Neurocomputing 201 (2016) 92–103

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Automatic non-parametric image parsing via hierarchical semantic voting based on sparse–dense reconstruction and spatial–contextual cues

Xinyi An [a], Shuai Li [a,*], Hong Qin [b], Aimin Hao [a,*]

[a] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[b] Stony Brook University, United States

ARTICLE INFO

Article history:
Received 11 September 2015
Received in revised form
1 December 2015
Accepted 17 March 2016
Communicated by Deng Cheng
Available online 20 April 2016

Keywords:
Non-parametric image parsing
Sparse–dense reconstruction
Hierarchical semantic voting
String with semantic spatial–contextual cue

ABSTRACT

Image parsing is vital for many high-level image understanding tasks. Although both parametric and non-parametric approaches have achieved remarkable success, many technical challenges still prevail for images containing things/objects with broad-coverage and high-variability, because it still lacks versatile and effective strategies to seamlessly integrate local–global features selection, contextual cues exploitation, spatial layout encoding, data-driven coherency exploration, and flexible accommodation of newly annotated labels. To ameliorate, this paper develops a novel automatic non-parametric image parsing method with advantages of both parametric and non-parametric methodologies by resorting to new modeling and inferring strategies. The originality of our new approach is to employ sparse–dense reconstruction as a latent learning model to conduct candidate-label probability analysis over multi-level local regions, and synchronously leverage context-specific local–global label confidence propagation and global semantic spatial–contextual cues to guide holistic scene parsing. Towards this goal, we devise several novel technical components to comprise a lightweight parsing framework, including local region representation integrating complementary features, anisotropic consistency propagation based on bi-harmonic distance metric, bottom-up label voting, semantic string generation of image-level spatial–contextual cues based on Hilbert space-filling curve, and co-occurrence priors analysis based on relaxed string matching algorithm, which collectively enable us to effectively combat the aforementioned obstinate problems. Moreover, we conduct comprehensive experiments on public benchmarks, and make extensive and quantitative evaluations with state-of-the-art methods, which demonstrate the advantages of our method in accuracy, versatility, flexibility, and efficiency.

1. Introduction and motivation

Image parsing aims to decompose an image into non-overlapped consistent regions that correspond to a set of pre-defined semantic classes. It is one of the most active research subjects in computer vision nowadays, and can benefit many high-level image understanding tasks, such as image editing, context-based image (or image patch) retrieval, etc. Although human's perceptual grouping ability could guarantee to quickly distinguish many patterns and abstract high-level information from them to form meaningful parts, it is fundamentally challenging to equip the computer with such ability due to the diversity of natural scenes and the variability of object class instances.

Given training images with region-level annotations, most state-of-the-art image parsing methods either learn a region-based parametric model by combining appearance and scene geometry representations [1–4] or resort to non-parametric modeling to transfer the annotations from training images to testing images [2,5,6]. Generally speaking, segmentation hypotheses, feature encoding of candidate segments, label transferring and spatial–contextual consistency have commonly become some of the key factors of effective image parsing. Through a long time evolution, both parametric and non-parametric approaches have achieved great success. However, to better combine the traditional problems of detection, segmentation, and multi-label recognition into a unified parsing framework, the challenges, ranging from feature representation, model design, to spatial–contextual priors leverage, are still not fully resolved, which collectively hinder further performance improvement towards human perception. Now, we shall summarize some of the key challenges as follows.

* Corresponding authors.
E-mail addresses: lishuai@buaa.edu.cn (S. Li), ham@buaa.edu.cn (A. Hao).

http://dx.doi.org/10.1016/j.neucom.2016.03.034
0925-2312/

First, from the perspective of producing good internal representation and exploration of visual information, gestalt psychologists suggest that hierarchically grouping from low-level features to high-level structures can better embody the concept of psychological recognition, including proximity, similarity, continuation, closure, symmetry, etc. [7,8].

However, the ambiguous semantic definition of what determining a local region to be an object (or a meaningful part) makes the hierarchical grouping become an ill-posed problem, because a meaningful region may refer to a thing, a kind of texture, a stuff, or even a part of an object. Therefore, how to simultaneously exploit the low-level features (such as color and texture) and local–global structures is still urgently needed in image parsing.

Second, from the perspective of intrinsic cross-image semantic consistency interpretation/detection, current methods more or less suffer from the following problems. The correspondence mapping or instance detection based on straightforward local features combination gives rise to less discriminative coherency propagation, because cross-image co-occurring contents may vary in shapes, colors, scales, illuminations, occlusions, and local deformations. And prior knowledge based learning/regression significantly depends on the quality and contexts of the training samples as well as the sophisticated parameter tuning of the underlying classifier and/or structural models, which lacks desirable efficiency, flexibility, and expandability. Thus, considering the semantics—similar but appearance—varying things/objects, how to design a physics-meaningful model to analyze the intrinsic correlations among the cross-image feature representations is extremely essential for the consistent label propagation.

Third, from the perspective of the effective utility of spatial–contextual information in co-occurrence interpretation, various kinds of high-level spatial layouts and contextual interactions among different object classes have been proven effective in semantic parsing [9–12], because the co-occurrence relations can impose constraints on the likelihood that some object classes occur simultaneously in the same scene. Although the contextual cues should be taken into account in a relatively easy way via segmentation-by-detection like methods, some spatial-layout cues become messy and unreliable when photographing 2D images from real 3D scenes with arbitrary viewpoint. Thus, considering the uncertainty of co-occurring object classes and their projected 2D position relationships, it needs a relatively independent but closely coupled strategy to flexibly encode and analyze the spatial–contextual cues, with respect to the prime spatial separation scheme.

