multi-label classification case (where one of the views used in CCA is derived from the labels) can be formulated as a least squares problem under some conditions. Take Eq. 11 (unregularized form of CVFL) as an example, if  $\text{rank}(X) = M_1$  and  $\text{rank}(V) = N - 1$ , CCA and least squares regression are equivalent, i.e., the learned linear mappings of both the two approaches are the same.

However, with the growing number of images, the data size  $N$  is generally larger than feature dimensionalities  $M_1$  and  $M_2$  in practice. Hence, the condition  $\text{rank}(V) = N - 1$  does not hold, and CVFL is different from CCA. Fortunately, despite the difference, the technical soundness of CVFL can be justified by resorting to the theory of random projection (as shown in the previous section).

The advantage of CVFL is twofold. On one hand, CCA involves a time-consuming generalized eigenvalue problem (Watkins 2004), whereas it is more efficient to solve a least squares problem such as CVFL. On the other hand, it is challenging to derive the formulation if regularizers (such as smoothness<sup>2</sup>, sparsity and group sparsity) are added to CCA. In contrast, CVFL is generally compatible with all such regularizers, and thus can be made more descriptive by applying different regularizers.

Furthermore, we show how CVFL can be extended to deal with more than two views. To begin with, we relax the constraints in Eq. 5 to be

$$\min_{\hat{X}, U} \|\hat{X} - X\|_F^2 + R(U) + \mu \|\hat{X} - VU\|_F^2 \quad (13)$$

where  $R(U)$  denotes the regularizers on  $U$  (which can also be substituted to sparsity constraints, etc). In cases where there are more than two views (say  $X$ ,  $V_1$  and  $V_2$ ), Eq. 13 can be extended to the following form

$$\begin{aligned} \min_{\hat{X}, U_1, U_2} \|\hat{X} - X\|_F^2 + R(U_1) + \mu_1 \|\hat{X} - V_1 U_1\|_F^2 \\ + R(U_2) + \mu_2 \|\hat{X} - V_2 U_2\|_F^2 \end{aligned} \quad (14)$$

Consequently, the learned representation  $\hat{X}$  encodes the information conveyed by  $X$ ,  $V_1$  and  $V_2$ .

## Experiments

In this section, we evaluate the performance of the proposed CVFL in social image analysis. We begin by describing the experimental setup. As a next step, empirical results of CVFL and other related methods are reported. Finally, we present the parameter tuning details and discuss the complexity issues. It should be noted that, CVFL can be readily applied to other multi-view tasks in unsupervised setting (e.g., social image retrieval without labeled relevant/irrelevant examples), although the evaluation metric adopted in this paper is classification accuracy.

<sup>2</sup>This paper efficiently models the smoothness of visual features (as shown in Eq. 3), which is different from studies on semi-supervised CCA (Blaschko, Lampert, and Gretton 2008), whose time complexity is at least  $O(N^2)$ .

## Experimental Setup

We conduct experiments on two publicly-available datasets in the multi-class classification setting. The first dataset is Corel-5K (Corel for short) (Duygulu et al. 2002), which is composed of 50 categories and each containing 100 images collected from the larger COREL CD set. Following the partition of the dataset, 4,500 images are used for training and the rest are used for test. Tags in the dataset are from a dictionary of 374 keywords, with each image having been annotated by an average of 3.5 tags.

The second dataset is NUS-WIDE-Object (NUS for short) (Chua et al. 2009), which is collected from the photo sharing website Flickr. In order to fit for the multi-class classification setting, we only select the images with a single class label, and thus obtain 23,953 images from all 31 categories. Note that the same strategy has already been adopted in (Gao, Chia, and Tsang 2011), where the authors have selected images with a single class label from 26 classes. We follow the standard train/test partition of the dataset, where the training set and the test set contain 14,270 and 9,683 images, respectively. Moreover, the most frequent 1,000 tags are retained on this dataset, with each image having been annotated by an average of 6.6 tags.

