Social Image Parsing by Cross-Modal Data Refinement

Zhiwu Lu[1] and **Xin Gao**[2] and **Songfang Huang**[3] and **Liwei Wang**[4] and **Ji-Rong Wen**[1]
[1]School of Information, Renmin University of China, Beijing 100872, China
[2]CEMSE Division, KAUST, Thuwal, Jeddah 23955, Saudi Arabia
[3]IBM China Research Lab, Beijing, China
[4]School of EECS, Peking University, Beijing 100871, China
{zhiwu.lu, xin4gao, wangliwei.pku, jirong.wen}@gmail.com, huangsf@cn.ibm.com

Abstract

This paper presents a cross-modal data refinement algorithm for social image parsing, or segmenting all the objects within a social image and then identifying their categories. Different from the traditional fully supervised image parsing that takes pixel-level labels as strong supervisory information, our social image parsing is initially provided with the noisy tags of images (i.e. image-level labels), which are shared by social users. By oversegmenting each image into multiple regions, we formulate social image parsing as a cross-modal data refinement problem over a large set of regions, where the initial labels of each region are inferred from image-level labels. Furthermore, we develop an efficient algorithm to solve such cross-modal data refinement problem. The experimental results on several benchmark datasets show the effectiveness of our algorithm. More notably, our algorithm can be considered to provide an alternative and natural way to address the challenging problem of image parsing, since image-level labels are much easier to access than pixel-level labels.

1 Introduction

As a fundamental problem in computer vision, image parsing aims to segment all the objects within an image and then identify their categories. In the past years, image parsing has drawn much attention [Shotton *et al.*, 2006; Yang *et al.*, 2007; Shotton *et al.*, 2008; Kohli *et al.*, 2009; Ladicky *et al.*, 2009; 2010; Csurka and Perronnin, 2011; Lucchi *et al.*, 2012; Tighe and Lazebnik, 2013; Chang *et al.*, 2014; Yang *et al.*, 2014]. Although these methods have been reported to achieve promising results, most of them take pixel-level labels as the inputs of image parsing. In real-world applications, pixel-level labels are very expensive to access, and these fully supervised methods cannot be widely applied in practice.

Many recent efforts have been made to exploit image-level labels for image parsing [Verbeek and Triggs, 2007; Vezhnevets and Buhmann, 2010; Vezhnevets *et al.*, 2011; 2012; Liu *et al.*, 2013; Zhang *et al.*, 2013; Liu *et al.*, 2014; Xu *et al.*, 2014], considering that image-level labels are much easier to access than pixel-level labels. The main challenge thus lies in inferring pixel-level labels from this weak supervisory information. As compared to the traditional fully supervised image parsing that takes pixel-level labels as supervisory information, such weakly supervised image parsing is more flexible in real-world applications. However, it is still expensive to collect image-level labels for image parsing. Hence, we hope that such supervisory information can be provided in an automatic and natural way.

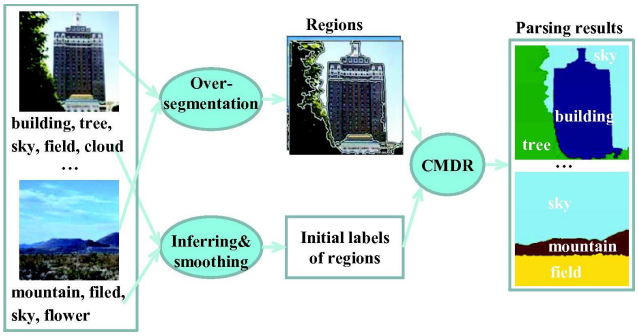

Figure 1: The flowchart of our cross-modal data refinement (CMDR) for social image parsing, where the initial labels of regions are inferred from image-level labels and then smoothed using some prior knowledge.

Due to the burgeoning growth of social images over photo-sharing websites (e.g. Flcikr), this automatic and natural setting becomes feasible for image parsing. That is, the tags of social images can be used as image-level labels for image parsing. It is worth noting that the tags of social images may be noisy (or incomplete) in practice [Tang *et al.*, 2009], although they are very easy to access. Hence, the main challenge lies in effectively exploiting the noisy tags for image parsing. However, such social image parsing has been rarely considered in recent work [Vezhnevets *et al.*, 2012; Liu *et al.*, 2013; Zhang *et al.*, 2013].