Fourth, from the perspective of practical use, facing different application backgrounds, it is indeed hard to make choices among various method-design alternatives, e.g., parametric or non-parametric, hand-crafted features or learned features, intrinsic cues or extrinsic cues, bottom-up or top-down, data-driven or prior/rule-guided, etc. The ideal state would be to find a perfect or near perfect method that could take full use of their respective advantages. Nonetheless, at the current stage it is only a viable option to design an efficient and effective framework that could partially couple their advantages in some sense.

To tackle the aforementioned challenges, we shall concentrate on the automatic non-parametric image parsing by incorporating region-level local–global complementary feature integration, per-exemplar candidate-label detection based on sparse–dense reconstruction, hierarchical semantic voting, and the statistical analysis of global spatial–contextual cues into a flexible and expandable framework, which can take full advantages of both parametric and non-parametric methods. As illustrated in Fig. 1, given the annotated images, instead of sophisticated learning, we only integrate multiple semi-local features to form dictionary words for each of the multi-level segmented regions, and construct a sematic string for each image. When handling testing images, we first retrieve globally similar images from the annotated ones to construct latent learning dictionary, conduct multi-level segmentation, and extract analogous region-level features. Then, we employ per-exemplar sparse–dense reconstruction for each region and determine the high-level region's label candidates via hierarchical voting. In addition, we construct image-level semantic string candidates for testing images, and finally, determine image-level parsing by taking into account the statistics of cross-image semantic-string matching results. Specifically, our salient contributions can be summarized as follows:

- We pioneer a hierarchical sparse–dense reconstruction based semantic region voting method, which in some sense is equivalent to the role of the structural models in parametric methods. It gives rise to the efficient and effective revealing of the intrinsic semantic consistence among training and testing images, while still being able to flexibly accommodate newly added annotated images.
- We propose a semi-local discriminative representation by combining bi-harmonic distance distribution (BDD) with low-level features based on multi-level super-pixel segmentation, which can capture both region-level local appearances and global geometric structures of the potential semantic entities in an image. Specifically, the BDD could also facilitate anisotropic label-confidence propagation.
- We propose an efficient semantic-string encoding and matching method to represent the image-level spatial–contextual cues based on Hilbert space-filling curve, wherein the high-level spatial layout and contextual co-occurrence are closely coupled in a smart way, and the relaxed string matching algorithm guarantees the effective exploration of the priors embedded in diverse scenes.

2. Related work

Closely relevant to the central theme of this paper, we now briefly review previous works in three subjects: parametric image parsing methods, non-parametric image parsing methods, and the exploitation of features and spatial–contextual cues.

2.1. Parametric image parsing methods

Parametric methods [13–15] usually leverage semantic learning to establish certain appearance representation or relationship representation from training dataset. With the hope that such a learning would facilitate the mapping from visual features to a semantically meaningful space, most of the state-of-the-art parametric methods usually propagate labels from annotated pixels to testing pixels by jointly considering the appearance and structure features based on manually designed structural models [2], such as Markov random field (MRF) and conditional random fields (CRF) based models [16]. For example, Yuan et al. [17] employed CRF and max-margin Markov networks (M3N) to perform scene understanding. Tu et al. [18] proposed a Bayesian framework to parse images into visual patterns. Besides, inspired by the deep Recursive Context Propagation Network, Abhishek et al. [3] proposed a learning-based scene parsing method by combining bottom-up and top-down context propagation within random binary parse trees, which makes the region-level feature representation be better classified into semantic categories. Modolo et al. [19] proposed a context forest to learn the relationship between the global image appearance and properties of objects by selecting the most relevant components to run on testing images. Humayun et al. [20] proposed to reuse the inference by pre-computing a graph for parametric

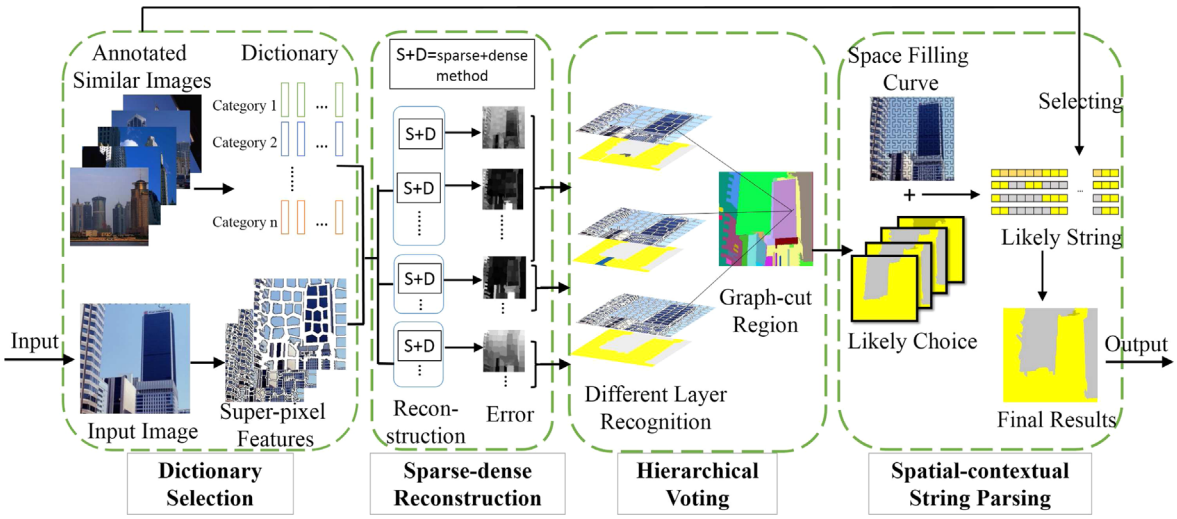


Fig. 1. The pipeline of our framework.

min-cuts. Most recently, convolutional neural network (CNN) based methods [21,22,4] gain great momentum in end-to-end image parsing and achieve high-accuracy, however, the involved deep learning models are extremely time-consuming and are hard to incorporate new semantic labels. In general, the performance of current parametric methods is commonly limited by the suboptimal performance of many sophisticated intermediate components. Besides, the number of object labels in such parsing models is limited, yet in contrast, the number of objects is actually unlimited in real world, combined with the randomly distributed objects to form a scene, so it is hard to build a complete parametric model, because it has to re-train/update the models when encountering newly added training dataset.

