Direct Semantic Analysis for Social Image Classification

Zhiwu Lu[1,2] and **Liwei Wang**[3] and **Ji-Rong Wen**[1]
[1]School of Information, Renmin University of China, Beijing 100872, China
[2]Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China
[3]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, China
{zhiwu.lu, wangliwei.pku, jirong.wen}@gmail.com

Abstract

This paper presents a direct semantic analysis method for learning the correlation matrix between visual and textual words from socially tagged images. In the literature, to improve the traditional visual bag-of-words (BOW) representation, latent semantic analysis has been studied extensively for learning a compact visual representation, where each visual word may be related to multiple latent topics. However, these latent topics do not convey any true semantic information which can be understood by human. In fact, it remains a challenging problem how to recover the relationships between visual and textual words. Motivated by the recent advances in dealing with socially tagged images, we develop a direct semantic analysis method which can explicitly learn the correlation matrix between visual and textual words for social image classification. To this end, we formulate our direct semantic analysis from a graph-based learning viewpoint. Once the correlation matrix is learnt, we can readily first obtain a semantically refined visual BOW representation and then apply it to social image classification. Experimental results on two benchmark image datasets show the promising performance of the proposed method.

Introduction

In image analysis and computer vision, the visual bag-of-words (BOW) representation has been widely applied to different challenging tasks such as image classification and annotation. Especially for image classification, many encouraging results (Lazebnik, Schmid, and Ponce 2006; Moosmann, Nowak, and Jurie 2008; Li et al. 2008; Guillaumin, Verbeek, and Schmid 2010; Stottinger et al. 2012) have been reported in the literature. However, as shown in previous work (Mallapragada, Jin, and Jain 2010; Ji et al. 2009; Liu, Yang, and Shah 2009; Lu and Peng 2011), the traditional visual BOW representation still suffers from the so-called semantic gap. That is, for efficiency purposes, the visual words (which play an important role in visual BOW representation) are commonly generated by directly quantizing the local visual descriptors extracted from images, without considering the high-level semantics of images.

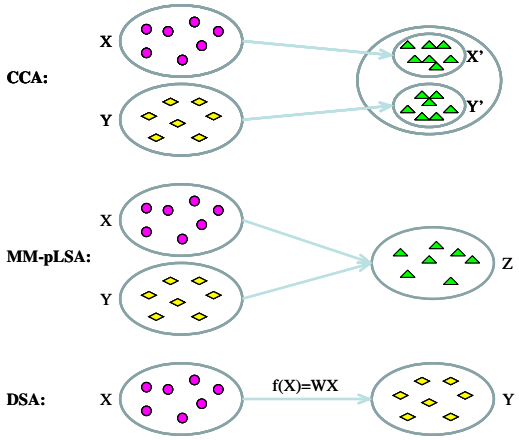

Figure 1: Illustration of the difference between our direct semantic analysis (DSA) and two representative methods for multi-modal data analysis.

To improve the traditional visual BOW representation, many approaches to latent semantic analysis (Quelhas et al. 2005; Bosch, Zisserman, and Muñoz 2006; Fei-Fei and Perona 2005; Cao and Fei-Fei 2007) have been developed based on topic models such as probabilistic latent semantic analysis (pLSA) (Hofmann 2001) and latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). A mixture of latent topics is used to model each image, and the latent topics are learnt by these approaches as multinomial distributions of visual words. Hence, each visual word may be related to multiple latent topics. However, the learnt latent topics *do not convey any true semantic information* which can be understood by human. This is also the reason why these approaches are called as "latent semantic analysis". In fact, it remains a challenging problem how to relate visual words to the high-level semantics in the literature.

Fortunately, with the burgeoning growth of shared images over online social networks, the aforementioned problem become less difficult to handle. That is, the tags of shared images provided by users can be regarded as the high-level semantics (or textual words) to be related to visual words. In this paper, motivated by the recent advances in dealing with these socially tagged images, we

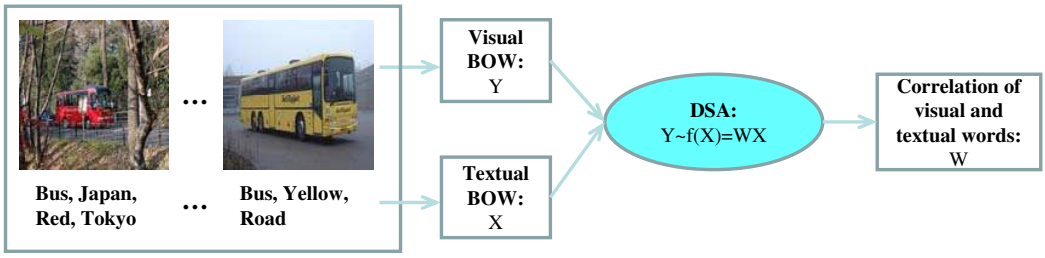

Figure 2: The flowchart of our direct semantic analysis (DSA) for learning the correlation matrix between visual and textual words from socially tagged images.

