Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Learning Descriptive Visual Representation by Semantic Regularized Matrix Factorization

Zhiwu Lu and Yuxin Peng*
Institute of Computer Science and Technology, Peking University,
Beijing 100871, China
{luzhiwu,pengyuxin}@pku.edu.cn

Abstract

This paper presents a novel semantic regularized matrix factorization method for learning descriptive visual bag-of-words (BOW) representation. Although very influential in image classification, the traditional visual BOW representation has one distinct drawback. That is, for efficiency purposes, this visual representation is often generated by directly clustering the low-level visual feature vectors extracted from local keypoints or regions, without considering the high-level semantics of images. In other words, this visual representation still suffers from the semantic gap and may lead to significant performance degradation in more challenging tasks (e.g., classification of community-contributed images with large intra-class variations). To overcome this drawback, we develop a semantic regularized matrix factorization method for learning descriptive visual BOW representation by adding Laplacian regularization defined with the tags (easy to access although noisy) of community-contributed images into matrix factorization. Experimental results on two benchmark datasets show the promising performance of the proposed method.

1 Introduction

Inspired by the success of bag-of-words (BOW) in text information retrieval, we can similarly represent an image as a histogram of visual words through quantizing the local keypoints or regions within the image into visual words, which is known as visual BOW in the areas of image analysis and computer vision. As an intermediate representation, it can help to reduce the semantic gap between the low-level visual features and the high-level semantics of images to some extent. Hence, in the literature, many efforts have been made to apply the visual BOW representation to image classification (one typical task in image analysis and computer vision). In fact, the visual BOW representation has been shown to give rise to encouraging results in image classification [Lazebnik *et al.*, 2006; Moosmann *et al.*, 2008; Li *et al.*, 2008; Guillaumin *et al.*, 2010; Stottinger *et al.*, 2012].

*Corresponding author.

However, as reported in previous work [Mallapragada *et al.*, 2010; Ji *et al.*, 2009; Liu *et al.*, 2009; Lu and Peng, 2011], the traditional visual BOW representation has one distinct drawback as follows. That is, for efficiency purposes, the visual vocabulary is commonly constructed for visual BOW generation by directly clustering the low-level visual feature vectors extracted from local keypoints or regions within images, without considering the high-level semantics of images. In other words, the traditional visual BOW representation still suffers from the problem of semantic gap and thus may lead to significant performance degradation in more challenging tasks such as classification of community-contributed images. Here, it is worth noting that the community-contributed images are shared in an unconstrained way and thus are more difficult to classify with larger intra-class variations (see examples in Figure 3). In this paper, our main motivation is to propose a new method for learning descriptive visual BOW representation to overcome the aforementioned drawback associated with the traditional visual BOW representation.

Considering that matrix factorization has been successfully applied to image representation [Cai *et al.*, 2011; Liu *et al.*, 2012], we develop a semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags (easy to access although noisy) of community-contributed images. The basic idea is to formulate the problem of learning descriptive visual BOW representation as low-rank matrix factorization (see Figure 2). We further define Laplacian regularization [Zhu *et al.*, 2003; Zhou *et al.*, 2004; Fergus *et al.*, 2010] with the tags of images (unlike [Liu *et al.*, 2012] that utilizes the class labels of images) and add this term into the objective function of matrix factorization. Due to the special definition of Laplacian regularization, our new SRMF problem can be solved efficiently based on the label propagation technique proposed in [Zhou *et al.*, 2004]. Although a Laplacian regularized matrix factorization method has also been proposed in [Cai *et al.*, 2011], our SRMF method has two distinct differences as follows: 1) we define Laplacian regularization in this paper mainly to guarantee a good approximation to the original visual BOW representation, while this term is defined in [Cai *et al.*, 2011] mainly to find a good dimension reduction; 2) we do not consider the nonnegative constraints and thus a sound initialization can be derived from eigenvalue decomposition, while for [Cai *et al.*, 2011] only a random initialization (which may