2.2. Non-parametric image parsing methods

Unlike parametric methods, non-parametric methods [5] almost do nothing at training time. At testing time, they commonly employ certain global image descriptors to identify a retrieval set of training images similar to the query, with the testing image being segmented into regions, then they transfer the labels from the retrieval set by matching segmentation regions, and finally fuse the transferred labels by heuristic aggregation schemes. For example, Liu et al. [23] employed SIFT Flows to construct the pixel–pixel correspondence and the dense deformation field between images, wherein the involved optimization problem is rather complex and expensive to be solved. David et al. [24] further improved Tighe and Lazebnik's work [5] by introducing per-descriptor weights and context-driven adaptation, which gives rise to a significant performance gain on sufficiently large datasets. Long et al. [6] proposed a hierarchical image model (HIM) to conduct parsing by segmentation and recognition, which can capture long-range dependency and different-level contextual information. And Tighe et al. [25] also proposed a parsing-by-detecting approach, wherein exemplar SVM is employed to estimate the bounding boxes of the objects and to facilitate the parsing labels' transfer. Cheng [26] formulated the image parsing as a problem of jointly estimating per-pixel object and attribute labels from training images by treating nouns as object labels and adjectives as visual attribute labels, however, such performance is limited by the ambiguity of language description. Similar to the weakly supervised graph propagation method [27], Wenxuan et al. [10] proposed to propagate class labels to image regions by only using image-level labels with the help of L_1 semantic graph and *k-NN* semantic graph, which gives rise to more semantic relevance. Generally speaking, benefitting from the flexibility of affording new dataset, non-parametric methods are more popular in practical applications, and theoretically they could utilize a huge amount of data with ever-improvement in accuracy. However, their practical parsing quality also tends to be influenced by the inaccurate matching and insufficient explicit semantic priors [3]. Inspired by the quasi-parametric method [28], which integrates KNN-based nonparametric method and CNN-based parametric method, this paper will focus on the new automatic non-parametric image parsing framework towards leveraging the advantages of both parametric and non-parametric methodologies.

2.3. Features and spatial–contextual cues

Image parsing methods are usually built on top of the feature representation and spatial–contextual cues exploration. Discriminative representations of object classes based on low-level image features are important for semantic parsing, wherein the commonly used features are bottom-up pixel-level features, such as color or texture patterns. For example, Long et al. [29] documented the better performance of convolutional activation features over traditional features (such as SIFT) for the involved correspondence calculation. And He et al. [30] suggested to simultaneously incorporate the region and image level features. Shotton et al. [31] employed spatial filters to represent the local information corresponding to different classes. Shotton et al. [11] learned a discriminative object-label model by incorporating appearance, shape, and context information efficiently. Instead of hand-crafted features, Farabet et al. [32] resorted to learning appropriate low-level and mid-level features via supervised training. And Zhu et al. [1] made a comprehensive survey on the feature selection involved in various kinds of image segmentation methods (refer to [1] for more details). Besides, various forms of co-occurrence, spatial adjacency and appearance have been proposed to serve as spatial–contextual cues in [5]. For example, Zhou et al. [9] decompose an image into four parts (the upper part, the lower part, the center, and the sides) to construct spatial layout relationships. Modolo et al. [19] modeled the context as a relationship between global image appearance and the properties of the objects within the same image. Vulee et al. [33] introduced the mutual spatial feature to obtain strong visual cue in image parsing. Tighe et al. [25] introduced region-based cues by transferring the trained segmentation mask into the testing image to form a segmentation hypothesis. Inspired by the above works, to make the image parsing become more holistic, this paper will also exploit the closely coupled and flexibly encoded spatial–contextual priors

by taking full consideration of the intrinsic cues and extrinsic cues simultaneously.

3. Semi-local feature construction and latent-learning dictionary selection

Since parsing images aims to transfer labels of training images to the unknown image, we should make full use of annotated descriptions to establish relationships with unknown parts, so we deal with training images and testing images respectively. For the training images, their pixels have been respectively annotated with specific labels, which gives rise to object-level region segmentation. Correspondingly, for the testing image, we employ an automatic graph-based method [34] to roughly conduct potentially semantic region segmentation, which will be finally refined and assigned to most likely labels. Meanwhile, both for the annotated image and testing image, we further employ SLIC [35] based over-segmentation to construct a spatial hierarchy for each potentially semantic region.

On that basis, for each-level local region, the description is an essential part for recognition. Different local regions should have the ability to distinguish themselves from others. However, regions from the same category may also have differences in certain aspects, so we should design our descriptions from different perspectives to be able to capture common properties of regions that have the same semantic meanings. Traditional descriptions like SIFT, and color histogram are good enough to represent local properties from different aspects, but that is not enough because they do not consider neighborhoods' influence. Entities in our daily life may look different from the local perspective, but we can regard them as the same category because they have something in common when considering adjacent parts. So we shall integrate not only multiple complementary local features, but also geometric distribution structure to semi-locally represent regions. Except for the low-level appearance features such as color histogram, local binary pattern, and histogram of gradient texture, we specifically propose a new semi-local structure representation, called bi-harmonic distance distribution (BDD), by computing bi-harmonic distance field over super-pixels, which shall be detailed as follows.

