Weakly Supervised Matrix Factorization for Noisily Tagged Image Parsing

Yulei Niu[1*] and **Zhiwu Lu**[1*†] and **Songfang Huang**[2] and **Peng Han**[1] and **Ji-Rong Wen**[1]
[1]School of Information, Renmin University of China, Beijing 100872, China
[2]IBM China Research Lab, Beijing, China
{zhiwu.lu, jirong.wen}@gmail.com, huangsf@cn.ibm.com

Abstract

In this paper, we propose a Weakly Supervised Matrix Factorization (WSMF) approach to the problem of image parsing with noisy tags, i.e., segmenting noisily tagged images and then classifying the regions only with image-level labels. Instead of requiring clean but expensive pixel-level labels as strong supervision in the traditional image parsing methods, we take noisy image-level labels as weakly-supervised constraints. Specifically, we first over-segment all the images into multiple regions which are initially labeled based upon the image-level labels. Moreover, from a low-rank matrix factorization viewpoint, we formulate noisily tagged image parsing as a weakly supervised matrix factorization problem. Finally, we develop an efficient algorithm to solve the matrix factorization problem. Experimental results show the promising performance of the proposed WSMF algorithm in comparison with the state-of-the-arts.

1 Introduction

Noisily tagged image parsing is an extensive problem of image parsing. The goal of image parsing is originally to segment images into multiple regions and identify their categories (i.e., infer pixel-level labels). In recent years, image parsing has become popular and drawn much attention. Traditional approaches to image parsing require pixel-level labels as fully or partial supervisory information [Li *et al.*, 2009; Socher and Li, 2010; Liu *et al.*, 2011; Achanta *et al.*, 2012]. Since collecting pixel-level labels is time-consuming, these approaches cannot be widely applied in practice. As a result, many efforts have been made to exploit image-level labels for image parsing [Vezhnevets and Buhmann, 2010; Vezhnevets *et al.*, 2011; 2012]. The goal of image parsing is now to infer pixel-level labels from image-level labels. Since lots of photo-sharing websites (e.g., Flickr) provide us with plenty of social images, it is easy to collect images-level labels from the tags provided by social users. It should be noted that the social tags may be noisy in practice [Tang *et al.*, 2009]. Image parsing only with noisy image-level labels is really an interesting but challenging problem. In fact, such noisily tagged setting has been rarely considered in recent works on image parsing [Vezhnevets and Buhmann, 2010; Vezhnevets *et al.*, 2012; Zhang *et al.*, 2013; Xu *et al.*, 2014]. Our motivation is thus to build a robust and efficient model for such noisy weakly-supervised setting.

Inspired by the successful use of matrix factorization for image representation [Cai *et al.*, 2011; Liu *et al.*, 2012], we propose a novel Weakly Supervised Matrix Factorization (WSMF) approach to solve the challenging image parsing problem under the noisy weakly-supervised setting (i.e., only noisy image-level labels are provided initially). The proposed WSMF approach has two main components: 1) oversegment each image into multiple regions, and 2) annotate the regions based upon the initial noisy image-level labels. Different from recent work [Liu *et al.*, 2013] that takes clean and complete labels of images as supervisory information, we do not impose any restriction on the initial image-level labels.

The proposed WSMF approach is illustrated in Figure 1. Here, we adopt the Blobworld method [Carson *et al.*, 2002] for automatic over-segmentation. To deal with the noisy image-level labels, we choose to decompose the matrix that collects the initial region-level labels into low-rank matrices. Furthermore, to guarantee a better solution to matrix factorization, we define a new Laplacian regularization term [Zhu *et al.*, 2003; Zhou *et al.*, 2003] and add it into the objective function of matrix factorization. The resulting WSMF problem is finally solved by developing an efficient algorithm based on the label propagation technique [Zhou *et al.*, 2003]. It should be noted that the proposed WSMF approach is quite different from [Cai *et al.*, 2011] that also proposed a Laplacian regularized matrix factorization method. Specifically, we do not consider the nonnegative constraints and thus can derive a sound initialization from eigenvalue decomposition, while only a random initialization can be provided for [Cai *et al.*, 2011] which may severely affect the performance. More notably, as shown in our later experiments, the proposed WSMF approach can provide a better solution to matrix factorization. In this paper, to verify the effectiveness of the proposed approach, we conduct experiments on two public benchmark datasets: MSRC [Shotton *et al.*, 2006] and LabelMe [Liu *et al.*, 2009]. The experimental results show the encouraging and robust performance of the proposed approach.