To generate the visual representation for the Corel dataset, we extract the SIFT descriptors (Lowe 2004) of  $16 \times 16$  pixel blocks computed over a regular grid with spacing of 8 pixels. We then perform k-means clustering on the extracted descriptors to form a vocabulary of 2,000 visual words. Based on the visual vocabulary, a 2,000-dimensional feature vector is obtained for each image. For the NUS dataset, we adopt the 500-dimensional bag-of-words representation available online for all the images. What is more, a binary matrix recording tag presence/absence is used as the textual representation for both datasets.

As the final step, we adopt linear SVM (Fan et al. 2008) to obtain the classification accuracy for evaluation.

## Empirical Results

To demonstrate the effectiveness of the proposed CVFL, we evaluate the performance of the following methods/features:

- **Visual Representation Only.**
- **Textual Representation Only.**
- **Textual Representation (Random Projection)**, which denotes the result derived from text representation with a random projection.
- **Visual + Textual Representation**, which denotes the combination of visual representation and textual representation.
- **PCA**, which denotes the principal component analysis (Hotelling 1933) approach. PCA is performed on the combined representation of image and tags.
- **CCA**, which denotes the canonical correlation analysis (Hotelling 1936). CCA is performed to maximize the correlation between visual representation and textual representation, and the learned representations are combined.

Table 1: Classification accuracy (%) of the proposed CVFL along with other methods/features on Corel dataset and NUS dataset.

Methods/Features	Corel	NUS
Visual Representation Only	46.2	31.0
Textual Representation Only	68.6	68.6
Textual Representation (Random Projection)	68.4	71.6
Visual + Textual Representation	70.6	69.5
Principal Component Analysis (PCA)	70.6	69.5
Canonical Correlation Analysis (CCA)	70.4	69.2
Partial Least Squares Regression (PLSR)	68.2	66.8
CVFL (proposed, without regularizer)	66.8	62.5
CVFL (proposed)	<b>71.8</b>	<b>74.7</b>

- **PLSR**, which denotes the partial least squares regression (Wold 1985) approach. PLSR is performed similarly with CCA, except that PLSR maximizes covariance while CCA maximizes correlation.
- **CVFL (without regularizer)**, which denotes the proposed cross-view feature learning approach without any regularizers (Eq. 1).
- **CVFL**, which denotes the proposed cross-view feature learning approach (Eq. 5). The learned representation  $\hat{X}$  is used to evaluate the final performance.

The classification results of all the aforementioned methods/features are listed in Table 1. Several observations can be made concerning the results. To begin with, CVFL performs better than a single representation, since the representation  $\hat{X}$  learned by CVFL encodes the information from both image and tags. Secondly, CVFL turns out to be more effective than a random mapping on textual representation, which shows that the mapping learned by CVFL is more descriptive than a randomly generated one, although CVFL is inspired by the idea of random projection. What is more, due to the explicit modeling of the relevance between image and tags, CVFL outperforms feature combination. Because the mapping learned by PCA is an orthogonal matrix, the results obtained by PCA and feature combination are the same. More importantly, being a reconstruction formulation, CVFL is compatible with many kinds of regularizers. With these regularizers, CVFL becomes more descriptive and obtains better results than CCA and PLSR. Finally, it can be learned from the underperformance of the unregularized CVFL that, regularizers are important for CVFL, and the result of overfitting can be catastrophic.

### Parameters and Complexity Issues

Parameters are determined by 5-fold cross-validation in the experiments. We empirically find that the performance of CVFL reaches its peak when both  $\lambda$  and  $\gamma$  are chosen between 1 and 2. Concretely, we choose  $\lambda = 1.5$  and  $\gamma = 1.8$  for the Corel dataset, and choose  $\lambda = 1.7$  and  $\gamma = 1.9$  for the NUS dataset. The detailed cross-validation results are illustrated in Fig. 2. It is noteworthy that CVFL is not sensitive to these parameters.