3.1. Semi-local region representation based on bi-harmonic distance distribution

In fact, bi-harmonic distance [36] is a kind of distance metric built on Riemannian manifold. Compared with traditional Euclidean distance, bi-harmonic distance has the ability of capturing both local and global information. Therefore, we extend it to image space to describe the geometric distribution structure, and we will also employ it to guide the anisotropic reconstruction error propagation in Section 4.3. The original bi-harmonic distance between the two positions x and y of certain manifold is defined as

$$d_B(x,y)^2 = \sum_{k=1}^{\infty} \frac{(\Phi_k(x)-\Phi_k(y))^2}{\lambda_k^2}, \tag{1}$$

where $\Phi_k(x)$ are eigenfunctions and λ_k are eigenvalues of Laplace–Beltrami matrix.

In order to measure the local image structure, based on the super-pixel over-segmentation, we use Delaunay triangle to construct a manifold mesh, wherein the super-pixel location centers $P=\{p_1,p_2,\ldots,p_K\}$ serve as vertices, K is the number of super-pixels. And we define bi-harmonic distance metric based on discrete Laplacian-matrix $L=A^{-1}M$, where A is a diagonal matrix and A_{ii} is proportional to the average area of the triangles sharing vertex p_i. And M is formulated as

$$M_{ij}=\begin{cases} \sum_k m_{i,j} & \text{if } i=j \\ -m_{ij} & \text{if } p_i \text{ and } p_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Here $m_{ij}=\cot\alpha_{ij}+\cot\beta_{ij}$, α_{ij} and β_{ij} are the opposite angles of two adjacent triangles sharing edge p_ip_j.

Since the lightness information is contained in all the three channels of RGB color space, such channels have strong correlations. However, LAB color space can separate lightness from colors, which is more appropriate in explaining the human's vision. Therefore, we use $|l_i-l_j|+|a_i-a_j|+|b_i-b_j|$ to calculate the color distance, which is used as the third dimension of the 3D coordinates for each vertex, so that the color components are embedded in the edge length calculation based on LAB color space. Here l, a, and b denote the average color value of the super-pixel p in LAB color space. And Fig. 2 shows the constructed Delaunay triangles and the bi-harmonic distance fields corresponding to different anchor super-pixels.

Since bi-harmonic distance can measure the differences in color and location between super-pixels, by calculating its distribution in the neighborhood, we can obtain semi-local geometric structures of each super-pixel. Given super-pixel sp_i and its corresponding bi-harmonic distances to other super-pixels $\{d_B(i,1),d_B(i,2),\ldots,d_B(i,K)\}$, we define $h(d_a,d_b)$ as the probability of such bi-harmonic distance set belonging to the range between d_a and d_b. Thus, $h(d_a,d_b)$ can be computed as

$$h(d_a,d_b)=\frac{\sum_{k=1}^{K}\delta(d_a\le d_B(i,k)<d_b)}{K}, \tag{3}$$

where we define $\delta(equation)=1$ when *equation* is true, otherwise, $\delta(equation)=0$. Thus, the l-dimension BDD histogram H_i of super-pixel i can be represented as

$$H_i=\left[h\left(0,\frac{1}{l}\right),h\left(\frac{1}{l},\frac{2}{l}\right),\ldots,h\left(\frac{l-1}{l},1\right)\right], \tag{4}$$

where l represents the dimension used for BDD histogram.

In this paper, we mainly care about the semi-local structures of each super-pixel, so we only take into account super-pixels with the three-ring neighborhood when calculating BDD. Therefore, by concatenating other low-level local features with BDD, we now have a new semi-local complementary feature, which gives rise to the informative description of the local appearance and global geometric structures of the potentially semantic regions.

3.2. Image-specific latent-learning dictionary selection

So far, the constructed semi-local representations of the local regions in annotated images can serve as candidate dictionary words to form per-exemplar dictionaries. Given a new testing image, similar to other data-driven methods, in order to efficiently find effective dictionary words from annotated images, we assume that globally similar images may have more intrinsically consistent information, i.e., the features involved in such images are more likely to fall into the same subspace. Thus, we retrieve such images based on global features, including spatial pyramid, gist and color histogram as in [5]. Then we take Euclidean distance as measurement of each feature's ranking, by averaging the ranks of each global descriptor, we can obtain the final rank of the annotated images. Here we set M as the number of the chosen annotated images. As a result, the combination of different kinds of global features afford complementary descriptive power.

Since the subset of annotated images is obtained by comparing their global-view similarity with respect to the testing image, in some sense, such process has roughly considered the appearance