In this paper, we focus on developing a cross-modal data refinement approach to solve the challenging problem of social image parsing. The basic idea is to first oversegment all the images into regions and then infer the labels of regions from the initial image-level labels. Since the initial labels of regions cannot be accurately estimated even from clean initial image-level labels, our main motivation is to continuously

suppress the noise in the labels of regions through an iterative procedure. By considering the initial labels and visual features as two modalities of regions, we formulate such iterative procedure as a cross-modal data refinement problem over all the regions. In contrast to [Liu *et al.*, 2013] which is limited to clean and complete image-level labels for image parsing, we do not impose any extra requirement on the initial image-level labels. Based on L_1-optimization [Elad and Aharon, 2006; Chen *et al.*, 2011] and label propagation [Zhu *et al.*, 2003; Zhou *et al.*, 2004], we develop an efficient algorithm to solve such cross-modal data refinement problem.

The flowchart of our cross-modal data refinement (CMDR) for social image parsing is illustrated in Figure 1. Here, we adopt the Blobworld method [Carson *et al.*, 2002] for oversegmentation, since it can automatically detect the number of regions within an image. Meanwhile, we utilize image-level labels to infer the initial labels of regions, which are further smoothed by using some prior knowledge about regions and object categories. In this paper, to study whether our CMDR algorithm can deal with noisily tagged images, we first conduct experiments on the MSRC [Shotton *et al.*, 2006] and LabelMe [Liu *et al.*, 2011] benchmark datasets by adding random noise to the initial image-level labels. These two datasets are originally used in recent work on image parsing [Csurka and Perronnin, 2011; Lucchi *et al.*, 2012; Liu *et al.*, 2013; Zhang *et al.*, 2013] without adding random noise. To obtain more convincing parsing results, we further conduct experiments by collecting a Flickr dataset with realistic noise, where the noisy image-level labels are directly downloaded from the Flickr website. As shown in later experiments, our investigation with random and realistic noise has provided an *alternative and natural* way to address the challenging problem of image parsing, given that the noisy image-level labels can be easily obtained from social image collections. In fact, our social image parsing is somewhat similar to verbal guided image parsing [Cheng *et al.*, 2014] that aims to perform automatic image parsing using the verbal guidance.

To emphasize our main contributions, we summarize the following distinct advantages of the present work:

- This is the first attempt to formulate social image parsing as cross-modal data refinement, to the best of our knowledge. In fact, the problem of social image parsing has been rarely considered in the literature.
- We have successfully developed an efficient algorithm for social image parsing, unlike many previous image parsing approaches that incur too large time cost.
- The proposed algorithm can be considered to provide an alternative and natural way for image parsing, since the noisy image-level labels are easy to access.

2 Cross-Modal Data Refinement

2.1 Problem Formulation

The problem of social image parsing is described as follows. Given a set of social images, we adopt the Blobworld method [Carson *et al.*, 2002] to oversegment each image and then extract a feature vector (including color and texture features) from each region. All the features vectors are collected into

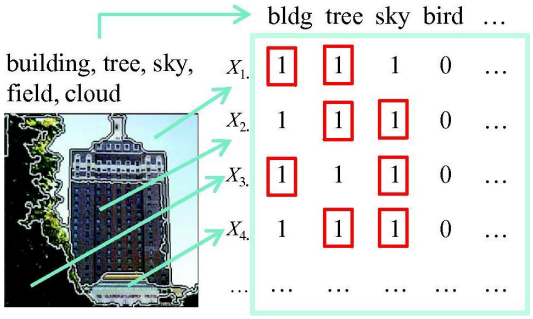

Figure 2: The initial estimation of Y by a simple inference from image-level labels. Here, the labels of regions wrongly estimated are marked by red boxes.

$X \in R^{N\times d}$ with each row $X_{i.}$ being a region, where N is the total number of regions and d is the dimension of feature vectors. Moreover, as illustrated in Figure 2, the initial labels of all the regions $Y = \{y_{ij}\}_{N\times C}$ are inferred from the image-level labels provided for image parsing as: $y_{ij} = 1$ if the region $X_{i.}$ belongs to an image which is labeled with category j and $y_{ij} = 0$ otherwise, where C is the number of object categories. Here, the initial labels of regions *cannot be accurately estimated by such simple inference*, even if clean image-level labels are provided initially. The noise issue becomes more severe when noisy image-level labels are used as supervisory information. Hence, we need to pay attention to noise reduction over Y. In the following, we will formulate it as a cross-modal data refinement problem.