[*]Co-first authors of equal contributions
[†]Corresponding author

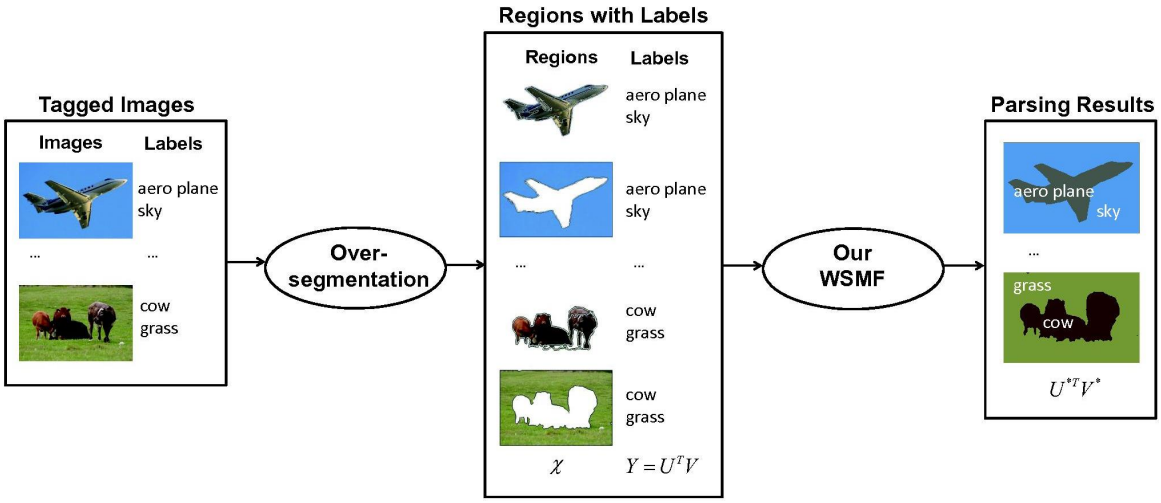

Figure 1: The flowchart of our Weakly Supervised Matrix Factorization (WSMF) for image parsing.

Our main contributions are summarized as follows:

- We propose a novel low-rank matrix factorization approach to address the challenging problem of noisily tagged image parsing. To the best of our knowledge, we are *the first to formulate image parsing as matrix factorization* from the viewpoint of noise reduction over the region-level labels, which is an active extension to the traditional image parsing problem.
- We define a new Laplacian regularized term in our problem formulation, which is shown to improve the performance of matrix factorization for image parsing under the noisy weakly-supervised setting.
- The encouraging results of the proposed approach show that it is much more flexible for image parsing in real-world applications, since only noisy image-level labels are used as supervisory information.

The reminder of this paper is organized as follows. Section 2 gives a brief review of related work. In Section 3, we formulate image parsing as a Weakly Supervised Matrix Factorization (WSMF) problem. In Section 4, we develop an efficient WSMF algorithm. In Section 5, we apply the proposed algorithm to noisily tagged image parsing. Section 6 provides the experimental results on two benchmark datasets. Finally, Section 7 draws our conclusions.

2 Related Work

Supervised Settings for Image Parsing. In recent works, fully-supervised and weakly-supervised settings are widely applied for image parsing. In [Shotton *et al.*, 2006; Lucchi *et al.*, 2012], a fully-supervised setting is considered for image parsing where pixel-level labels are provided at training time. However, pixel-level labels are time-consuming to obtain in practice. In [Verbeek and Triggs, 2007; Vezhnevets and Buhmann, 2010; Vezhnevets *et al.*, 2011; 2012; Liu *et al.*, 2013; Nguyen *et al.*, 2013; Huang *et al.*, 2014], image-level labels are considered for image parsing as weakly-supervised information, which are easier to obtain in many applications than pixel-level labels, and thus the annotation cost can be significantly reduced. However, extra constraints are imposed on the initial image-level labels. For example, image-level labels need to be clear and complete in [Liu *et al.*, 2013] due to the special model used for image parsing. It is worth noting that image-level labels provided by users may be noisy in social image collections [Tang *et al.*, 2009]. It remains a challenging task to effectively exploit the image-level labels for image parsing under this noisy weakly-supervised setting.

Matrix Factorization for Data Representation. Matrix factorization techniques have been successfully applied for learning data representation. The main point of matrix factorization is to find two or more low-rank matrices whose product is a good approximation to the original matrix [Liu *et al.*, 2012]. The traditional methods include Vector Quantization, LU-decomposition, QR-decomposition, and Singular Value Decomposition (SVD). Recently, [Cai *et al.*, 2011] proposed a Graph Regularized Non-negative Matrix Factorization (GNMF) approach based on the original Non-negative Matrix Factorization (NMF) algorithm. In NMF, non-negative constraints are imposed to promise two non-negative matrices results, and GNMF further encodes the geometrical information of the data space by constructing a nearest neighbor graph. This extension also promotes the application of NMF from unsupervised learning field to semi-supervised learning algorithm.