develop a direct semantic analysis (DSA) method which can explicitly learn the correlation matrix between visual and textual words for social image classification. The basic idea is to formulate the problem of learning the correlation matrix between visual and textual words from a graph-based learning viewpoint. We define Laplacian regularization (Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004; Fu et al. 2011) over the textual words (i.e. social tags) of images and add this regularization term into the objective function of graph-based learning. Due to the special definition of Laplacian regularization, our new DSA problem can be solved efficiently based on the label propagation technique proposed in (Zhou et al. 2004). Besides learning the correlation matrix between visual and textual words, we also pay much attention to dealing with the noise issue caused by the inaccurate quantization and the nosily socially tagging in social image classification. Although image annotation based on relevance models (Jeon, Lavrenko, and Manmatha 2003; Feng, Manmatha, and Lavrenko 2004) can similarly learn the relationships between image regions and textual words, the tags of images are required to be exactly correct and this approach is not suitable for nosily socially tagged images.

When the visual BOW representation and the social tag information are considered as two modalities of images, our direct semantic analysis actually belongs to multi-modal data analysis. However, our present work is distinctly different from previous work on multi-modal data analysis. In (Rasiwasia et al. 2010), the two modalities of images are mapped into two latent spaces of the same dimension by canonical correlation analysis (CCA) (Hotelling 1936) so that the correlation matrix can be estimated by the product operation across the two latent spaces. In (Chandrika and Jawahar 2010), the two modalities of images are mapped into the same latent space by multi-modal pLSA (MM-pLSA). In contrast, our direct semantic analysis can map one modality into another modality *without using any latent space*, which makes it convenient not only to learn the correlation matrix but also to deal with the noise issue. The difference between our direct semantic analysis and these two representative methods is also illustrated in Figure 1.

In summary, we propose a direct semantic analysis (DSA) method for learning the correlation matrix between visual and textual words from socially tagged images, as illustrated in Figure 2. Once the correlation matrix between visual and textual words is learnt, we can readily first obtain a semantically refined visual BOW representation and then apply it to social image classification. Here, it is worth noting that our DSA method is efficient even for large image datasets. More notably, when the global visual features are fused for social image classification, our DSA method can *achieve very impressive results* on the PASCAL VOC'07 (Everingham et al. 2007) and MIR FLICKR (Huiskes and Lew 2008) benchmark datasets, as shown in our later experiments. Although only evaluated in social image classification, our DSA method can be readily extended to other challenging tasks such as semantic image segmentation.

The remainder of this paper is organized as follows. In Section 2, we develop a novel direct semantic analysis (DSA) method for learning the correlation matrix between visual and textual words from a graph-based learning viewpoint. In Section 3, the semantically refined visual BOW representation is evaluated on two benchmark datasets by applying it to social image classification. Finally, Section 4 gives the conclusions drawn from our experimental results.

Direct Semantic Analysis

This section presents direct semantic analysis (DSA) in detail. We first give our problem formulation for learning the correlation matrix between visual and textual words from a graph-based learning viewpoint, and then develop an efficient DSA algorithm based on the label propagation technique (Zhou et al. 2004). Finally, we discuss the out-of-sample extension and illustrative explanation of our DSA.

Problem Formulation

In this paper, our goal is to learn the correlation matrix between visual and textual words from social tagged images. To this end, we need to first generate the visual and textual BOW representation for each social tagged image. More concretely, the visual BOW representation is formed by quantizing local visual descriptors extracted from images, while the textual BOW representation is derived from the social tags of images. Due to the inaccurate quantization and the nosily socially tagging, both visual and textual BOW representation suffer from the noise issue. Hence, the main challenge in learning the correlation matrix between visual and textual words is actually how to deal with the noise issue. In the following, our problem formulation is elaborated from a graph-based learning viewpoint.