We first model the whole set of regions as a graph $\mathcal{G} = \{\mathcal{V}, A\}$ with its vertex set $\mathcal{V}$ being the set of regions and affinity matrix $A = \{a_{ij}\}_{N\times N}$, where a_{ij} denotes the affinity between region $X_{i.}$ and region $X_{j.}$. The affinity matrix A is usually defined by a Gaussian kernel in the literature. The normalized Laplacian matrix L of $\mathcal{G}$ is given by

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}, \tag{1}$$

where I is an $N \times N$ identity matrix, and D is an $N \times N$ diagonal matrix with its i-th diagonal element being equal to the sum of the i-th row of A (i.e. $\sum_j a_{ij}$).

Based on these notations, we formulate social image parsing as a cross-modal data refinement problem:

$$\min_{\hat{Y}\geq 0,\hat{X},W} \frac{1}{2}||\hat{Y} - \hat{X}W||_F^2 + \frac{\lambda}{2}\mathrm{tr}(W^T\hat{X}^T L\hat{X}W) + \gamma||\hat{Y} - Y||_1, \tag{2}$$

where $\hat{Y} \in R^{N\times C}$ stores the ideal labels of regions, $\hat{X} \in R^{N\times d}$ denotes the ideal visual representation of regions (initialized as X), $W \in R^{d\times C}$ denotes the correlation matrix between $\hat{Y}$ and $\hat{X}$, λ and γ denote the positive regularization parameters, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Since the two modalities of regions (i.e. $\hat{X}$ and $\hat{Y}$) can be optimized alternately (see our later explanation), solving Eq. (2) is called as *cross-modal data refinement* (CMDR) 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 $\hat{X}W$ 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; Fu *et al.*, 2011], which means that $\hat{X}W$ should not change too much between similar regions. 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; Figueiredo *et al.*, 2007; Mairal *et al.*, 2008; Wright *et al.*, 2009; Chen *et al.*, 2011].

Although we have given the problem formulation for social image parsing, there remain two key problems to be addressed: 1) *how to smooth the initial labels of regions stored in* Y, and 2) *how to efficiently solve the cross-modal data refinement problem in Eq. (2)*. In the next two subsections, we will address these two key problems, respectively.

2.2 Initial Label Smoothing

It should be noted that there exists much noise in the initial labels of regions stored in Y which are estimated simply from the image-level labels. In the following, we adopt two smoothing techniques to suppress the noise in Y and reestimate Y as accurately as possible for image parsing.

We first smooth the initial labels of regions stored in Y by considering the relative size of regions. Let ρ_i be the ratio of the region $X_{i\cdot}$ occupying the image that $X_{i\cdot}$ belongs to, where $i = 1, ..., N$. We directly define the smoothed labels of regions $\tilde{Y} = \{\tilde{y}_{ij}\}_{N\times C}$ as follows:

$$\tilde{y}_{ij} = \rho_i y_{ij}, \tag{3}$$

which means that a larger region is more important for our cross-modal data refinement. More notably, such smoothing technique can suppress the negative effect of tiny regions produced by oversegmentation.

We further refine the smoothed $\tilde{Y}$ by exploiting the semantic context of object categories. In the task of image parsing, some object categories may be semantically correlated, e.g., "water" and "sky" are at a high probability to occur in the same image while "water" and "book" are less possible to occur together. In fact, such semantic context can be defined using a single matrix $S = \{s_{jj'}\}_{C\times C}$ based on the Pearson product moment correlation. Let the image-level labels of M images be collected as $Z = \{z_{ij}\}_{M\times C}$, where $z_{ij} = 1$ if image i is labeled with category j and $z_{ij} = 0$ otherwise. We then define $S = \{s_{jj'}\}_{C\times C}$ by:

$$s_{jj'} = \frac{\sum_{i=1}^{M}(z_{ij} - \mu_j)(z_{ij'} - \mu_{j'})}{(M-1)\sigma_j\sigma_{j'}}, \tag{4}$$

where μ_j and σ_j are the mean and standard deviation of column j of Z, respectively. By directly using the label propagation technique [Zhou *et al.*, 2004], we further smooth $\tilde{Y}$ with the semantic context as follows:

$$\bar{Y} = \tilde{Y}(I - \alpha_c D_c^{-1/2} S D_c^{-1/2})^{-1}, \tag{5}$$

where α_c is a positive parameter to control the strength of smoothing, and D_c is a diagonal matrix with its j-th diagonal element being $\sum_{j'} s_{jj'}$. In this paper, we directly set $\alpha_c = 0.04$ in all the experiments.