3 Weakly Supervised Matrix Factorization

In this section, we give our problem formulation for image parsing from a low-rank matrix factorization viewpoint under the noisy weakly-supervised setting.

3.1 Notations

Given a dataset of images as the inputs, we adopt the Blobworld method [Carson *et al.*, 2002] to over-segment each image into multiple regions, and collect the set of regions into

$\mathcal{X} = \{x_1, ..., x_N\}$, where N is the total number of regions and x_i is the feature descriptor of the i-th region. Such over-segmentation step will be described in detail in Section 5.

Now we have a set of regions $\mathcal{X}$. Let $Y = \{y_{ij}\}_{N\times C}$ denote the initial labels of regions, where C is the number of object categories. The initial region-level labels are inferred from the initial image-level labels as follows:

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to an image from category } j \\ 0 & \text{otherwise} \end{cases}$$

3.2 Graph Model

Since visually similar regions have higher probability to share the same label, we represent the relationships of regions using a weight matrix W defined as

$$w_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|_2^2}{\sigma^2}) & x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(x)$ is the set of k-nearest regions of x. We actually define the weight matrix based on the Gaussian kernel.

Then we can model the set of regions as a graph $\mathcal{G} = \{\mathcal{V}, W\}$. The vertex set $\mathcal{V}$ is defined as $\mathcal{X}$ and the weight matrix is defined as $W = \{w_{ij}\}_{N\times N}$. The normalized Laplacian matrix $\mathcal{L}$ of $\mathcal{G}$ is given by

$$\mathcal{L} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{1}$$

where I is an $N \times N$ identity matrix, and D is an $N \times N$ diagonal matrix with $D_{ii} = \sum_{j=1}^{N} w_{ij}$.

3.3 Fitting Constraints and Regularization

In this paper, we choose to find two new matrices U and V to obtain the optimal approximation of Y as

$$\hat{Y} = U^T V \approx Y,$$

where $U \in R^{m\times N}$ and $V \in R^{m\times C}$ denote the two low-rank matrix factors. The Frobenius-norm fitting constraint can then be defined as follows:

$$\left\| \hat{Y} - U^T V \right\|_F^2 \tag{2}$$

This means that the product of U and V should not change too much from $\hat{Y}$, which can be considered as an intermediate representation of Y.

Considering the advantage of L_1-norm optimization in noise reduction [Elad and Aharon, 2006; Mairal *et al.*, 2008; Wright *et al.*, 2009; Lu *et al.*, 2015], we define a L_1-norm fitting constraint term as follows:

$$\left\| \hat{Y} - Y \right\|_1 \tag{3}$$

which can impose direct noise reduction on Y.

To guarantee that the product of U and V should not change too much between similar regions, we define the following smoothness constraint term related to the well-known Laplacian regularization [Zhou *et al.*, 2003; Zhu *et al.*, 2003]:

$$\frac{1}{2}\sum_{i,j=1}^{N} w_{i,j} \left\| \frac{y_i}{\sqrt{D_{ii}}} - \frac{y_j}{\sqrt{D_{jj}}} \right\|_2^2 = \mathrm{tr}(V^T U L U^T V) \tag{4}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

3.4 The Proposed Formulation

By combining the above two fitting-constraint terms and one regularization term together, we formulate noisily tagged image parsing as the following Weakly Supervised Matrix Factorization problem from the viewpoint of noise reduction over the labels of regions:

$$\min_{U,V,\hat{Y}} \frac{1}{2}\|\hat{Y} - U^T V\|_F^2 + \frac{\lambda}{2}\mathrm{tr}(V^T U L U^T V) + \gamma\|\hat{Y} - Y\|_1 \tag{5}$$

where λ and γ are the regularization parameters. Our main motivation is to impose direct noise reduction on Y by using the L_1-norm fitting constraint $\|\hat{Y} - Y\|_1$. With $\hat{Y}$ being an intermediate representation, we can transfer the effect of noise reduction to $U^T V$ by solving Eq. (5).

After we have formulated noisily tagged image parsing from the noise reduction viewpoint, we will further discuss how to solve the WSMF problem efficiently. Considering the special definition of Laplacian regularization in Eq. (5), the WSMF problem can be solved efficiently by using the label propagation technique [Zhou *et al.*, 2003] based on k-nearest neighbors (k-NN) graph.