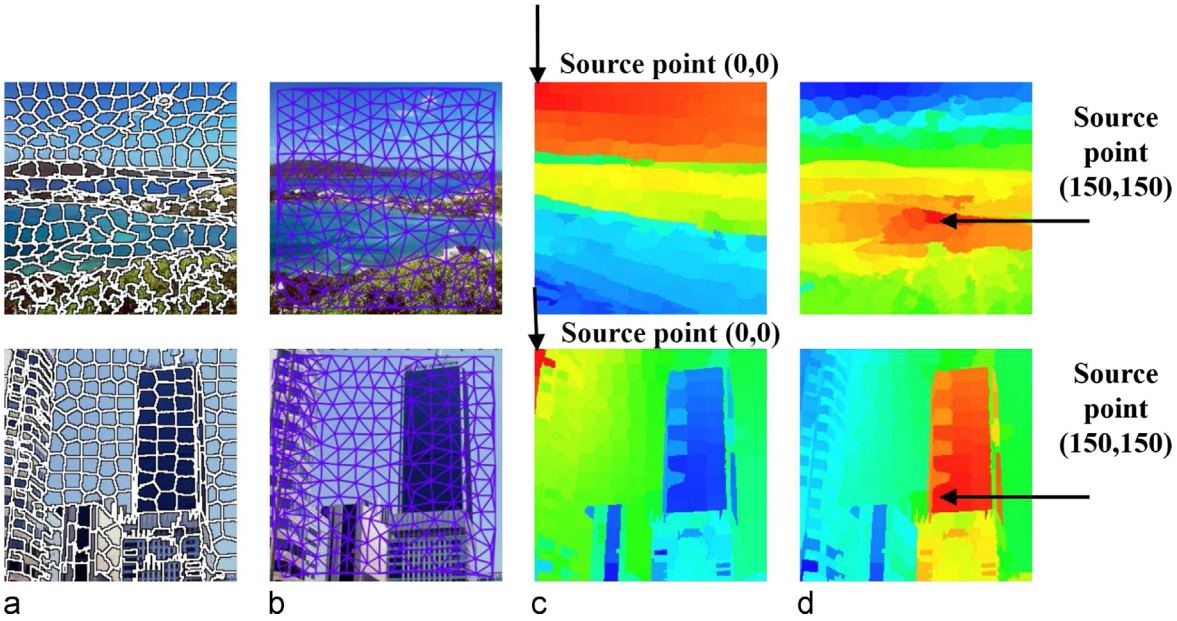


Fig. 2. Demonstration of the bi-harmonic distance field construction. (a) Super-pixel over-segmentation. (b) Delaunay triangle construction based on super-pixels. (c) Bi-harmonic distance distribution with respect to source point (0, 0). (d) Bi-harmonic distance distribution with respect to source point (150,150).

contextual cues, and they should contribute more than other images when selecting dictionary words. After choosing M most similar images, we construct per-exemplar dictionaries $D=\{D_1, D_2, \ldots D_N\}$ from those images, where D_i indicates the dictionary of the i-th category, N is the total number of the pre-defined object category. And each dictionary is represented as a concatenation of words expressed as $D_i=[d_i^1, d_i^2, \ldots, d_i^M]$, where d_i^j is the word selected from the j-th image for the dictionary of the i-th category. The number of the words in each dictionary is mainly determined by the corresponding words appearing in those similar images. However, considering that a large number of words may influence efficiency, we can also limit the max number of the words in each dictionary to balance the tradeoff between accuracy and efficiency.

Such latent-learning dictionary enables us to select useful annotated image features efficiently, especially when having a large number of candidate annotated images. On that basis, our image-specific dictionary selection method also gives rise to another advantage, that is, our method has more flexibility to accommodate new complementary annotated labels while requiring no tedious re-training or updating process.

4. Region-label candidates generation based on sparse–dense reconstruction

According to the semi-local representations obtained from the testing image and corresponding per-exemplar dictionary, we explore their intrinsic consistency by using per-exemplar dictionary to represent the tested local region in each level within the hierarchy. Since regions from the same category may have something in common, the corresponding features must have some relationships. Here, we let per-exemplar dictionary represent unknown features. If a region can be reconstructed with minimal error, we assume that it belongs to corresponding object class with larger probability. After the reconstruction process, each region in each level within the hierarchy should be assigned with a few candidate category labels according to the reconstruction accuracy. To make such reconstruction process more reliable, we use two kinds of reconstruction method based on different models which have respective advantages. Sparse reconstruction can generate unique and compact representation but gives rise to less stableness, while dense reconstruction can reveal more expressive and generic properties but is more sensitive to outliers. Therefore, we resort to the utility of both sparse and dense reconstruction errors to measure the semantic probability of each region, which are detailed as follows.

4.1. Sparse reconstruction

The goal of sparse reconstruction is to compute each category's probability on specific over-segmented super-pixels, and find the most likely category to form each hierarchical basis. We use such method to find intrinsically consistent relationships based on the assumption that features of the same category can be linearly constructed by a few words in corresponding dictionary D_i while other category dictionary cannot, so the reconstruction residue is the measurement of similarity between feature and per-exemplar dictionary. Our main idea is to conduct sparse representation and use the reconstruction error to judge unknown representation.

Since sparse coding can effectively reveal the relationship among similar features, we follow the method introduced in [37] to compare the difference between per-exemplar dictionary and current-region representation. Given one category for example, we take its corresponding dictionary $D_i=[d_i^1, d_i^2, \ldots d_i^M]$ as the bases of sparse representation. For a testing image with K regions (super-pixels) $F=[f_1, f_2, .., f_K]$, where f_j indicates the local representation computed on super-pixel j, the sparse reconstruction coefficients corresponding to K super-pixels $A=[\alpha_1, \alpha_2, \ldots, \alpha_K]$ can be computed via

$$\alpha_j = \underset{\alpha_j}{\operatorname{argmin}} \|f_j - D_i\alpha_j\|_2^2 + \lambda\|\alpha_j\|_1. \quad (5)$$

Here, we solve the above equation with LARS algorithm, a variant for solving the Lasso [38]. Thus the sparse reconstruction error r_{ij}^s, indicating to what extent each region's local representation f_j belongs to category dictionary i, can be defined as

$$r_{ij}^s = \|f_j - D_i\alpha_j\|_2^2. \quad (6)$$

The sparse reconstruction error r_{ij}^s is regarded as a measurement of corresponding super-pixel category similarity. Compared with purely mapping or instance-detecting method, sparse reconstruction

can make full use of underlying consistent information and reveal the intrinsic cross-image correlations, which can well handle the variety of shape, color, illumination, and so on. In addition, considering that the process should be executed over the hierarchical regions, in nature it is an efficient way to simultaneously explore the intra-image spatial coherency and inter-image semantic consistence.

4.2. Dense reconstruction

In order to make complement of sparse reconstruction error, we conduct dense reconstruction based on Principal Component Analysis (PCA) method. For dictionary of each category, the corresponding eigenvectors of orthogonal bases in the dictionary capture the common property of features in the same dictionary, as well as excluding diversity of such features. Therefore, such new bases have great potentials to globally represent new local representation that belongs to the same category semantically. So we reconstruct unknown features by principal component bases.