When we have obtained the final smoothed $\bar{Y}$ using the above two smoothing techniques, we reformulate our cross-modal data refinement problem as:

$$\min_{\hat{Y}\geq 0, \hat{X}, W} \frac{1}{2}||\hat{Y} - \hat{X}W||_F^2 + \frac{\lambda}{2}\mathrm{tr}(W^T\hat{X}^T L\hat{X}W) + \gamma||\hat{Y} - \bar{Y}||_1, \tag{6}$$

In the following, we will develop an efficient algorithm to solve the above problem for image parsing.

2.3 Efficient CMDR Algorithm

In fact, the CMDR problem in Eq. (6) can be solved in two alternate optimization steps as follows:

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

$$\hat{Y}^* = \arg\min_{\hat{Y}\geq 0} \frac{1}{2}||\hat{Y} - \hat{X}^*W^*||_F^2 + \gamma||\hat{Y} - \bar{Y}||_1,$$

where $\mathrm{LAP}(W, \hat{X}) = \mathrm{tr}(W^T\hat{X}^T L\hat{X}W)$. We set $\hat{X}^* = X$ and $\hat{Y}^* = \bar{Y}$ initially. As a basic L_1-norm optimization problem, the second subproblem has an explicit solution:

$$\hat{Y}^* = \mathrm{soft\_thr}(\hat{X}^*W^*, \bar{Y}, \gamma), \tag{7}$$

where $\mathrm{soft\_thr}(\cdot, \cdot, \gamma)$ is a soft-thresholding function. Here, we directly define $z = \mathrm{soft\_thr}(x, y, \gamma)$ as:

$$z = \begin{cases} z_1 = \max(x - \gamma, y), & f_1 \leq f_2 \\ z_2 = \max(0, \min(x + \gamma, y)), & f_1 > f_2 \end{cases},$$

where $f_1 = (z_1 - x)^2 + 2\gamma|z_1 - y|$ and $f_2 = (z_2 - x)^2 + 2\gamma|z_2 - y|$. In the following, we focus on efficiently solving the first quadratic optimization subproblem.

Let $\mathcal{Q}(W, \hat{X}) = \frac{1}{2}||\hat{Y}^* - \hat{X}W||_F^2 + \frac{\lambda}{2}\mathrm{LAP}(W, \hat{X})$. 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 $\hat{X}^*$, $\min_W \mathcal{Q}(W, \hat{X}^*)$ can be solved by setting the gradients of $\mathcal{Q}(W, \hat{X}^*)$ to zeros:

$$((\hat{X}^*)^T(I + \lambda L)\hat{X}^*)W = (\hat{X}^*)^T\hat{Y}^*. \tag{8}$$

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

Updating $\hat{X}$: When W is fixed at W^*, $\min_{\hat{X}} \mathcal{Q}(W^*, \hat{X})$ can be solved by setting the gradients of $\mathcal{Q}(W^*, \hat{X})$ to zeros:

$$(I + \lambda L)\hat{X}W^*W^{*T} = \hat{Y}^*W^{*T}. \tag{9}$$

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

$$F^*(\hat{X}) = (I + \lambda L)^{-1}\hat{Y}^*W^{*T}. \tag{10}$$

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:

$$\hat{X}(W^* W^{*T}) = F^*(\hat{X}). \quad (11)$$

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

The complete CMDR algorithm is outlined as follows:

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

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

(3) Initialize $\hat{X}^* = X$ and $\hat{Y}^* = \bar{Y}$;