4 Efficient WSMF Algorithm

The optimization problem in Eq. (5) can be solved in two alternate steps as follows:

$$U^*, V^* = \arg\min_{U,V} \frac{1}{2}\|\hat{Y}^* - U^T V\|_F^2 + \frac{\lambda}{2}\mathrm{tr}(V^T U L U^T V) \tag{6}$$

$$\hat{Y}^* = \arg\min_{\hat{Y}} \frac{1}{2}\|\hat{Y} - U^{*T} V^*\|_F^2 + \gamma\|\hat{Y} - Y\|_1 \tag{7}$$

Here, $\hat{Y}^*$ is initialized with Y. As a basic L_1-norm optimization problem, the second problem can be solved based on the soft-thresholding function:

$$\hat{Y}^* = \mathrm{soft}(U^{*T} V^* - Y, \gamma) + Y \tag{8}$$

where $\mathrm{soft}(y, \gamma) = \mathrm{sign}(y)\max\{|y| - \gamma, 0\}$. To solve the first quadratic optimization subproblem, we develop an efficient algorithm as follows.

Let $\mathcal{Q}(U, V) = \frac{1}{2}\|\hat{Y}^* - U^T V\|_F^2 + \frac{\lambda}{2}\mathrm{tr}(V^T U L U^T V)$. For the first subproblem $\min_{U,V} \mathcal{Q}(U, V)$, we adopt the alternate optimization technique as follows: 1) fix $U = U^*$,and update V by $V^* = \arg\min_V \mathcal{Q}(U^*, V)$; 2) fix $V = V^*$, and update U by $U^* = \arg\min_U \mathcal{Q}(U, V^*)$.

Updating V: When we set $U = U^*$, the solution of $\min_V \mathcal{Q}(U^*, V)$ can be found by solving the linear equation

$$\frac{\partial \mathcal{Q}(U^*, V)}{\partial V} = -U^*(\hat{Y}^* - U^{*T} V) + \lambda U^* L U^{*T} V = 0 \tag{9}$$

which is actually equivalent to the linear equation

$$(U^*(I + \lambda L)U^{*T})V = U^*\hat{Y}^* \tag{10}$$

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

Updating U: When we set $V = V^*$, the solution of $\min_U \mathcal{Q}(U, V^*)$ can be found by solving the linear equation:

$$\frac{\partial \mathcal{Q}(U, V^*)}{\partial U} = -V^*(\hat{Y}^{*T} - V^{*T} U) + \lambda V^* V^{*T} U L = 0 \tag{11}$$

Algorithm 1 Weakly Supervised Matrix Factorization

Input:

Regions $\mathcal{X} = \{x_1, ..., x_N\}$;
Initial image-level labels $P \in R^{M \times C}$;
Parameters k, m, γ, λ.

Output:

Labels of regions $U^{*T}V^*$.

1: Construct a weight matrix W on k-NN graph;
2: Compute the normalized Laplacian matrix $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ according to Eq. (1);
3: Initialize the labels of regions Y referring to the image-level labels;
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. (10) 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. (13) with $\alpha = \lambda/(1 + \lambda) \in (0, 1)$;
7: Find the best solution U^* by solving Eq. (14): $(V^*V^{*T})U = X^*(U)$, 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 $\hat{Y}^*$ 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 labels of regions $U^{*T}V^*$.

which can be further transformed into the linear equation

$$V^*V^{*T}U(I + \lambda L) = V^*\hat{Y}^{*T} \tag{12}$$

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

$$X^*(U) = V^*\hat{Y}^{*T}(I + \lambda L)^{-1} \tag{13}$$

However, due to an $O(N^3)$ time complexity of matrix inverse, this analytical solution is not suitable for large image datasets. Fortunately, this solution can also be *efficiently found using the label propagation technique* proposed in [Zhou *et al.*, 2003] based on k-NN graph. Finally, the solution of $\min_U \mathcal{Q}(U, V^*)$ is found by solving:

$$(V^*V^{*T})U = X^*(U) \tag{14}$$

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

The complete WSMF algorithm for noisily tagged image parsing is outlined as Algorithm 1. Similar to the convergence analysis in [Zhou *et al.*, 2003], the iteration in Step (7) converges to $X^*(U) = V^*Y^{*T}(1-\alpha)(I-\alpha(I-L))^{-1}$, which is equal to the solution given by Eq. (13) with $\alpha = \lambda/(1+\lambda)$. Moreover, in our later experiments, we find that iterations in Steps (6), (8), (9) generally converge in very limited number of iteration steps (≤ 10). Finally, collecting m smallest eigenvectors of sparse L in Step (4) has a time complexity of $O(m^2N + kmN)$. Given that $m, k \ll \min(N, C)$, the time complexity of Steps (6-9) is respectively $O(m^2M + mMN + m^2N + kmN)$, $O(mMN + kmN)$, $O(m^2M + m^2N)$, and $O(mMN)$, the proposed WSMF algorithm can be applied to a large set of regions in practice.