Also, given one category for example, we compute the L largest eigenvalues of normalized covariance matrix of the corresponding dictionary $D_i=[d_i^1, d_i^2, \ldots, d_i^M]$, and form the bases from the eigenvectors $U_{D_i}=[u_{i1}, u_{i2}, \ldots, u_{iL}]$ with corresponding decreasing order of the eigenvalues. U_{D_i} represents a dense basis for category i in dense reconstruction process. For a testing image with K super-pixels, $F=[f_1, f_2, .., f_K]$ are the features to be labeled, where f_j indicates the local representation of super-pixel j, and the dense reconstructing coefficients corresponding to K super-pixels $\beta=[\beta_1, \beta_2, \ldots, \beta_K]$ can be computed via

$$\beta_j = U_{D_i}^{\top}(f_j - \overline{f}), \tag{7}$$

where $\overline{f}$ is the mean local representation of F. Thus, the dense reconstruction error $r_{ij}^{\ d}$ of super-pixel f_j, indicating to what extent it belongs to category dictionary i, is represented as

$$r_{ij}^d = \|f_j - (U_{D_i}\beta_j + \overline{f})\|_2^2. \tag{8}$$

We determine the candidate label of each region according to the weighted sum of the sparse reconstruction error and dense reconstruction error via

$$r_{ij} = \lambda r_{ij}^s + (1-\lambda) r_{ij}^d, \tag{9}$$

where λ is used to balance the effects of sparse and dense reconstruction process. Thus, the region can be preliminarily assigned to the same label as the category producing minimum errors, and it may be further adjusted by taking into account the same-level spatial coherency, which is detailed below.

Since both sparse and dense reconstruction errors facilitate to judge the unknown representation from different view points. In fact, the reasonability of sparse reconstruction error lies in that the unknown representation [39] can be linearly reconstructed by only a few of known representations. And the dense reconstruction error mainly considers the principle that the unknown representation can be roughly reconstructed by the principal components of certain category dictionary, which is expected to reveal the internal structure of the data in a way that best explains the variance in the data. Meanwhile, the sparse reconstruction error tends to be sensitive to the noise involved in the dictionary, while the dense reconstruction error may also behave badly when certain category has high variety. Therefore, by combining sparse and dense reconstruction, their disadvantages can be greatly weakened in a mutually complementary way.

4.3. Anisotropic reconstruction error propagation

Since adjacent neighbors may belong to the same category with high probability, or certain categories can be neighbors while others become neighbors without any semantic meanings. We cannot ignore the influence of neighborhoods. So neighbors can serve as complement description, we should consider reconstruction error both from itself and its neighborhoods. Here we let reconstruction error spread among neighborhoods to adjust final results.

To emphasize cross validation while suppressing the misleading interaction among neighboring regions, we assume that regions with similar appearance belong to the same category in all likelihood. So we should limit the per-exemplar reconstruction error to propagate only among the similar neighborhoods as much as possible. The propagation process can also facilitate the smoothing of possible discontinuous reconstruction errors within the local regions of certain potential semantic entity by introducing bi-harmonic distance measurement (refer to Section 3.1). Therefore, at first we determine the diffusion domain, which consists of adjacent super-pixels having smaller bi-harmonic distance than predefined threshold. The propagation only occurs from one super-pixel to another when they are in the same diffusion domain. This process is equal to locally conducting anisotropic convolution over sparse and dense reconstruction error fields respectively, wherein the local domains are handled in descending order with respect to the corresponding reconstruction error. The final reconstruction error of each region/super-pixel consists of two parts: original error and diffused error from other regions/super-pixels, which is formulated as

$$r'_{ij} = \lambda r_{ij} + (1-\lambda)\frac{\sum_{k=1,k\neq i}^{K} W_{jk} r_{ik}}{\sum_{k=1,k\neq i} W_{jk}}. \tag{10}$$

Here r'_{ij} is the new reconstruction error after the diffusion process, λ is used to balance between two kinds of reconstruction errors, i indicates the category of dictionary, and the weight W_{jk} is defined as

$$W_{jk} = \begin{cases} 0 & \text{if } S_j = S_k \\ \exp\left(-\dfrac{d_B(j,k)}{\beta}\right) & \text{if } S_j \neq S_k \end{cases}, \tag{11}$$

where S_j indicates the domain that the super-pixel j belongs to, $d_B(j,k)$ is the bi-harmonic distance between super-pixel j and super-pixel k, and β is the parameter used to control to what extent the bi-harmonic distance would influence the weight. The propagation process gives rise to better discriminative coherency. Fig. 3 demonstrates the sparse reconstruction error, dense reconstruction error, the diffusion domains, and the corresponding error-diffusing results respectively.

4.4. Hierarchical recognition voting for potentially semantic regions

After obtaining each super-pixel's reconstruction error, we can probably infer its semantics. Since the finer-levels' super-pixels represent the over-segmentation of the image, it cannot provide the convincing mid-level meanings, which is important for image parsing. In order to comprehensively infer the semantics within higher-level regions, we further employ graph-based image segmentation [34] to guide pixel-wise semantic parsing, because such image segmentation has the ability to keep the details in regions of low-variability while trying to overlook the details in regions of high-variability. Each of the high-level segmented regions can be regarded as an individual entity to be recognized. Therefore, we parse each region independently according to the probable semantics of the hierarchical super-pixels covered by it.