(4) Find the best solution W^* by solving $((\hat{X}^*)^T (I + \frac{\alpha}{1-\alpha} L)\hat{X}^*)W = (\hat{X}^*)^T \hat{Y}^*$, which is exactly Eq. (8) with $\alpha = \lambda/(1+\lambda) \in (0, 1)$;

(5) Iterate $F_{t+1}(\hat{X}) = \alpha(I - L)F_t(\hat{X}) + (1 - \alpha)\hat{Y}^* W^{*T}$ until convergence, where a solution can thus be found just the same as Eq. (10) with $\alpha = \lambda/(1+\lambda)$;

(6) Find the best solution $\hat{X}^*$ by solving Eq. (11): $\hat{X}(W^* W^{*T}) = F^*(\hat{X})$, 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 $\hat{Y}^* = \text{soft\_thr}(\hat{X}^* W^*, \bar{Y}, \gamma)$;

(8) Iterate Steps (4)–(7) until the stopping condition is satisfied, and output the final parsing results $\hat{Y}^*$.

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

3 Social Image Parsing

In the previous section, we assume that the large set of regions have been provided in advance. In the following, we discuss how to generate this input for our CMDR algorithm.

Given a set of images, we adopt the Blobworld method [Carson *et al.*, 2002] for oversegmentation. Concretely, we first 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. The pixels within this image are then grouped into regions, and the number of regions is automatically detected by a model selection principle. To ensure an oversegmentation of each image, we slightly modify the original Blobworld method in two ways: 1) the number of regions is initially set to a large value; 2) model selection is forced to be less important during segmentation.

After we have oversegmented all the images into regions, we collect them into a single matrix $X \in R^{N \times d}$. Here, each region $X_{i\cdot}$ is denoted as a 137-dimensional feature vector by concatenating color and texture features: three mean color features with their standard deviations (6-dimensional), three mean texture features with their standard deviations (6-dimensional), and color histogram (125-dimensional). Finally, we compute a Gaussian kernel over X for our cross-modal data refinement algorithm.

The full algorithm for social image parsing is as follows:

(1) Oversegment each image into multiple regions using the modified Blobworld method;

(2) Extract a 137-dimensional feature vector from each region by concatenating color and texture features;

(3) Perform image parsing by running our CMDR algorithm over the large set of regions.

4 Experimental Results

In this section, our CMDR algorithm for image parsing is evaluated by conducting two groups of experiments: test with noisily tagged images and test with social images.

4.1 Test with Noisily Tagged Images

Experimental Setup

We select two benchmark datasets for performance evaluation: MSRC [Shotton *et al.*, 2006] and LabelMe [Liu *et al.*, 2011]. The MSRC dataset contains 591 images, accompanied with a hand labeled object segmentation of 21 object categories. Pixels on the boundaries of objects are usually labeled as background and not taken into account in these segmentations. The LabelMe dataset (also called as the SIFT Flow dataset) contains 2,688 outdoor images, densely labeled with 33 object categories using the LabelMe online annotation tool. For these two benchmark datasets, we oversegment each image into multiple regions and then totally obtain about 7,500 regions and 33,000 regions, respectively.

To study whether our CMDR algorithm can deal with noisily tagged images, we add random noise into the image-level labels that are initially provided for image parsing. More concretely, we randomly select certain percent of images and then attach each selected image with one extra wrong label. This means that the noise strength is determined by the percent of noisily tagged images (see Table 1) in the following experiments. It should be noted that *the noise level is actually high*, although we only add one wrong label for each selected image. That is, since most of the images used in our experiments are segmented into three or four objects in the ground-truth segmentations, one wrong label induces 1/5 to 1/4 noise within an image. More notably, we will also make evaluation with realistic noise in the next subsection.

To verify the effectiveness of our cross-modal data refinement for image parsing, we first consider a baseline method that is a variant of our CMDR algorithm without optimizing $\hat{X}$ in Eq. (6) (thus denoted as DR). Moreover, we compare our CMDR algorithm with two representative methods [Tang *et al.*, 2009; Chen *et al.*, 2011] for data refinement by

Table 1: Average accuracies (%) of different image parsing methods on the two benchmark datasets. The standard deviations are also provided along with average accuracies.