5 Noisily Tagged Image Parsing

In the previous section, we have just developed an efficient WSMF algorithm for noisily tagged image parsing. As for the inputs of our WSMF algorithm, we assume that we have collected a large set of regions with image-level labels in advance. In this section, we will focus on how to generate the large set of regions for our WSMF algorithm.

Given a set of images, we first adopt the Blobworld method [Carson *et al.*, 2002] for over-segmentation. Specifically, we extract a 6-dimensional vector of color and texture features for each pixel of an image and then model this image as a Gaussian mixture model. With all pixels grouped into different regions, the number of regions can be automatically detected by a model selection principle. To ensure an over-segmentation of each image, we modify the original Blobworld method slightly: 1) the number of regions is initially set to a relatively large value; 2) model selection is forced to be less important during over-segmentation.

After over-segmenting all the images, we collect a large set of regions $\mathcal{X} = \{x_1, ..., x_N\}$. Each region is represented as a 137-dimensional feature vector by concatenating color and textual features, which includes three mean color features with their standard deviations (6-dimensional), three mean texture features with their standard deviations (6-dimensional), and 125-dimensional color histogram. Finally, we apply a Gaussian kernel over $\mathcal{X}$ to produce the weight matrix W in our WSMF algorithm.

Different from many previous image parsing methods [Shotton *et al.*, 2006; Ladicky *et al.*, 2009; Vezhnevets *et al.*, 2012] that consume too much time during training the generative or discriminative model, our WSMF algorithm for image parsing runs very efficiently on a large set of regions. As for the time-consuming over-segmentation, we can readily speed it up by running in a distributed way.

6 Experimental Evaluation

In this section, we evaluate our WSMF algorithm on two benchmark datasets. Experiments are conducted to answer the following questions: 1) How do matrix factorization methods perform in the image parsing task? 2) How does our WSMF algorithm perform when more noisy image-level labels are considered? 3) How does our WSMF algorithm perform when compared to the state-of-the-arts?

6.1 Experimental Setup

To evaluate the effectiveness of our WSMF algorithm, we conduct experiments on two public datasets, i.e., MSRC [Shotton *et al.*, 2006] and LabelMe [Liu *et al.*, 2009].

MSRC: The MSRC dataset contains 591 images with 21 different object categories. The dataset is split into 276 training images and 256 test images.

LabelMe: The LabelMe dataset contains 2688 images with 33 different categories. The dataset is split into 2488

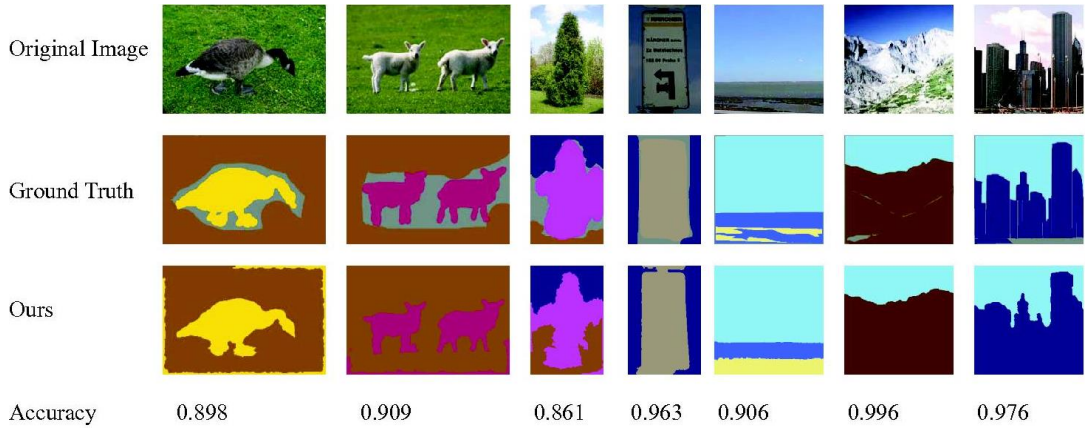

Figure 2: Example results obtained by our WSMF algorithm in comparison with the ground-truth segmentations on the MSRC and LabelMe benchmark datasets. Colors correspond to object categories.

Table 1: Average per-class accuracies (%) of different learning methods for noisily tagged image parsing on the MSRC dataset. The standard deviations are also provided here.