Let $Y \in R^{M\times N}$ denote the visual BOW representation and $X \in R^{K\times N}$ denote the textual BOW representation, where N is the number of images, M is the number of visual words, and K is the number of textual words. We compute the kernel matrix $A \in R^{N\times N}$ over the textual BOW representation X. In this paper, we only consider linear kernel for the textual BOW representation. By directly using A as the affinity matrix, we construct an undirected graph $\mathcal{G} = \{\mathcal{V}, A\}$ with its vertex set $\mathcal{V}$ being the set of images. The normalized Laplacian matrix of $\mathcal{G}$ is given by

$$L = I - D^{-1/2}AD^{-1/2}, \tag{1}$$

where I is an identity matrix and D is a diagonal matrix with its i-th diagonal entry being the sum of the i-th row of A.

Based on the above notations, the problem of learning the correlation matrix between visual and textual words can be formulated from a graph-based learning viewpoint:

$$\min_{W,\hat{X},\hat{Y}} \frac{1}{2}||\hat{Y} - W\hat{X}||_F^2 + \frac{\lambda}{2}\mathrm{tr}(W\hat{X}L\hat{X}^TW^T) + \gamma||\hat{Y} - Y||_1, \tag{2}$$

where $W \in R^{M\times K}$ denotes the correlation matrix between visual and textual words, $\hat{X} \in R^{K\times N}$ denotes the ideal textual BOW representation, $\hat{Y} \in R^{M\times N}$ denotes the ideal visual BOW representation, λ and γ denote the positive regularization parameters, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. It should be noted that although our goal is to find the correlation matrix W between visual and textual words, $\hat{X}$ and $\hat{Y}$ are also optimized since both original X and Y suffer from the noise issue. That is, we expect to find optimal W by simultaneously dealing with the noise issue associated with X and Y. Moreover, considering that the correlation matrix W can be found directly, solving Eq. (2) is called as *direct semantic analysis* (DSA) in this paper.

The objective function given by Eq. (2) is further discussed as follows. The first term denotes the Frobenius-norm fitting constraint, which means that $W\hat{X}$ should not change too much from $\hat{Y}$. The second term denotes the smoothness constraint, also known as Laplacian regularization (Zhu, Ghahramani, and Lafferty 2003; Zhou et al. 2004; Fu et al. 2011), which means that $W\hat{X}$ should not change too much between similar images. The third term denotes the L_1-norm fitting constraint, which can impose direct noise reduction on the original Y due to the nice property of L_1-norm optimization (Elad and Aharon 2006; Mairal, Elad, and Sapiro 2008; Wright et al. 2009). Here, besides the ideal textual BOW representation $\hat{X}$, we also introduce the ideal visual BOW representation $\hat{Y}$ into our problem formulation. Our main motivation is to impose direct noise reduction on Y by extra consideration of the L_1-norm fitting constraint $||\hat{Y} - Y||_1$. Although this L_1-norm fitting constraint is only defined with respect to $\hat{Y}$, the effect of noise reduction can be transferred to $W\hat{X}$ by solving Eq. (2) with $\hat{Y}$ being an intermediate representation.

To apply our DSA to large image datasets, we have to concern the following key problem: *how to solve Eq. (2) efficiently*. Fortunately, due to the special definition of Laplacian regularization in Eq. (2), the problem of learning the correlation matrix W can be solved efficiently using the label propagation technique (Zhou et al. 2004) based on k-nearest neighbors (k-NN) graph constructed with the textual BOW representation. The proposed efficient DSA algorithm will be elaborated in the next subsection.

Efficient DSA Algorithm

In fact, the DSA problem (2) can be solved in two alternate optimization steps as follows:

$$W^*, \hat{X}^* = \arg\min_{W,\hat{X}} \frac{1}{2}||\hat{Y}^* - W\hat{X}||_F^2 + \frac{\lambda}{2}\mathrm{tr}(W\hat{X}L\hat{X}^TW^T),$$

$$\hat{Y}^* = \arg\min_{\hat{Y}} \frac{1}{2}||\hat{Y} - W^*\hat{X}^*||_F^2 + \gamma||\hat{Y} - Y||_1.$$

Here, we set $\hat{X}^* = X$ and $\hat{Y}^* = Y$ initially. Concretely, as a basic L_1-norm optimization problem, the second subproblem has an explicit solution based on the following soft-thresholding function:

$$\hat{Y}^* = \mathrm{soft}(W^*\hat{X}^* - Y, \gamma) + Y, \tag{3}$$

where $\mathrm{soft}(y, \gamma) = \mathrm{sign}(y)\max\{|y| - \gamma, 0\}$. In the following, we focus on developing an efficient algorithm to solve the first quadratic optimization subproblem.