Datasets	Methods	Noisily tagged images				
		0%	25%	50%	75%	100%
MSRC	CMDR (ours)	**74**	**70**±1	**66**±1	**63**±2	**59**±2
	DR (ours)	61	57±1	54±1	51±2	48±2
	[Tang *et al.*, 2009]	67	61±1	59±1	53±1	51±1
	[Chen *et al.*, 2011]	57	53±1	48±1	46±1	44±1
	[Liu *et al.*, 2013]	70	59±2	52±3	44±3	38±3
LabelMe	CMDR (ours)	**29**	**28**±1	**26**±1	**24**±2	**23**±1
	DR (ours)	18	16±1	15±1	14±1	13±1
	[Tang *et al.*, 2009]	22	20±1	17±1	16±1	15±1
	[Chen *et al.*, 2011]	18	17±1	16±1	15±1	15±1
	[Liu *et al.*, 2013]	26	24±5	18±4	14±2	13±2

sparse coding. Finally, we make direct comparison to the recent work [Liu *et al.*, 2013][1] on image parsing by inputting noisily tagged images into the algorithm developed in [Liu *et al.*, 2013]. For fair comparison, we make use of the same initial label smoothing techniques proposed in Section 2.2 for all the methods compared here. We evaluate the parsing results on a subset of images equivalent to the test set used in [Vezhnevets *et al.*, 2011; Liu *et al.*, 2013]. The accuracy is computed by comparing the parsing results to the ground truth segmentations for each category and then averaged over all the categories. Each trial is randomly repeated 25 times.

It should be noted that *the ground-truth pixel-level labels of all the images are unknown* in our setting for image parsing. Hence, it is not possible to select the parameters by cross-validation for our CMDR algorithm. In this paper, we thus uniformly set the parameters of our CMDR algorithm as $k = 110$, $\alpha = 0.45$ (equally $\lambda = 0.82$), and $\gamma = 0.12$ for the two benchmark datasets. Moreover, we construct k-NN graphs over all the regions for other related methods to speed up image parsing. The parameters of these methods are also set to their respective optimal values.

Parsing Results

We compare our CMDR algorithm for image parsing with other related methods when different percents of noisily tagged images are provided initially. The comparison results on the two benchmark datasets are listed in Table 1. The immediate observation is that our CMDR algorithm achieves the best results in all cases (see example results in Figure 3). This means that our CMDR algorithm is more effective for noisily tagged image parsing than the two data refinement methods [Tang *et al.*, 2009; Chen *et al.*, 2011] based on sparse coding. The comparison of CMDR vs. DR shows that the "cross-modal" idea plays an important role in our algorithm for image parsing. More notably, our CMDR algorithm is shown to obviously outperform the recent work [Liu *et al.*, 2013], mainly due to the fact that extra requirements are imposed on the initial image-level labels in [Liu *et al.*, 2013] while our problem formulation is much more flexible.

[1]Originally developed for weakly supervised setting, but applied to noisily tagged setting here for fair comparison.

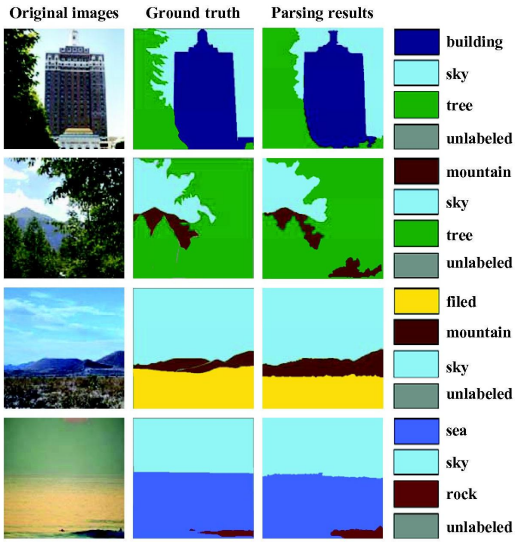

Figure 3: Example parsing results (in the third column) by our CMDR algorithm in comparison with the ground-truth (in the second column) on the LabelMe dataset.

Table 2: Overview of the state-of-the-art results in different image parsing settings on the MSRC dataset.