Besides parameter tuning, to systematically investigate the complexity issues, we report in Table 2 the running time

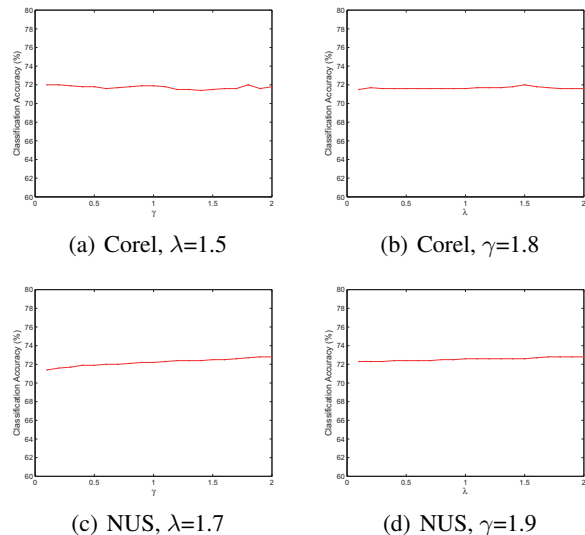


Figure 2: The 5-fold cross-validation results of  $\lambda$  and  $\gamma$  on: (a,b) Corel; (c,d) NUS.

Table 2: Running time (measured in seconds) of CCA and the proposed CVFL on Corel dataset and NUS dataset.

Methods	Corel	NUS
CCA	28.7	47.1
CVFL (proposed)	<b>17.1</b>	<b>11.0</b>

(measured in seconds) of the following two closely-related approaches: CCA and CVFL. Note that we run MATLAB codes on a server with 2.20GHz CPU and 128GB RAM. It can be observed from Table 2 that, CVFL performs more efficiently than CCA. The above results may be due to the fact that CCA involves a generalized eigenvalue problem which is computationally more expensive to solve, whereas it is more efficient to solve a least squares problem like CVFL. More notably, although the NUS dataset contains more images, CVFL remains to be efficient.

### Conclusion and Future Work

Due to numerous concepts in the real world and the scalability issues, a suitable approach to practical social image analysis is constrained to be unsupervised and computationally efficient. In this paper, we propose a cross-view feature learning (CVFL) framework to handle this task. As the promising results shown in the experiments, the proposed CVFL is an effective algorithm, although some simple settings are adopted in this paper (e.g., using a simple  $L_2$  regularizer instead of a sparsity or group sparsity regularizer, and using inner products to model the affinity of visual features). Taking into account that CVFL can be made more descriptive by using other types of regularizers, and that CVFL is only defined over feature representations, we will conduct a deeper analysis on the effectiveness of CVFL by investigating different regularizers, and apply CVFL to other large-scale multi-view tasks in unsupervised setting.

## Acknowledgments

This work was supported by National Hi-Tech Research and Development Program (863 Program) of China under Grants 2014AA015102 and 2012AA012503, National Natural Science Foundation of China under Grant 61371128, and Ph.D. Programs Foundation of Ministry of Education of China under Grant 20120001110097.