Let $\mathcal{Q}(W, \hat{X}) = \frac{1}{2}||\hat{Y}^* - W\hat{X}||_F^2 + \frac{\lambda}{2}\mathrm{tr}(W\hat{X}L\hat{X}^TW^T)$. We can still adopt the alternate optimization technique for the first subproblem $\min_{W,\hat{X}} \mathcal{Q}(W, \hat{X})$: 1) fix $\hat{X} = \hat{X}^*$, and update W by $W^* = \arg\min_W \mathcal{Q}(W, \hat{X}^*)$; 2) fix $W = W^*$, and update $\hat{X}$ by $\hat{X}^* = \arg\min_{\hat{X}} \mathcal{Q}(W^*, \hat{X})$.

Updating W: When $\hat{X}$ is fixed at $\hat{X}^*$, the solution of $\min_W \mathcal{Q}(W, \hat{X}^*)$ can be found by solving

$$\frac{\partial \mathcal{Q}(W, \hat{X}^*)}{\partial W} = (W\hat{X}^* - \hat{Y}^*)(\hat{X}^*)^T + \lambda W\hat{X}^*L(\hat{X}^*)^T = 0,$$

which can be further transformed into

$$W(\hat{X}^*(I + \lambda L)(\hat{X}^*)^T) = \hat{Y}^*(\hat{X}^*)^T. \tag{4}$$

Since $\hat{X}^*(I + \lambda L)(\hat{X}^*)^T \in R^{K\times K}$ and $K \ll \min(N, M)$, the above linear equation can be solved very efficiently.

Updating $\hat{X}$: When W is fixed at W^*, the solution of $\min_{\hat{X}} \mathcal{Q}(W^*, \hat{X})$ can be found by solving

$$\frac{\partial \mathcal{Q}(W^*, \hat{X})}{\partial \hat{X}} = W^{*T}(W^*\hat{X} - \hat{Y}^*) + \lambda W^{*T}W^*\hat{X}L = 0,$$

which is actually equivalent to

$$W^{*T}W^*\hat{X}(I + \lambda L) = W^{*T}\hat{Y}^*. \tag{5}$$

Let $F(\hat{X}) = W^{*T}W^*\hat{X}$. Since $I + \lambda L$ is a positive definite matrix, the above linear equation has an analytical solution:

$$F^*(\hat{X}) = W^{*T}\hat{Y}^*(I + \lambda L)^{-1}. \tag{6}$$

However, this analytical solution is not efficient for large image datasets, since matrix inverse has a time complexity of $O(N^3)$. Fortunately, this solution can also be *efficiently found using the label propagation technique* proposed in

(Zhou et al. 2004) based on k-NN graph. Finally, the solution of $\min_{\hat{X}} \mathcal{Q}(W^*, \hat{X})$ is found by solving:

$$(W^{*T}W^*)\hat{X} = F^*(\hat{X}). \quad (7)$$

Since $W^{*T}W^* \in R^{K\times K}$ and $K \ll \min(N, M)$, the above linear equation can be solved very efficiently.

The complete DSA algorithm is outlined as follows:

(1) Construct a k-NN graph with its affinity matrix A being defined over the textual BOW representation X;

(2) Compute the normalized Laplacian matrix $L = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ according to Eq. (1);

(3) Initialize the deal textual and visual BOW representation as $\hat{X}^* = X$ and $\hat{Y}^* = Y$, respectively;

(4) Find the best solution W^* by solving $W(\hat{X}^*(I + \frac{\alpha}{1-\alpha}L)(\hat{X}^*)^T) = \hat{Y}^*(\hat{X}^*)^T$, which is exactly Eq. (4) with $\alpha = \lambda/(1+\lambda) \in (0, 1)$;

(5) Iterate $F_{t+1}(\hat{X}) = \alpha F_t(\hat{X})(I - L) + (1-\alpha)W^{*T}\hat{Y}^*$ until convergence, where a solution can thus be found just the same as Eq. (6) with $\alpha = \lambda/(1+\lambda)$ (see more explanation below);

(6) Find the best solution $\hat{X}^*$ by solving Eq. (7): $(W^{*T}W^*)\hat{X} = F^*(\hat{X})$, where $F^*(\hat{X})$ denotes the limit of the sequence $\{F_t(\hat{X})\}$;

(7) Iterate Steps (4)–(6) until the stopping condition is satisfied, and update the deal visual BOW representation as: $\hat{Y}^* = \text{soft}(W^*\hat{X}^* - Y, \gamma) + Y$;

(8) Iterate Steps (4)–(7) until the stopping condition is satisfied, and output the final semantically refined visual BOW representation $\hat{Y}^*$.