Noisily tagged images	0%	25%	50%	75%	100%
WSMF	**71±0**	**66±1**	**62±1**	**58±1**	**55±1**
[Liu *et al.*, 2013]	67±0	59±1	52±1	46±3	37±2
[Cai *et al.*, 2011]	69±0	64±2	57±3	53±3	47±3
QR	70±0	64±2	57±3	52±3	45±2
SVD	68±0	63±2	57±2	52±2	47±2

Table 2: Average per-class accuracies (%) of different learning methods for noisily tagged image parsing on the LabelMe dataset. The standard deviations are also provided here.

Noisily tagged images	0%	25%	50%	75%	100%
WSMF	**33±0**	**31±1**	**30±1**	**28±1**	**27±1**
[Liu *et al.*, 2013]	27±0	22±3	21±2	17±3	13±1
[Cai *et al.*, 2011]	22±0	21±1	21±2	19±4	17±1
QR	23±0	17±3	17±1	18±2	14±2
SVD	16±0	16±2	17±2	16±1	16±1

training images and 200 test images. It is more challenging than MSRC in image parsing tasks.

To verify the effectiveness of our WSMF algorithm for noisily tagged image parsing, we add random noise to the image-level labels as a simulation of social images. Concretely, we randomly select certain percent of images and then attach an extra wrong label to each selected image. Since most images have about 3 or 4 labels according to the ground-truth segmentations, one extra label for each selected image induce relatively strong noise into the inputs of image parsing. We over-segment each image into multiple regions and then totally obtain about 7,000 regions and 33,000 regions for the two benchmark datasets.

We evaluate the performance of noisily tagged image parsing by using average per-class accuracy, which measures the percentage of correctly classified pixels for a class then averaged over all classes. In the experiments, we make comparison to closely related algorithms under the same noisy weakly-supervised setting. Concretely, since a Laplacian regularization term is considered in our problem formulation, we compare our WSMF algorithm with Graph Regularized Non-negative Matrix Factorization (GNMF) [Cai *et al.*, 2011], which considers a similar Laplacian regularization term with non-negative constraints and we extend it to the noisy weakly-supervised setting for fair comparison. Although Weakly-Supervised Dual Clustering (WSDC) [Liu *et al.*, 2013] is originally designed for weakly-supervised image parsing, we can readily extend it to the noisy weakly-supervised setting for comparison with our WSMF algorithm. Moreover, we make comparison to the traditional matrix-factorization methods such as QR-decomposition (QR) and Singular Value Decomposition (SVD).

It should be noted that *the ground-truth pixel-level labels of all the images are unknown* in our image parsing setting. Hence, it is not possible to select the parameters by cross-validation. In this paper, we thus uniformly set the parameters of our WSMF algorithm as $k = 30$, $\alpha = 0.1$, $\gamma = 0.013$, and $m = 30$ for the two datasets. The parameters of other closely related methods are also set their respective optimal values.

6.2 Parsing Results

The comparison of our WSMF algorithm to other closely related methods is shown in Tables 1 and 2 (see some example results in Figure 2). We can see that our WSMF algorithm performs the best in all cases. That is, our WSMF algorithm is more effective for noisily tagged image parsing. In particular, the improvements achieved by our WSMF algorithm over GNMF are mainly due to the fact that our new Laplacian regularized term is quite different from that used in GNMF. Specifically, our Laplacian regularized term is used to guarantee a good approximation to the intermediate representation of region-level labels, while this term is defined in GNMF to find a good dimension reduction. More importantly, without the nonegative constraints (considered in GNMF), we can drive a sound initialization from eigenvalue decomposition, while GNMF can only take a random initialization. In ad-

Table 3: Accuracy(%) for each category on the MSRC dataset. The last column is the average per-class accuracy.

Supervision	Methods	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	average
Full	[Ladicky *et al.*, 2009]	**80**	**96**	86	74	87	**99**	74	**87**	86	87	82	97	95	30	86	31	**95**	51	69	**66**	9	75
Full	[Csurka and Perronnin, 2011]	75	93	78	70	79	88	66	63	75	76	81	74	44	25	75	24	79	54	55	43	18	64
Full	[Lucchi *et al.*, 2012]	59	90	**92**	82	83	94	**91**	80	85	**88**	96	89	73	48	96	62	81	87	33	44	30	**76**
Weak	[Vezhnevets *et al.*, 2011]	12	83	70	81	**93**	84	**91**	55	**97**	87	92	82	69	51	61	59	66	53	44	9	58	67
Weak	[Zhang *et al.*, 2013]	63	93	**92**	62	75	78	79	64	95	79	93	62	76	32	95	48	83	63	38	68	15	69
Weak	[Akbas and Ahuja, 2014]	74	93	84	61	60	80	55	75	75	62	75	81	71	36	72	25	75	52	39	49	10	62
Weak	WSMF (0% noise)	20	54	55	**96**	85	61	57	40	73	73	**97**	**100**	**99**	**95**	**100**	**99**	26	**100**	**97**	10	**62**	71
Weak	WSMF (25% noise)	18	64	49	85	68	63	51	49	69	59	79	97	90	83	92	83	37	100	85	12	56	66
Weak	WSMF (50% noise)	16	68	47	78	60	67	50	56	58	50	72	87	85	70	86	76	42	94	77	16	54	62