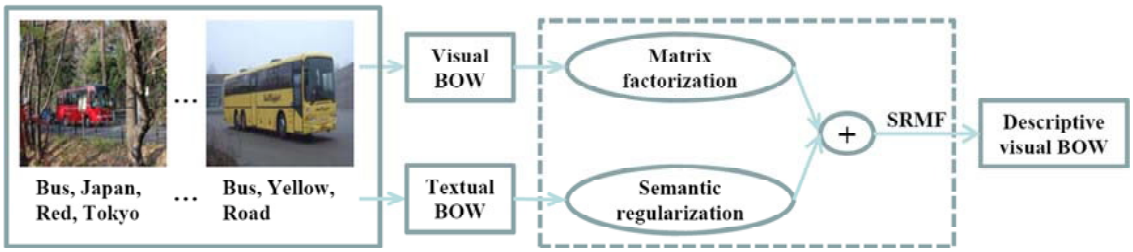

Figure 1: Illustration of our semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags of images.

severely affect the performance) can be provided without any prior knowledge of the original visual BOW representation. More importantly, our later experiments also show that our SRMF method obviously outperforms [Cai *et al.*, 2011].

It should be noted that the tags of images used for learning descriptive visual BOW representation are very easy to access (although noisy) for community-contributed image collections (e.g. Flickr). In contrast, the class labels of images used for learning image representation based on non-negative matrix factorization in [Liu *et al.*, 2012] are commonly very expensive to obtain in practice. Similarly, it is also very expensive to obtain the constraints with respect to local keypoints, although this kind of high-level semantics has been successfully used for visual vocabulary optimization in [Mallapragada *et al.*, 2010]. In addition, other than many previous approaches [Ji *et al.*, 2009; Liu *et al.*, 2009; Lu and Peng, 2011] to visual vocabulary optimization that have ignored the high-level semantics, our SRMF method can explicitly utilize the tags of images (i.e. semantics) for learning descriptive visual BOW representation.

In summary, we propose a novel semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags of images, which can effectively reduce the semantic gap associated with the traditional visual BOW representation. As illustrated in Figure 1, the proposed SRMF method consists of two key steps: matrix factorization over visual BOW representation, and semantic regularization with textual BOW representation (derived from the tags of images). Here, it is worth noting that our SRMF method is efficient even for large image datasets. More notably, when the global visual features are also utilized for image classification, we can *obtain the best results so far* (to our best knowledge) 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 image classification tasks, our SRMF method can be readily extended to other challenging tasks such as image annotation and retrieval.

The remainder of this paper is organized as follows. In Section 2, we develop a novel semantic regularized matrix factorization (SRMF) method for learning descriptive visual BOW representation by exploiting the tags of images. In Section 3, the learnt descriptive visual BOW representation is evaluated on two benchmark datasets by directly applying it to image classification tasks. Finally, Section 4 gives the conclusions drawn from our experimental results.

2 Semantic Regularized Matrix Factorization

This section presents our semantic regularized matrix factorization method in detail. We first give our problem formulation for learning descriptive visual BOW representation from a low-rank matrix factorization viewpoint, and then develop an efficient SRMF algorithm based on the label propagation technique proposed in [Zhou *et al.*, 2004].

2.1 Problem Formulation

In this paper, to reduce the semantic gap associated with the original visual BOW representation, we focus on learning descriptive visual BOW representation by exploiting the tags of images, which are easy to access (although noisy) for community-contributed image collections (e.g. Flickr). Here, it should be noted that the tags of images are used as the high-level semantic information for learning descriptive visual BOW representation. Similar to the formation of the visual BOW representation, we generate a new textual BOW representation with the tags of images, i.e., the high-level semantic information has been encoded into the new textual BOW representation. This means that our problem is actually how to learn more descriptive visual BOW representation from the original one by exploiting the textual BOW representation. More importantly, as illustrated in Figure 2, we can transform it into a low-rank matrix factorization problem [Singh and Gordon, 2008; Lee and Seung, 1999]. Our problem formulation will be elaborated as follows.