Similar to the convergence analysis in (Zhou et al. 2004), the iteration in Step (5) converges to $F^*(\hat{X}) = W^{*T}\hat{Y}^*(1-\alpha)(I - \alpha(I - L))^{-1}$, which is equal to the solution given by Eq. (6) with $\alpha = \lambda/(1+\lambda)$. Moreover, in our later experiments, we find that the iterations in Steps (5) , (7), and (8) generally converge in very limited number of iteration steps (<10). Finally, since the time complexity of Steps (4-7) is respectively $O(K^2M + KMN + K^2N + kKN)$, $O(KMN + kKN)$, $O(K^2M + K^2N)$, and $O(KMN)$ ($k, K \ll \min(N, M)$), the proposed DSA algorithm can be applied to large image datasets.

Out-of-Sample Extension

In this subsection, we discuss the out-of-sample extension issue. In fact, since we have found the best correlated matrix $W^* \in R^{M\times K}$, our DSA algorithm can readily deal with this issue when a new image is coming. Let $y \in R^{M\times 1}$ be the visual BOW representation of this new image. The problem of learning the semantically refined visual BOW representation for this new image can be formulated as follows:

$$\min_{\hat{x},\hat{y}} \frac{1}{2}||\hat{y} - W^*\hat{x}||^2 + \gamma||\hat{y} - y||_1, \quad (8)$$

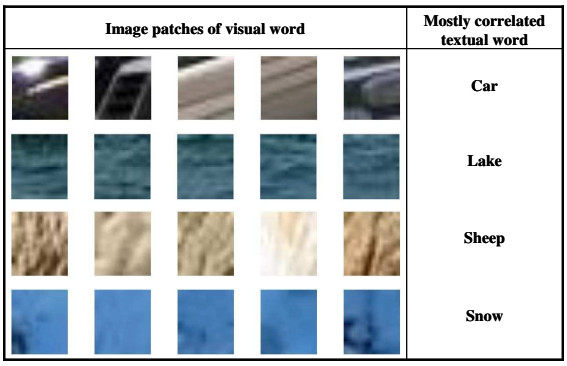

Figure 3: Illustration of the mostly correlated textual words found by our DSA for several examples of visual words, where each visual word is denoted by image patches.

where $\hat{x} \in R^{K\times 1}$ and $\hat{y} \in R^{M\times 1}$ denote the ideal textual and visual BOW representation of this new image. The above problem can be solved by alternate optimization:

$$\hat{x}^* = \arg\min_{\hat{x}} \frac{1}{2}||\hat{y}^* - W^*\hat{x}|^2,$$

$$\hat{y}^* = \arg\min_{\hat{y}} \frac{1}{2}||\hat{y} - W^*\hat{x}^*||^2 + \gamma||\hat{y} - y||_1.$$

Here, we set $\hat{y}^* = y$ initially. The first subproblem can be solved by the standard quadratic optimization technique, while the second subproblem has an explicit solution:

$$\hat{y}^* = \text{soft}(W^*\hat{x}^* - y, \gamma) + y, \quad (9)$$

where $\text{soft}(y, \gamma) = \text{sign}(y)\max\{|y| - \gamma, 0\}$. Since both of the above two subproblems are solved at a linear time cost with respect to M, we can learn the semantically refined visual representation for the new image very efficiently.

Discussion

As we have mentioned in the introduction, the traditional latent semantic analysis (Quelhas et al. 2005; Bosch, Zisserman, and Muñoz 2006; Fei-Fei and Perona 2005; Cao and Fei-Fei 2007) can only relate visual words to latent topics which convey no true semantic information, while our direct semantic analysis (DSA) can learn the correlation matrix between visual and textual words explicitly. To make this clearer, we show the mostly correlated textual words found by our DSA for several examples of visual words in Figure 3. Here, the experiment is conducted on a subset of the PASCAL VOC'07 dataset (Everingham et al. 2007), and each example of visual word is denoted by a set of image patches. We can observe that the mostly correlated textual word found by our DSA is *consistent with what we should understand each visual word as* it really is. This means that the correlation matrix between visual and textual words has been effectively learnt by our DSA from socially tagged images. To end this section, we want to emphasize that the experimental results illustrated in Figure 3 actually pave the way to apply our DSA to the challenging task of semantic image segmentation, since each image patch has been explicitly attached with a textual word.