Table 4: Accuracy(%) for each category on the LabelMe dataset. The last column is the average per-class accuracy.

Supervision	Methods	awning	balcony	bird	boat	bridge	building	bus	car	crosswalk	door	fence	field	grass	mountain	person	plant	pole	river	road	rock	sand	sea	sidewalk	sign	sky	staircase	streetlight	sun	tree	window	average
Full	[Tighe and Lazebnik, 2010]	0	0	0	0	2	**87**	0	48	16	4	27	52	**47**	74	1	1	0	12	77	5	**26**	78	28	0	**92**	0	0	**100**	78	7	29
Full	[Liu *et al.*, 2011]	0	0	0	2	**91**	0	39	0	0	7	21	45	32	71	0	12	0	30	**88**	2	3	**87**	36	9	90	2	3	22	78	1	26
Full	[Myeong *et al.*, 2012]	0	3	0	0	9	86	0	56	32	23	22	**57**	30	69	1	3	0	36	79	4	25	82	40	6	**92**	0	0	**100**	**82**	22	32
Full	[Tighe and Lazebnik, 2013]	15	71	0	21	6	84	0	**62**	41	12	**42**	42	44	**80**	4	1	0	19	85	3	17	83	**42**	**42**	**92**	**51**	0	**100**	78	22	**39**
Weak	[Liu *et al.*, 2013]	**32**	**18**	**100**	**85**	10	16	**100**	18	54	17	8	21	11	7	0	**26**	**49**	18	5	21	6	10	4	0	8	18	0	**100**	20	**26**	27
Weak	WSMF (0% noise)	19	**18**	0	62	57	31	29	18	**57**	**44**	37	31	13	37	6	23	**49**	**39**	10	**57**	20	18	7	29	66	6	17	**100**	61	20	33
Weak	WSMF (25% noise)	15	15	0	63	57	27	27	16	48	42	28	31	14	33	**13**	20	49	34	9	49	20	19	7	28	67	6	**18**	100	58	21	31
Weak	WSMF (50% noise)	17	15	0	63	57	26	29	14	57	36	27	30	12	28	11	17	53	28	15	42	20	14	5	27	67	6	15	100	56	19	30

dition, when we add noise into image-level labels for the MSRC dataset, the average per-class accuracy of our WSMF decreases by 4% with the ratio of noisy images increasing by 25%, while the average per-class accuracy of WSDC decreases by 7.5%. And for LabelMe, the average per-class accuracies of WSMF and WSDC decrease by 1.5% and 3.5%, respectively. Such observation shows the effectiveness of our WSMF in noisily tagged image parsing. These results can be used to answer Question 1) and 2).

Besides the above advantages in noisily tagged image parsing, our WSMF algorithm runs efficiently on a large set of regions. For example, the running time of WSMF, WSDC, GNMF, QR, and SVD on MSRC is 16, 52, 22, 10, and 8 seconds, respectively. Here, we run all the algorithms (Matlab code) on a computer with 3.4GHz CPU and 8GB RAM.

6.3 Comparison to the State-of-the-Arts

We show the comparison to more full-supervised methods and weakly-supervised methods for image parsing in Tables 3 and 4. Firstly, our WSMF algorithm achieves higher average per-class accuracies than other weakly-supervised approaches, and even outperforms some fully-supervised approaches on the two benchmark datasets. Secondly, our WSMF algorithm obtains the best results for 10 out of 21 categories on the MSRC dataset and for 9 out of 30 categories on the LabelMe dataset, especially for animals such as cow, sheep, bird, cat and dog, which are difficult to distinguish from the backgrounds by other methods. Thirdly, the number of categories that have zero accuracies is only one for our WSMF algorithm, much lower than other methods on the LabelMe dataset. These results demonstrate that our WSMF algorithm is more effective in more challenging tasks such as image parsing on the LabelMe dataset. This evaluation can be a good answer to Question 3).