Settings	Methods	Avg accuracies
FS	[Shotton *et al.*, 2008]	67
	[Ladicky *et al.*, 2009]	75
	[Csurka and Perronnin, 2011]	64
	[Lucchi *et al.*, 2012]	**76**
WS	[Verbeek and Triggs, 2007]	50
	[Vezhnevets and Buhmann, 2010]	37
	[Vezhnevets *et al.*, 2011]	67
	[Zhang *et al.*, 2013]	**69**
	[Akbas and Ahuja, 2014]	62
NT	[Liu *et al.*, 2013] (0% noise)	70
	[Liu *et al.*, 2013] (100% noise)	38
	CMDR (0% noise)	**74**
	CMDR (100% noise)	59

We further give an overview of the parsing results obtained by our CMDR algorithm and the state-of-the-art results on the two benchmark datasets in Tables 2 and 3, respectively. Here, three settings for image parsing are considered: FS–fully supervised image parsing using pixel-level labels, WS–weakly supervised image parsing using clean image-level labels, and NT–noisily tagged image parsing using noisy image-level labels. Strictly speaking, the three settings for image parsing cannot be compared directly. Here, we mainly want to give an overview of them in Tables 2 and 3. In fact, from such overview, we find that our CMDR algorithm can provide an *alternative and natural* way to address the challenging problem of image parsing. That is, when the image-level labels (maybe noisy) are easy to access in practice, we are able to achieve promising results by our CMDR algorithm, without the need to directly collect pixel-level labels as supervisory information at too expensive cost. However, this conclusion

Table 3: Overview of the state-of-the-art results in different image parsing settings on the LabelMe dataset.

Settings	Methods	Avg accuracies
FS	[Liu *et al.*, 2011]	24
	[Tighe and Lazebnik, 2010]	29
	[Myeong *et al.*, 2012]	**32**
	[Tighe and Lazebnik, 2013]	30
WS	[Vezhnevets *et al.*, 2012]	**21**
NT	[Liu *et al.*, 2013] (0% noise)	26
	[Liu *et al.*, 2013] (100% noise)	13
	Our CMDR (0% noise)	**29**
	Our CMDR (100% noise)	23

does not hold for the recent work on image parsing [Liu *et al.*, 2013], which means that our noisily tagged image parsing is indeed a very challenging problem and the algorithm design in this setting is very important.

Besides the above advantages, our CMDR algorithm has another distinct advantage, i.e., it runs efficiently on a large set of regions. For example, the running time of data refinement over $\bar{Y}$ taken by our CMDR algorithm, [Tang *et al.*, 2009], [Chen *et al.*, 2011], and [Liu *et al.*, 2013] on the MSRC dataset ($N \approx 7,500$) is 23, 63, 30, and 68 seconds, respectively. We run all the algorithms (Matlab code) on a computer with 3.9GHz CPU and 32GB RAM. It can be clearly observed that our CMDR algorithm runs the fastest among the four related methods for image parsing.

4.2 Test with Social Images

Experimental Setup

In this paper, we actually derive a Flickr dataset with realistic noise from the PASCAL VOC2007 benchmark dataset [Everingham *et al.*, 2007]. The original VOC2007 dataset contains 632 images, well segmented with 20 object categories. Since images in the VOC2007 dataset are originally downloaded from the Flickr website, we choose to construct a Flickr dataset based on this benchmark dataset. Concretely, we directly copy all the images from the VOC2007 dataset and then collect the image-level labels from the Flickr website, instead of the original clean image-level labels from the VOC2007 dataset. In our experiments, we only keep 100 object categories that most frequently occur in this Flickr dataset. Moreover, the ground-truth pixel-level labels derived from the original VOC2007 dataset are used for performance evaluation. To this end, we only consider the 20 object categories that occur in the ground-truth segmentations in the evaluation step (while 100 object categories are still considered in the other steps of image parsing).

To verify the effectiveness of our CMDR algorithm for social image parsing with realistic noise, we compare it with several closely related methods [Tang *et al.*, 2009; Chen *et al.*, 2011; Liu *et al.*, 2013]. Here, we just take the same strategy of parameter selection and performance evaluation as that on the MSRC and LabelMe datasets. In fact, the experimental setup of this subsection is the same as that of the previous subsection except that we consider the realistic noise in image-level labels instead of random noise.