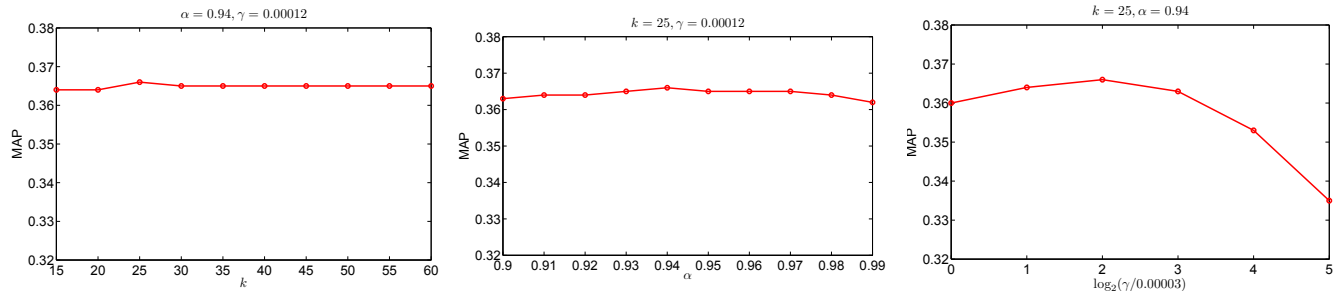

Figure 4: The cross-validation classification results using the semantically refined visual BOW representations learnt by our DSA algorithm on the training set of the PASCAL VOC'07 dataset.

Experimental Results

In this section, the proposed DSA algorithm is evaluated in social image classification on two benchmark datasets. We first describe the experimental setup, including information of the two benchmark datasets and the implementation details. Moreover, our DSA algorithm is compared with other closely related methods on the two benchmark datasets.

Experimental Setup

We select two benchmark datasets for performance evaluation. The first dataset is PASCAL VOC'07 (Everingham et al. 2007) that contains around 10,000 images. Each image is annotated by users with a set of tags, and the total number of tags used here is reduced to 804 by the same preprocessing step as (Guillaumin, Verbeek, and Schmid 2010). This dataset is organized into 20 classes. Moreover, the second dataset is MIR FLICKR (Huiskes and Lew 2008) that contains 25,000 images annotated with 457 tags. This dataset is organized into 38 classes. For the PASCAL VOC'07 dataset, we use the standard training/test split, while for the MIR FLICKR dataset we split it into 12,500 training/test images just as (Guillaumin, Verbeek, and Schmid 2010).

For each dataset, we extract the same feature set as (Guillaumin, Verbeek, and Schmid 2010). That is, we use local SIFT features and local hue histograms, both computed on a dense regular grid and on regions found with a Harris interest-point detector. We quantize the four types of local descriptors using k-means clustering, and represent each image using four visual word histograms. Moreover, following the idea of (Lazebnik, Schmid, and Ponce 2006), each visual BOW representation is also computed over a 3×1 horizontal decomposition of the image, and concatenated to form a new representation that encodes some of the spatial layout of the image. Finally, by concatenating all the visual BOW representations into a single representation, we generate a large visual vocabulary of about 10,000 visual words exactly the same as (Guillaumin, Verbeek, and Schmid 2010).

To evaluate the semantically refined visual BOW representation learnt by our DSA algorithm, we apply it directly to social image classification using SVM with χ^2 kernel. Since we actually perform multi-label classification on the two benchmark datasets, the classification results are measured by mean average precision (MAP) just the same as (Guillaumin, Verbeek, and Schmid 2010). In the following, we compare our DSA algorithm with three closely related methods: CCA (Rasiwasia et al. 2010), MM-pLSA (Chandrika and Jawahar 2010), and standard pLSA over the concatenation of visual and textual BOW representation. Although there exist other multi-modal pLSA methods such as (Lienhart, Romberg, and Hörster 2009), we *only make comparison to MM-pLSA* which has been shown to have superior performance in (Chandrika and Jawahar 2010).

In the experiments, the parameters of our DSA algorithm are selected by cross-validation on the training set. For example, according to Figure 4, we set the three parameters of our DSA algorithm on the PASCAL VOC'07 dataset as: $k = 25$, $\alpha = 0.94$ and $\gamma = 0.00012$ (which appear in Steps 1, 4 (or 5), 7 of our DSA algorithm proposed in Section 2, respectively). The same parameter selection strategy is adopted by other closely related methods.