7 Conclusion

We have proposed a Weakly Supervised Matrix Factorization approach to image parsing with noisy image-level labels under the weakly-supervised setting. Concretely, we first formulate the problem of noisily tagged image parsing as matrix factorization by defining a novel Laplacian regularized term. Moreover, an efficient WSMF algorithm is developed based on the label propagation technique. The experimental results have demonstrated the promising performance of our WSMF algorithm. In the future work, we plan to 1) apply other methods to obtain initial segmentation labels such as Markov random field-type optimization; and 2) evaluate directly over social image collections.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 61202231 and 61222307, National Key Basic Research Program (973 Program) of China under Grant 2014CB340403, Beijing Natural Science Foundation of China under Grant 4132037, the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China under Grant 15XNLQ01, and IBM Global Faculty Award Program.

References

[Achanta *et al.*, 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[Akbas and Ahuja, 2014] E. Akbas and N. Ahuja. Low-level hierarchical multiscale segmentation statistics of natural images. *PAMI*, 36(9):1900–1906, 2014.

[Cai *et al.*, 2011] D. Cai, X. He, J. Han, and T. Huang. Graph regularized nonnegative matrix factorization for data representation. *PAMI*, 33(8):1548–1560, 2011.

[Carson *et al.*, 2002] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *PAMI*, 24(8):1026–1038, 2002.

[Csurka and Perronnin, 2011] G. Csurka and F. Perronnin. An efficient approach to semantic segmentation. *IJCV*, 95(2):198–212, 2011.

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

[Huang *et al.*, 2014] S.-J. Huang, W. Gao, and Z.-H. Zhou. Fast multi-instance multi-label learning. In *AAAI*, pages 1868–1874, 2014.

[Ladicky *et al.*, 2009] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746, 2009.

[Li *et al.*, 2009] L.-J. Li, R. Socher, and F.-F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009.

[Liu *et al.*, 2009] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, pages 1972–1979, 2009.

[Liu *et al.*, 2011] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33(12):2368–2382, 2011.

[Liu *et al.*, 2012] H. Liu, Z. Wu, D. Cai, and T. Huang. Constrained nonnegative matrix factorization for image representation. *PAMI*, 34(7):1299–1311, 2012.

[Liu *et al.*, 2013] Y. Liu, J. Liu, Z. Li, J. Tang, and H. Lu. Weakly-supervised dual clustering for image semantic segmentation. In *CVPR*, pages 2075–2082, 2013.

[Lu *et al.*, 2015] Zhiwu Lu, Peng Han, Liwei Wang, and Ji-Rong Wen. Semantic sparse recoding of visual content for image applications. *IEEE Trans. Image Processing*, 24(1):176–188, 2015.

[Lucchi *et al.*, 2012] A. Lucchi, Y. Li, K. Smith, and P. Fua. Structured image segmentation using kernelized features. In *ECCV*, pages 400–413. 2012.

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

[Myeong *et al.*, 2012] H. Myeong, J. Chang, and K. Lee. Learning object relationships via graph-based context model. In *CVPR*, pages 2727–2734, 2012.

[Nguyen *et al.*, 2013] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou. Multi-modal image annotation with multi-instance multi-label LDA. In *IJCAI*, pages 1558–1564, 2013.

[Shotton *et al.*, 2006] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, pages 1–15. 2006.

[Socher and Li, 2010] R. Socher and F.-F. Li. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, pages 966–973, 2010.

[Tang *et al.*, 2009] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *ACM Multimedia*, pages 223–232, 2009.

[Tighe and Lazebnik, 2010] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV*, pages 352–365, 2010.

[Tighe and Lazebnik, 2013] J. Tighe and S. Lazebnik. Superparsing. *IJCV*, 101(2):329–349, 2013.

[Verbeek and Triggs, 2007] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, pages 1–8, 2007.

[Vezhnevets and Buhmann, 2010] A. Vezhnevets and J. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *CVPR*, pages 3249–3256, 2010.

[Vezhnevets *et al.*, 2011] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, pages 643–650, 2011.

[Vezhnevets *et al.*, 2012] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised structured output learning for semantic segmentation. In *CVPR*, pages 845–852, 2012.

[Wright *et al.*, 2009] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Yi Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

[Xu *et al.*, 2014] J. Xu, A. Schwing, and R. Urtasun. Tell me what you see and I will show you where it is. In *CVPR*, pages 3190–3197, 2014.

[Zhang *et al.*, 2013] K. Zhang, W. Zhang, Y. Zheng, and X. Xue. Sparse reconstruction for weakly supervised semantic segmentation. In *IJCAI*, pages 1889–1895, 2013.

[Zhou *et al.*, 2003] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2003.

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