Let $Y \in R^{N\times M}$ denote the visual BOW representation and $A \in R^{N\times N}$ denote the kernel (affinity) matrix computed over the textual BOW representation, where N is the number of images and M is the number of visual words. In this paper, we only adopt linear kernel to define the similarity matrix over the textual BOW representation. By directly setting the weight matrix $W = A$, we construct an undirected graph $\mathcal{G} = \{\mathcal{V}, W\}$ with its vertex set $\mathcal{V}$ being the set of images. The normalized Laplacian matrix of $\mathcal{G}$ is given by

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

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 W.

Based on the above preliminary notations, the problem of learning descriptive visual BOW representation can be formulated from a low-rank matrix factorization viewpoint:

$$\min_{U,V,\hat{Y}} \frac{1}{2}||\hat{Y} - U^TV||_F^2 + \frac{\lambda}{2}\mathrm{tr}(V^TULU^TV) + \gamma||\hat{Y} - Y||_1, \tag{2}$$

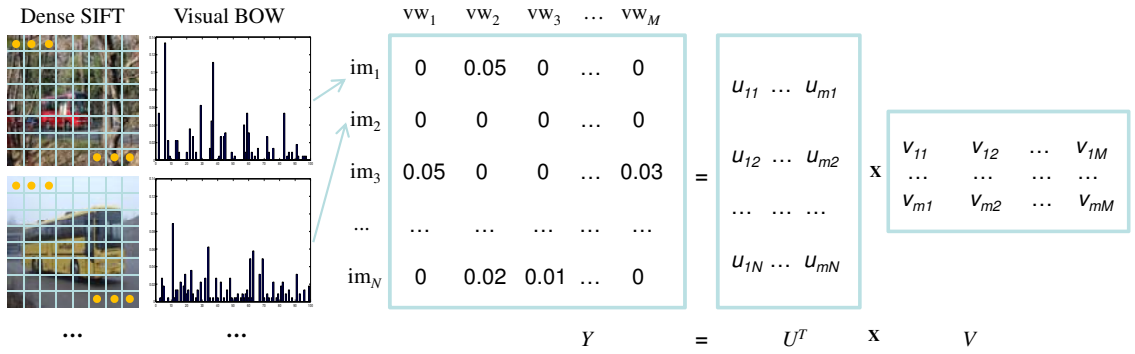

Figure 2: Illustration of learning descriptive visual BOW representation from a low-rank matrix factorization viewpoint ($m \ll \min(N, M)$). Each column denote a visual word vw_j ($j = 1, ..., M$), while each row denotes an image im_i ($i = 1, ..., N$).

where $U \in R^{m\times N}$ and $V \in R^{m\times M}$ denote the low-rank factors, $\hat{Y} \in R^{N\times M}$ denotes the ideal visual BOW representation, λ and γ denote the positive regularization parameters, and $\text{tr}(\cdot)$ denotes the trace of a matrix. The immediate observation is that our problem formulation is quite different from [Cai *et al.*, 2011] which takes the form of $\min_{U,V} \frac{1}{2}||Y - U^TV||_F^2 + \frac{\lambda}{2}\text{tr}(ULU^T)$. Moreover, since the semantic information has been encoded to the normalized Laplacian matrix L, solving Eq. (2) is called as *semantic regularized matrix factorization* (SRMF) in this paper.