Classification Results

We first show the comparison between different BOW representations on the two benchmark datasets in Figure 5(a). The immediate observation is that the semantically refined visual BOW representation learnt by our DSA algorithm significantly outperforms the original visual BOW representation. That is, the social tags of images have been effectively added to the refined visual BOW representation and thus the semantic gap associated with the original visual BOW representation has been reduced effectively. More notably, our semantically refined visual BOW representation is even shown to achieve more than 39% gains over the original textual BOW representation on both of the two benchmark datasets. The significant gains over the original visual and textual BOW representation are due to the fact that our DSA algorithm can deal with the noise issue associated with these two types of BOW representations during learning the semantically refined visual BOW representation.

The comparison between different methods for learning semantically refined visual BOW representation is further shown in Figure 5(b). Here, our DSA method do not use any latent space for learning semantically refined visual BOW representation, while the other three methods all consider one or more latent spaces. From Figure 5(b), we find that our DSA method obviously outperforms the other three methods in the challenging task of social image classification. These impressive results mean that directly learning the mapping

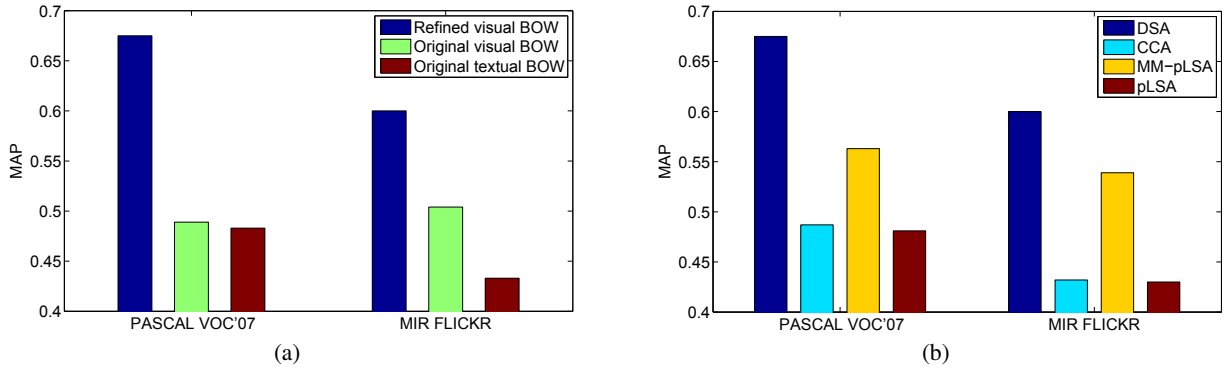

Figure 5: The test classification results using semantically refined visual BOW representations on the two benchmark datasets: (a) comparison between different BOW representations; (b) comparison between different methods.

Table 1: Comparison of our DSA method with the state-of-the-art on the two benchmark datasets (LVF: local visual features; GVF: global visual features).

Methods	LVF	GVF	Tags	VOC	MIR
Winner	yes	yes	no	0.594	–
(Guillaumin et al. 2010)	yes	yes	yes	0.667	0.623
Ours (LVF only)	yes	no	yes	0.675	0.600
Ours (LVF+GVF)	yes	yes	yes	**0.701**	**0.646**

between visual and textual BOW representation by our DSA method is more suitable for social image classification than those methods that make use of latent spaces. More importantly, without using any latent space, our DSA method can readily deal with the noise issue associated with the original visual and textual BOW representation, which is especially crucial for social image classification.

The comparison of our DSA method with the state-of-the-art on the two benchmark datasets is shown in Table 1. To the best of our knowledge, the recent work (Guillaumin, Verbeek, and Schmid 2010) has reported the best results so far for social image classification on the PASCAL VOC'07 and MIR FLICKR datasets. However, when the semantically refined visual BOW representation (i.e. local visual features) obtained by our method is fused with the global visual features (i.e. color histogram and GIST descriptor (Oliva and Torralba 2001)), our method is shown to achieve better results than (Guillaumin, Verbeek, and Schmid 2010) on both benchmark datasets. This becomes more impressive given that the present work makes use of *much weaker* global visual features than (Guillaumin, Verbeek, and Schmid 2010) (i.e. two types vs. seven types). Moreover, since (Guillaumin, Verbeek, and Schmid 2010) makes a direct fusion of visual features and tag information, the gain achieved by our method also means that our method *outperforms direct fusion*. Finally, from Table 2, we observe that both (Guillaumin, Verbeek, and Schmid 2010) and our method obviously outperform the winner of PASCAL VOC'07 due to the effective use of extra tags for social image classification.