## References

- Achlioptas, D. 2001. Database-friendly random projections. In *PODS*, 274–281.
- Amaldi, E., and Kann, V. 1998. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science* 209(1):237–260.
- Atrey, P. K.; Hossain, M. A.; El Saddik, A.; and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16(6):345–379.
- Bach, F. R., and Jordan, M. I. 2003. Kernel independent component analysis. *JMLR* 3:1–48.
- Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS* 14:585–591.
- Bilenko, M.; Basu, S.; and Mooney, R. J. 2004. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 11.
- Bingham, E., and Mannila, H. 2001. Random projection in dimensionality reduction: applications to image and text data. In *ACM KDD*, 245–250.
- Blaschko, M. B.; Lampert, C. H.; and Gretton, A. 2008. Semi-supervised laplacian regularization of kernel canonical correlation analysis. In *ECML PKDD*. 133–145.
- Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *ACM KDD*, 333–342.
- Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, 48.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.
- Duygulu, P.; Barnard, K.; Freitas, J. F. G. d.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV* 97–112.
- Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; and Lin, C. 2008. Liblinear: A library for large linear classification. *JMLR* 9:1871–1874.
- Gao, S.; Chia, L.; and Tsang, I. 2011. Multi-layer group sparse coding—For concurrent image classification and annotation. In *CVPR*, 2809–2816.
- Gionis, A.; Indyk, P.; and Motwani, R. 1999. Similarity search in high dimensions via hashing. In *VLDB*, 518–529.
- Guyon, I., and Elisseeff, A. 2003. An introduction to variable and feature selection. *JMLR* 3:1157–1182.
- Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24(6):417–441.
- Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.
- Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *ICCV*, 2407–2414.
- Johnson, W., and Lindenstrauss, J. 1984. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26, 189–206.
- Lanckriet, G.; Cristianini, N.; Bartlett, P.; Ghaoui, L.; and Jordan, M. 2004. Learning the kernel matrix with semidefinite programming. *JMLR* 5:27–72.
- Li, Z.; Liu, J.; Jiang, Y.; Tang, J.; and Lu, H. 2012. Low rank metric learning for social image retrieval. In *ACM MM*, 853–856.
- Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60(2):91–110.
- Lu, Z., and Peng, Y. 2013. Latent semantic learning with structured sparse representation for human action recognition. *PR* 46(7):1799–1809.
- Putthivithy, D.; Attias, H.; and Nagarajan, S. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 3408–3415.
- Smeulders, A. W.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *TPAMI* 22(12):1349–1380.
- Sriperumbudur, B. K.; Torres, D. A.; and Lanckriet, G. R. 2007. Sparse eigen methods by dc programming. In *ICML*, 831–838.
- Sun, L.; Ji, S.; and Ye, J. 2011. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *TPAMI* 33(1):194–200.
- Torres, R. d. S.; Falcão, A. X.; Gonçalves, M. A.; Papa, J. P.; Zhang, B.; Fan, W.; and Fox, E. A. 2009. A genetic programming framework for content-based image retrieval. *PR* 42(2):283–292.
- Vedaldi, A.; Gulshan, V.; Varma, M.; and Zisserman, A. 2009. Multiple kernels for object detection. In *ICCV*, 606–613.
- Vempala, S. S. 2004. *The random projection method*. AMS Bookstore.
- Wang, S.; Jiang, S.; Huang, Q.; and Tian, Q. 2010. S3mkl: scalable semi-supervised multiple kernel learning for image data mining. In *ACM MM*, 163–172.
- Watkins, D. S. 2004. *Fundamentals of matrix computations*. John Wiley & Sons.
- Wold, H. 1985. Partial least squares. *Encyclopedia of statistical sciences*.
- Wu, P.; Hoi, S. C.; Xia, H.; Zhao, P.; Wang, D.; and Miao, C. 2013. Online multimodal deep similarity learning with application to image retrieval. In *ACM MM*, 153–162.
- Xia, H.; Wu, P.; and Hoi, S. C. 2013. Online multi-modal distance learning for scalable multimedia retrieval. In *ACM WSDM*, 455–464.
- Xing, E. P., and Karp, R. M. 2001. Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* 17(suppl 1):S306–S315.
- Xu, Z.; King, I.; Lyu, M.-T.; and Jin, R. 2010. Discriminative semi-supervised feature selection via manifold regularization. *TNN* 21(7):1033–1047.
- Yan, S., and Wang, H. 2009. Semi-supervised learning by sparse representation. In *SDM*, 792–801.
- Yang, L., and Jin, R. 2006. Distance metric learning: A comprehensive survey. *Michigan State University* 1–51.