The objective function given by Eq. (2) is further discussed as follows. The first term denotes the Frobenius-norm fitting constraint, also used in the standard matrix factorization, which means that the product of U and V should not change too much from $\hat{Y}$. The second term denotes the smoothness constraint, also known as Laplacian regularization [Zhu *et al.*, 2003; Zhou *et al.*, 2004; Fergus *et al.*, 2010], which means that the product of U and V 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 Y due to the nice property of L_1-norm optimization [Elad and Aharon, 2006; Mairal *et al.*, 2008; Wright *et al.*, 2009]. Here, besides the two low-rank factors U and V, 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 U^TV by solving Eq. (2) with $\hat{Y}$ being an intermediate representation. In this paper, we focus on learning descriptive visual BOW representation (i.e. U^TV) based on SRMF, without considering other types of sparsity penalties (e.g. $||U||_1$).

To apply our SRMF 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 descriptive visual BOW representation 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 SRMF algorithm will be elaborated in the next subsection.

2.2 Efficient SRMF Algorithm

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

$$U^*, V^* = \arg\min_{U,V} \frac{1}{2}||\hat{Y}^* - U^TV||_F^2 + \frac{\lambda}{2}\text{tr}(V^TULU^TV),$$

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

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

$$\hat{Y}^* = \text{soft}(U^{*T}V^* - Y, \gamma) + Y, \tag{3}$$

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

Let $\mathcal{Q}(U, V) = \frac{1}{2}||\hat{Y}^* - U^TV||_F^2 + \frac{\lambda}{2}\text{tr}(V^TULU^TV)$. We can still adopt the alternate optimization technique for the first subproblem $\min_{U,V} \mathcal{Q}(U, V)$ as follows: 1) fix $U = U^*$, and learn V by $V^* = \arg\min_V \mathcal{Q}(U^*, V)$; 2) fix $V = V^*$, and learn U by $U^* = \arg\min_U \mathcal{Q}(U, V^*)$.

Learning V: When U is fixed at U^*, the solution of $\min_V \mathcal{Q}(U^*, V)$ can be found by solving

$$\frac{\partial\mathcal{Q}(U^*, V)}{\partial V} = -U^*(\hat{Y}^* - U^{*T}V) + \lambda U^*LU^{*T}V = 0,$$

which can be further transformed into

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

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

Learning U: When V is fixed at V^*, the solution of $\min_U \mathcal{Q}(U, V^*)$ can be found by solving

$$\frac{\partial\mathcal{Q}(U, V^*)}{\partial U} = -V^*((\hat{Y}^*)^T - V^{*T}U) + \lambda V^*V^{*T}UL = 0,$$

which is actually equivalent to

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

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

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

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 $\min_U \mathcal{Q}(U, V^*)$ is found by solving:

$$(V^* V^{*T})U = X^*(U). \tag{7}$$

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

The complete SRMF algorithm for learning descriptive visual BOW representation is outlined as follows:

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

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

(3) Initialize the deal visual BOW representation as $\hat{Y}^* = Y$, where Y is the original visual BOW representation;

(4) Initialize $U = U^* \in R^{m \times N}$ using the m smallest eigenvectors of the normalized Laplacian matrix L, where each row of U^* corresponds to an eigenvector of L;

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

(6) Iterate $X_{t+1}(U) = \alpha X_t(U)(I - L) + (1 - \alpha)V^*(\hat{Y}^*)^T$ until convergence, where a solution can thus be found just as Eq. (6) with $\alpha = \lambda/(1+\lambda)$ (see more explanation below);

(7) Find the best solution U^* by solving Eq. (7): $(V^* V^{*T})U = X^*(U)$, where $X^*(U)$ denotes the limit of the sequence $\{X_t(U)\}$;

(8) Iterate Steps (5)–(7) until the stopping condition is satisfied, and update the deal visual BOW representation as: $\hat{Y}^* = \text{soft}(U^{*T} V^* - Y, \gamma) + Y$;

(9) Iterate Steps (5)–(8) until the stopping condition is satisfied, and output the final descriptive visual BOW representation $U^{*T} V^*$.