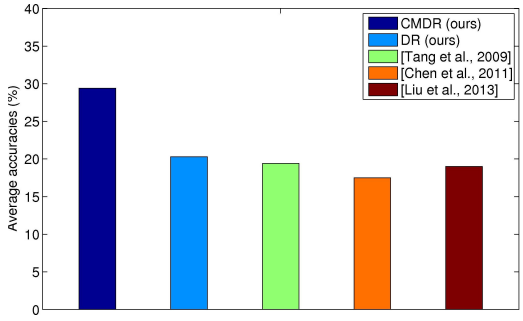

Figure 4: Comparison of our CMDR algorithm to related methods for social image parsing with realistic noise.

Parsing Results

We show the comparison of our CMDR algorithm to related methods in Figure 4. Overall, we can make the same observation on this new dataset as we have done with the MSRC and LabelMe datasets in the previous subsection. That is, our CMDR algorithm is still shown to achieve the best results in social image parsing with realistic noise. This means that our CMDR algorithm is more effective for data refinement over the initial labels of regions even when the noisy image-level labels are directly collected from the Flickr website. In fact, our WSSL algorithm can be considered to provide an *alternative and natural* way to address the challenging problem of image parsing, since the noisy image-level labels collected from the Flickr website are much easier to access than pixel-level labels (used in fully supervised setting). In fact, this is also the place where we stand in the present work.

5 Conclusions

In this paper, we have investigated the challenging problem of social image parsing. From the viewpoint of data refinement over the labels of regions, we have formulated social image parsing as a cross-modal data refinement problem. Based on L_1-optimization and label propagation techniques, we have further developed an efficient algorithm to solve such cross-modal data refinement problem. The experimental results have demonstrated the effectiveness of our cross-modal data refinement algorithm for image parsing. In the future work, we will extend our algorithm to other challenging tasks in computer vision for the need of data refinement.

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

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

[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. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

[Chang *et al.*, 2014] F.-J. Chang, Y.-Y. Lin, and K.-J. Hsu. Multiple structured-instance learning for semantic segmentation with uncertain training data. In *Proc. CVPR*, pages 360–367, 2014.

[Chen *et al.*, 2011] X. Chen, Q. Lin, S. Kim, J.G. Carbonell, and E.P. Xing. Smoothing proximal gradient method for general structured sparse learning. In *Proc. UAI*, pages 105–114, 2011.

[Cheng *et al.*, 2014] M.-M. Cheng, S. Zheng, W.-Y. Lin, V. Vineet, P. Sturgess, N. Crook, N. Mitra, and P. Torr. Imagespirit: Verbal guided image parsing. *ACM Trans. Graphics*, (In Press), 2014.

[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.

[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/, 2007.

[Figueiredo *et al.*, 2007] M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[Fu *et al.*, 2011] Z. Fu, Z. Lu, H. Ip, Y. Peng, and H. Lu. Symmetric graph regularized constraint propagation. In *Proc. AAAI*, pages 350–355, 2011.

[Kohli *et al.*, 2009] P. Kohli, L. Ladicky, and P.H.S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

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

[Ladicky *et al.*, 2010] L. Ladicky, C. Russell, P. Kohli, and P.H.S. Torr. Graph cut based inference with co-occurrence statistics. In *Proc. ECCV*, pages 239–253, 2010.

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

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

[Liu *et al.*, 2014] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Trans. Multimedia*, 16(1):253–265, 2014.

[Lucchi *et al.*, 2012] A. Lucchi, Y. Li, K. Smith, and P. Fua. Structured image segmentation using kernelized features. In *Proc. 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] Heesoo Myeong, Ju Yong Chang, and Kyoung Mu Lee. Learning object relationships via graph-based context model. In *Proc. CVPR*, pages 2727–2734, 2012.

[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 *Proc. ECCV*, pages 1–15, 2006.

[Shotton *et al.*, 2008] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. CVPR*, pages 1–8, 2008.

[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 *Proc. ACM Multimedia*, pages 223–232, 2009.

[Tighe and Lazebnik, 2010] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *Proc. 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 *Proc. CVPR*, pages 1–8, 2007.

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

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

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

[Wright *et al.*, 2009] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 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 *Proc. CVPR*, pages 3190–3197, 2014.

[Yang *et al.*, 2007] L. Yang, P. Meer, and D.J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *Proc. CVPR*, pages 1–8, 2007.

[Yang *et al.*, 2014] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *Proc. CVPR*, pages 3294–3301, 2014.

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

[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 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.