Given a high-level segmented region, if its super-pixels have large overlap in semantics, we will assign a candidate label to this region according to the corresponding semantics, because we suppose that most super-pixels should have the right labels after

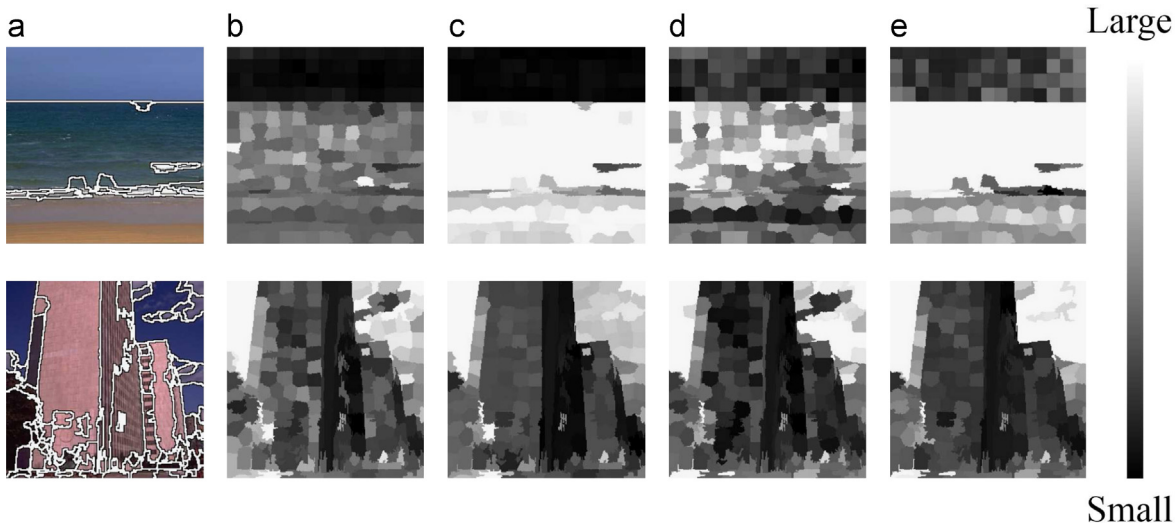


Fig. 3. Demonstration of the anisotropic reconstruction error propagation. The top row illustrates the reconstruction error about "sky" category, and the bottom is about "tall building" category. Deeper color denotes smaller reconstruction error, which means being closer to the corresponding category. (a) Reconstruction error in diffusion domains; (b) original sparse reconstruction error; (c) diffused sparse reconstruction error; (d) original dense reconstruction error; and (e) diffused dense reconstruction error. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

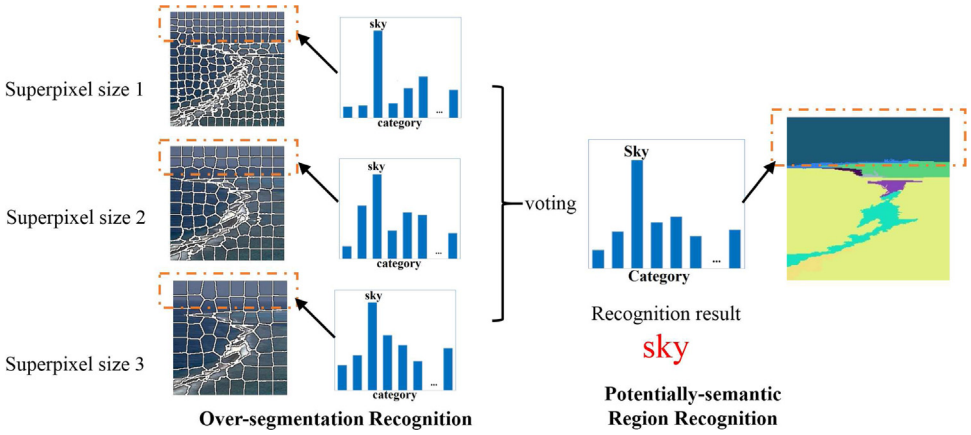


Fig. 4. Illustration of the hierarchical voting process from over-segmentation recognition to potentially semantic region recognition.

the error-diffusion process. However, there may inevitably exist some wrong labels. Therefore, similar to the common practice that we do when we have different ideas in daily life, we solve this problem by resorting to the voting-based strategy. We give each super-pixel a vote to decide which category this region should belong to, wherein the lower level the super-pixel locates in, the smaller its weight will be. Given a super-pixel located at (x,y), the probability $Pro_i(x,y)$ it belongs to the i-th category can be calculated as

$$Pro_i(x,y) = \frac{1}{L}\sum_{k=1}^{L} p_{ik}(x,y), \tag{12}$$

where L indicates the super-pixel's levels, $p_{ik}(x,y)$ represents the probability that the super-pixel (located at (x,y) in *k-level*) belongs to the i-th label. By summing the weighted votes of each category in certain region, we get the probability of i-th label p_{ij} that the j-th region may belong to. The hierarchical semantic voting process is illustrated in Fig. 4. Thus, this process can correct most of the wrong labels by excluding the influence of a small part of incorrect choices. For each high-level region in the testing image, we choose the higher probability labels as its candidate labels.

5. Parsing refinement assisted by spatial–contextual strings

In natural images, certain object classes may appear simultaneously and have certain location correlation. We further exploit spatial–contextual cues to remove the unreasonable candidate label choice. Since two-dimensional images are hard to compare directly due to variations in scale, position, and so on, we convert two-dimensional images with semantic meanings into one-dimensional semantic strings in order to reduce complex structure. By encoding the spatial–contextual cues into a semantic string, we can get the refined parsing result via statistical analysis on the string matching results.

Because dimensionality reduction can cause information loss, we should preserve useful information like relative relationships. One solution is to use Hilbert space-filling curve, which is traditionally used to traverse two-dimensional space with a goal to get one-dimensional continuous path while still preserving point neighborhoods as much as possible. When applying this technique in images, Hilbert curve traverses pixels according to certain pattern. Since traversing route goes through adjacent image regions in a pre-defined order, it can well reflect the image layout and objects' relative location, comparing with the naive row-by-row traversing methods and the prime spatial-subdivision scheme.