Similar to the convergence analysis in [Zhou *et al.*, 2004], the iteration in Step (6) converges to $X^*(U) = V^*(\hat{Y}^*)^T(1 - \alpha)(I - \alpha(I - L))^{-1}$, 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 (6) , (8), and (9) generally converge in very limited number of iteration steps (<10). Finally, since L is computed over k-NN graph, finding m smallest eigenvectors of sparse L in Step (4) has a time complexity of $O(m^2 N + kmN)$. Meanwhile, the time complexity of Steps (5-8) is respectively $O(m^2 M + mMN + m^2 N + kmN)$, $O(mMN + kmN)$, $O(m^2 M + m^2 N)$, and $O(mMN)$. Hence, given that $m, k \ll \min(N, M)$, the proposed algorithm can be applied to large datasets.

To give an explicit explanation of the descriptive visual BOW representation, we show the comparison between the descriptive and original visual BOW representations in Figure 3. Here, we conduct the experiments on a subset of the

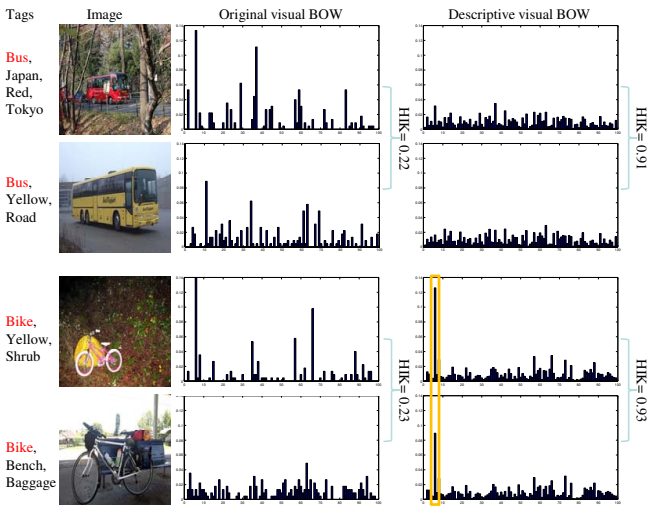

Figure 3: Comparison between the descriptive and original visual BOW representations. The similarity of two histograms is measured by histogram-intersection kernel (HIK). It can be observed that the intra-class variations due to scale changes and cluttered backgrounds can be reduced through learning descriptive visual BOW representation. That is, the semantic gap associated with the original visual BOW representation is indeed reduced to some extent by our semantic regularized matrix factorization.

PASCAL VOC'07 dataset [Everingham *et al.*, 2007]. We can observe that the intra-class variations due to scale changes and cluttered backgrounds can be reduced by our semantic regularized matrix factorization in terms of the histogram-intersection kernel (HIK) values [Barla *et al.*, 2003]. This means that the semantic gap associated with the original visual BOW representation is indeed reduced to some extent by exploiting the tags of images. Another interesting observation is that some visual words of the descriptive visual BOW representation tend to be explicitly related to the high-level semantics of images. For example, the visual word marked by an orange box is clearly shown to be related to "bike", considering that this visual word becomes dominative (*originally far from dominative*) in the histogram of the fourth image after learning descriptive visual BOW representation. This observation further verifies the effectiveness of our semantic regularized matrix factorization. In the next section, we will conduct more extensive experiments in image classification to evaluate the performance of our algorithm.

3 Experimental Results

In this section, the proposed SRMF algorithm for learning descriptive visual BOW representation is evaluated in 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 algorithm is compared with other closely related methods on the two benchmark datasets.