Table 2: The running time (minutes) of learning semantically refined visual BOW representation taken by different methods on the MIR FLICKR dataset.

Methods	DSA	CCA	MM-pLSA	pLSA
Running time	5	1	83	62

Besides the above advantages, our DSA method has another advantage, i.e., it runs very fast even on large datasets. For example, the running time of learning semantically refined visual BOW representation taken by different methods on MIR FLICKR ($N = 25,000$) is listed in Table 2. We run the algorithms (Matlab code) on a computer with 3GHz CPU and 32GB RAM. It can be observed that our DSA *runs much faster* than MM-pLSA and pLSA, while CCA runs the fastest without considering noise reduction.

Conclusions

In this paper, we have proposed novel direct semantic analysis for learning the correlation matrix between visual and textual words from socially tagged images. To deal with the noise issue associated with the original visual and textual BOW representation, we have developed an efficient graph-based learning algorithm for our direct semantic analysis. The effectiveness of the proposed method has been verified by the extensive experimental results on two benchmark datasets. More importantly, since our experimental results have actually paved the way to apply the proposed method to semantic image segmentation, we will pay much attention to this challenging task in the future work.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 61202231, National Basic Research Program of China (973 Program) under Grant 2012CB316205, and Beijing Natural Science Foundation of China under Grant 4132037.

References

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bosch, A.; Zisserman, A.; and Muñoz, X. 2006. Scene classification via pLSA. In *Proc. European Conference on Computer Vision (ECCV)*, 517–530.

Cao, L., and Fei-Fei, L. 2007. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

Chandrika, P., and Jawahar, C. V. 2010. Multi modal semantic indexing for image retrieval. In *Proc. ACM International Conference on Image and Video Retrieval*, 342–349.

Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Processing* 15(12):3736–3745.

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; and Zisserman, A. 2007. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007.

Fei-Fei, L., and Perona, P. 2005. A Bayesian hierarchical model for learning natural scene. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 524–531.

Feng, S.; Manmatha, R.; and Lavrenko, V. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 1002–1009.

Fu, Z.; Lu, Z.; Ip, H. H.; Peng, Y.; and Lu, H. 2011. Symmetric graph regularized constraint propagation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 350–355.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multimodal semi-supervised learning for image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 902–909.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 41:177–196.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3-4):321–377.

Huiskes, M., and Lew, M. 2008. The MIR Flickr retrieval evaluation. In *Proc. ACM International Conference on Multimedia Information Retrieval (MIR)*, 39–43.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 119–126.

Ji, R.; Xie, X.; Yao, H.; and Ma, W.-Y. 2009. Vocabulary hierarchy optimization for effective and transferable retrieval. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1161–1168.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2169–2178.

Li, J.; Wu, W.; Wang, T.; and Zhang, Y. 2008. One step beyond histograms: Image representation using Markov stationary features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lienhart, R.; Romberg, S.; and Hörster, E. 2009. Multilayer pLSA for multimodal image retrieval. In *Proc. ACM International Conference on Image and Video Retrieval*.

Liu, J.; Yang, Y.; and Shah, M. 2009. Learning semantic visual vocabularies using diffusion distance. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 461–468.

Lu, Z., and Peng, Y. 2011. Latent semantic learning by efficient sparse coding with hypergraph regularization. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 411–416.

Mairal, J.; Elad, M.; and Sapiro, G. 2008. Sparse representation for color image restoration. *IEEE Trans. Image Processing* 17(1):53–69.

Mallapragada, P.; Jin, R.; and Jain, A. 2010. Online visual vocabulary pruning using pairwise constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3073–3080.

Moosmann, F.; Nowak, E.; and Jurie, F. 2008. Randomized clustering forests for image classification. *IEEE Trans. Pattern Analysis and Machine Intelligence* 30(9):1632–1646.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42(3):145–175.

Quelhas, P.; Monay, F.; Odobez, J.-M.; Gatica-Perez, D.; Tuytelaars, T.; and Gool, L. V. 2005. Modeling scenes with local descriptors and latent aspects. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 883–890.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proc. ACM Internatial Conference on Multimedia*, 251–260.

Stottinger, J.; Hanbury, A.; Sebe, N.; and Gevers, T. 2012. Sparse color interest points for image retrieval and object categorization. *IEEE Trans. Image Processing* 21(5):2681–2692.

Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence* 31(2):210–227.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 321–328.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. International Conference on Machine Learning (ICML)*, 912–919.