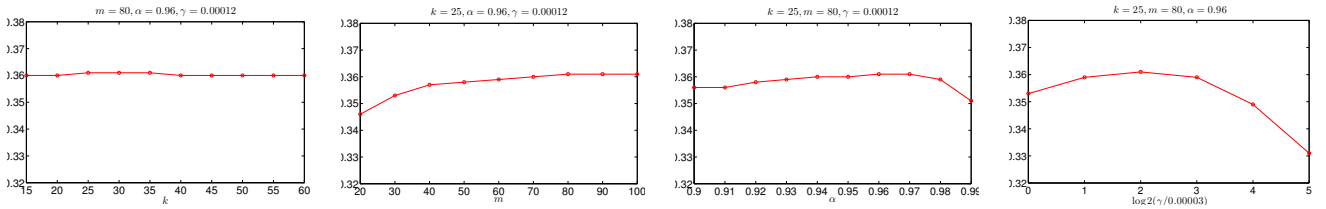

Figure 4: The cross-validation classification results using the descriptive visual BOW representations learnt by our SRMF algorithm on the training set of the PASCAL VOC'07 dataset.

3.1 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 in this paper is reduced to 804 by the same preprocessing step as [Guillaumin *et al.*, 2010]. This benchmark 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 benchmark 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 *et al.*, 2010]. The task of image classification on these two datasets is rather challenging, given that each image may belong to multiple classes and each class may have large intra-class variations.

For each benchmark dataset, we extract the same feature set as [Guillaumin *et al.*, 2010]. That is, we use local SIFT features [Lowe, 2004] and local hue histograms [van de Weijer and Schmid, 2006], 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 *et al.*, 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 the same as [Guillaumin *et al.*, 2010]. Here, we only adopt k-means clustering for visual BOW generation, regardless of other more advanced clustering techniques [Bagirov, 2008; Moosmann *et al.*, 2008].

To evaluate the learnt descriptive visual BOW representation, we apply it directly to 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 *et al.*, 2010]. To verify the effectiveness of learning descriptive visual BOW representation for image classification based on our semantic regularized matrix factorization, we make comparison among the following four closely related matrix factorization methods:

(1) Semantic Regularized Matrix Factorization (SRMF): $\min_{U,V,\hat{Y}} \frac{1}{2}||\hat{Y}-U^TV||_F^2+\frac{\lambda}{2}\mathrm{tr}(V^TULU^TV)+\gamma||\hat{Y}-Y||_1$;

(2) Regularized Matrix Factorization (RMF) [Takács *et al.*, 2008]: $\min_{U,V} \frac{1}{2}||Y-U^TV||_F^2+\frac{\lambda}{2}(||U||_F^2+||V||_F^2)$;

(3) Graph Regularized Nonnegative Matrix Factorization (GRNMF) [Cai *et al.*, 2011]: $\min_{U,V>0} \frac{1}{2}||Y-U^TV||_F^2+\frac{\lambda}{2}\mathrm{tr}(ULU^T)$;

(4) Sparse Nonnegative Matrix Factorization (SPNMF) [Kim and Park, 2007]: $\min_{U,V>0} \frac{1}{2}||Y-U^TV||_F^2+\frac{\lambda}{2}||V||_F^2+\gamma||U||_1$.

Here, both GRNMF and our SRMF exploit the tags of images for learning descriptive visual BOW representation through Laplacian regularization, while the other two methods do not consider this semantic information for matrix factorization. In the experiments, the parameters of our SRMF algorithm are selected by cross-validation on the training set. For example, according to Figure 4, we set the two parameters of our SRMF algorithm on the PASCAL VOC'07 dataset as: $k = 25$, $m = 80$, $\alpha = 0.96$ and $\gamma = 0.00012$ (which appear in Steps 1, 4, 5 (or 6), 8 of our algorithm proposed in Section 2.2, respectively). In particular, the performance of our SRMF algorithm is shown to only become slightly better (or keep unchanged) when the parameter m is increased to certain level (e.g. $\geq$80 here). The same parameter selection strategy is adopted by the other matrix factorization methods for learning descriptive visual BOW representation.

3.2 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 descriptive visual BOW representation learnt by our SRMF algorithm significantly outperforms the original visual BOW representation. That is, the tags of images have been successfully added to the descriptive visual BOW representation and thus the semantic gap associated with the original visual BOW representation has been reduced effectively. More notably, our descriptive visual BOW representation is even shown to achieve more than 37% gains over the original textual BOW representation for both of the two benchmark datasets. The significant gains over both of the original visual and textual BOW representations are due to the fact that our SRMF algorithm can exploit these two types of BOW representations simultaneously for learning descriptive visual BOW representation. In other words, the original visual and textual BOW representations can complement each other well for image classification.

The comparison between different matrix factorization methods for learning descriptive visual BOW representation is further shown in Figure 5(b). Here, it should be noted

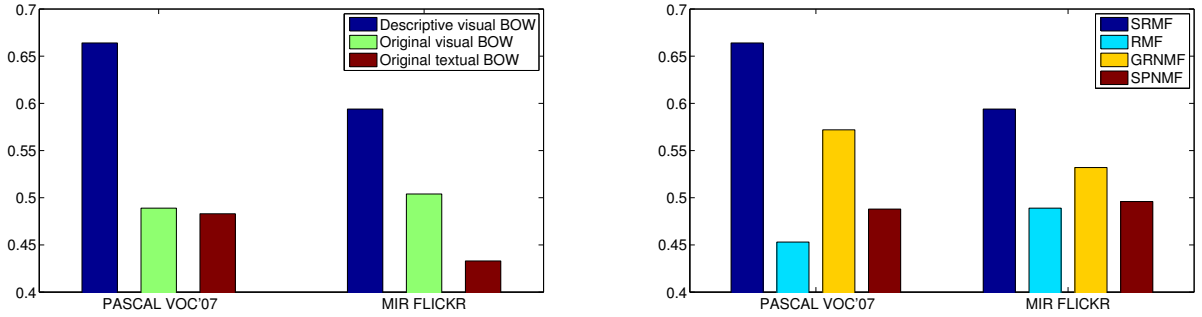

Figure 5: The test classification results using descriptive visual BOW representations on the two benchmark datasets: (a) comparison between different BOW representations; (b) comparison between different matrix factorization methods.

that only GRNMF and our SRMF can exploit the tags of images for learning descriptive visual BOW representation through Laplacian regularization, while the other two methods do not consider this semantic information for matrix factorization. From Figure 5(b), we find that both GRNMF and our SRMF obviously outperform the other two matrix factorization methods (i.e. RMF and SPNMF). These impressive results are due to the fact that the tags of images can be exploited by Laplacian regularization to reduce the semantic gap associated with the original visual BOW representation, while such important semantic information is completely ignored by the other two matrix factorization methods. Moreover, when the two Laplacian regularized matrix factorization methods (i.e. GRNMF and SRMF) are compared to each other, our SRMF is shown to achieve obvious improvements over GRNMF, since we define a different Laplacian regularization term to guarantee a good approximation to the original visual BOW representation (other than only a good dimension reduction guaranteed by GRNMF) and also provide a sound initialization based on eigenvalue decomposition (instead of random initialization used by GRNMF).

Besides the above advantages in image classification, our SRMF method has another advantage, i.e., it can run very fast even on large image datasets. For example, the running time of learning descriptive visual BOW representation taken by SRMF, RMF, GRNMF and SPNMF on PASCAL VOC'07 is 3.0, 2.1, 6.9 and 3.5 minutes, respectively. Here, we run the algorithms (Matlab code) on a computer with 3GHz CPU and 32GB RAM. It can be clearly observed that our SRMF runs faster than GRNMF and SPNMF, while RMF runs the fastest without considering Laplacian regularization.

The comparison of our SRMF 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 *et al.*, 2010] has reported the best results so far for image classification on the PASCAL VOC'07 and MIR FLICKR datasets. However, when the descriptive 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 *et al.*, 2010] on both of the two benchmark datasets. This becomes more impressive given that the present work makes use of *much weaker* global visual features than [Guillaumin *et al.*, 2010] (i.e. two types vs. seven types). Moreover, from Table 1, we can also observe that both [Guillaumin *et al.*, 2010] and our method obviously outperform the winner of PASCAL VOC'07 due to the effective use of extra tags for image classification.

Table 1: Comparison of our SRMF 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	0.664	0.594
Ours (LVF+GVF)	yes	yes	yes	**0.698**	**0.643**

4 Conclusions

In this paper, to reduce the semantic gap associated with the traditional visual BOW representation, we have developed an efficient semantic regularized matrix factorization method for learning descriptive visual BOW representation by adding Laplacian regularization defined with the tags of community-contributed images into matrix factorization. The effectiveness of the proposed method has been verified by the extensive experimental results on the PASCAL VOC'07 and MIR FLICKR datasets. When we also consider the global visual features for image classification, we can obtain even more impressive results on these two benchmark datasets. For future work, the proposed method will be extended to more challenging tasks such as video content analysis.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61073084 and 61202231, Beijing Natural Science Foundation of China under Grants 4122035 and 4132037, Ph.D. Programs Foundation of Ministry of Education of China under Grants 20120001110097 and 20120001120130, and National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503.

References

[Bagirov, 2008] A.M. Bagirov. Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition*, 41(10):3192–3199, 2008.

[Barla *et al.*, 2003] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *Proceedings of International Conference on Image Processing*, volume 3, pages 513–516, 2003.

[Cai *et al.*, 2011] D. Cai, X. He, J. Han, and T.S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

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

[Everingham *et al.*, 2007] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[Fergus *et al.*, 2010] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *Advances in Neural Information Processing Systems 22*, pages 522–530, 2010.

[Guillaumin *et al.*, 2010] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *Proc. CVPR*, pages 902–909, 2010.

[Huiskes and Lew, 2008] M.J. Huiskes and M.S. Lew. The MIR Flickr retrieval evaluation. In *Proceedings of ACM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[Ji *et al.*, 2009] R. Ji, X. Xie, H. Yao, and W.-Y. Ma. Vocabulary hierarchy optimization for effective and transferable retrieval. In *Proc. CVPR*, pages 1161–1168, 2009.

[Kim and Park, 2007] H. Kim and H. Park. Sparse non-negatice matrix factorization via alternating non-negativity-constrained least aquares for microarray data analysis. *SIAM Journal on Matrix Analysis and Applications*, 23(12):1495–1502, 2007.

[Lazebnik *et al.*, 2006] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, pages 2169–2178, 2006.

[Lee and Seung, 1999] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[Li *et al.*, 2008] J. Li, W. Wu, T. Wang, and Y. Zhang. One step beyond histograms: Image representation using Markov stationary features. In *Proc. CVPR*, pages 1–8, 2008.

[Liu *et al.*, 2009] J. Liu, Y. Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *Proc. CVPR*, pages 461–468, 2009.

[Liu *et al.*, 2012] H. Liu, Z. Wu, X. Li, D. Cai, and T.S. Huang. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1299–1311, 2012.

[Lowe, 2004] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Lu and Peng, 2011] Z. Lu and Y. Peng. Combining latent semantic learning and reduced hypergraph learning for semi-supervised image categorization. In *Proceedings of ACM International Conference on Multimedia*, pages 1409–1412, 2011.

[Mairal *et al.*, 2008] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[Mallapragada *et al.*, 2010] P. Mallapragada, R. Jin, and A. Jain. Online visual vocabulary pruning using pairwise constraints. In *Proc. CVPR*, pages 3073–3080, 2010.

[Moosmann *et al.*, 2008] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1632–1646, 2008.

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

[Singh and Gordon, 2008] A. Singh and G. Gordon. A unified view of matrix factorization models. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 2, pages 358–373, 2008.

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

[Takács *et al.*, 2008] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Investigation of various matrix factorization methods for large recommender systems. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, pages 6:1–6:8, 2008.

[van de Weijer and Schmid, 2006] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proc. ECCV*, pages 334–348, 2006.

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

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

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. ICML*, pages 912–919, 2003.