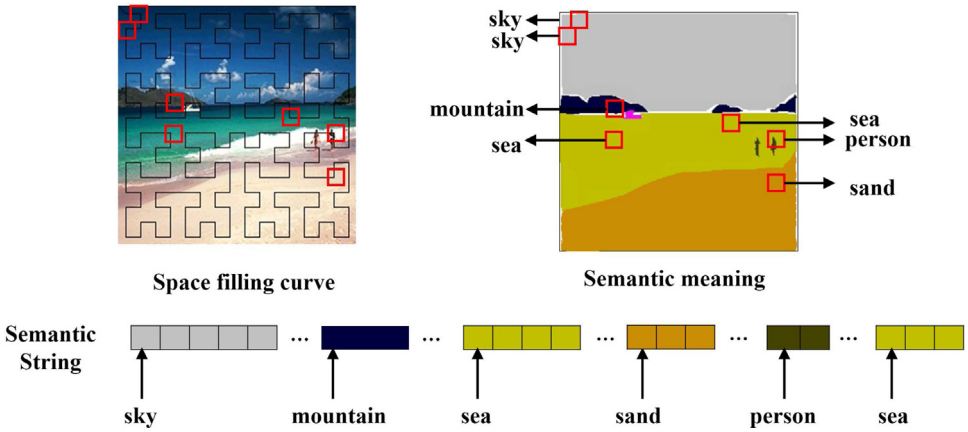


Fig. 5. Illustration of the encoded semantic string based on Hilbert space-filling curve.

Therefore, the Hilbert curve route provides us a simple yet effective way to encode the spatial–contextual cues as a semantic string. The string is obtained by walking along the Hilbert curve route and adding corresponding labels in its corresponding locations at the same time. Fig. 5 illustrates the encoded semantic string over super-pixels with different-levels.

Since our semantic string implicitly encodes the contextual information of the image, the specific string match algorithm is expected to efficiently and flexibly find similar content structures during comparing image contents. However, the strict string matching method (like KMP) cannot handle the problem of noise or finding the similar partial-structure between two strings. By conducting survey on the existing approximate string matching algorithm [40], we find that the dynamic programming method has great potential to solve this problem by keeping local optimal solution, which facilitates to efficiently find the most similar structure between two strings.

Therefore, we define the similarity of two semantic strings based on the dynamic programming method introduced in [41], which can capture the largest common structures of the two strings while ignoring noise-like perturbation in strings. Two strings with largest common structures will have the most similar meanings in corresponding images. Given two strings A_1 and A_2, supposing the length n of A_1 is shorter than the length m of A_2, we construct $(n+1)\times(m+1)$ dynamic programming matrix D. The matrix is initialized as

$$D_{i,j}=\begin{cases} W_{A_1(i),A_2(1)} & \text{if } j=0 \\ 0 & \text{elsewhere} \end{cases}, \tag{13}$$

and the dynamic programming matrix is updated according to

$$D_{i,j}=\min(D_{i-1,j-1},D_{i-1,j},D_{i,j-1})+W_{A_1(i),A_2(j)}. \tag{14}$$

The final $D_{n,m}$ represents the distance between A_1 and A_2, which measures the similarity of A_1 and A_2.

Since we have obtained the label probabilities of each region, some regions may have several candidate labels that are hard for us to decide arbitrarily. To solve this problem, we compare the semantic string of the testing image with those of the training images, because the annotated images can form semantic strings reflecting the common and reasonable layouts. For each testing image, based on the region-wise candidate labels, we can respectively construct corresponding semantic strings. As aforementioned, we have already selected out the similar images from annotated ones, by comparing the candidate testing image strings with those from annotated images, we can conduct statistics over the most similar strings, and assume the candidate string that mostly relates to the high frequencies of object assembly as the reasonable one, and this way we can finally get the refined parsing result.

6. Experiments and evaluations

6.1. Experiment settings

We have implemented our framework on a PC with Geforce GTX 770 GPU, Intel Core I7 CPU and 24G memory using C++ and the necessary invoking interface of MATLAB. We demonstrate the advantages of our method via extensive experiments on the popular SIFT Flow dataset [42], which consists of 200 test images and 2488 images from LabelMe, as well as 8 scenes and 33 semantic categories. In the experiment, we employ the average LAB color value and the corresponding standard deviation, uniform LBP in [43], average location, histogram of gradient texture and bi-harmonic distance distribution (BDD) as region-wise semi-local features.

As for the evaluation, we compare our method with several state-of-the-art image parsing methods, including Tighe and Lazebnik [5], Myeong et al. [44], and so on. We mainly use the pixel-level parsing accuracy and per-class accuracy published in their papers as quantitative indicators. In addition, because of the lack of per-class accuracy data in Liu et al. [42], we run their source codes to get pixel-level parsing accuracy and per-class accuracy.

6.2. Experiment comparisons and evaluations

Parameter analysis of our method: The performance of our method is mainly affected by several parameters, including retrieval set size when selecting similar images, average-pixel size when making over-segmentation and semantic graph-cut segment parameters.

The initial step of image parsing is to find similar images that contain useful segments to guide semantic annotation. The retrieval set should contain enough corresponding segments rather than introducing a large number of less useful segments. Therefore, the size of retrieval set is important. In our experiment, we adjust the size of retrieval set to find the proper number of the similar images. To verify the importance of similar image process, we also compare our method with k-means method about selecting candidate features. Table 1 lists the final accuracy of different sizes of retrieval set. From the accuracy, we can conclude that using too many images as candidate features will greatly influence our accuracy because more noise candidate features will be involved in the reconstruction process. Compared with averaging all features in one category using

Table 1
The image parsing per-pixel accuracy of different sizes of retrieval set.

Size of retrieval set	Accuracy	Accuracy of k-means
50	74.9 (19.5)	
100	74.3 (23.0)	
200	72.5 (24.6)	70.37
500	68.9 (27.8)	
1000	66.9 (28.1)	

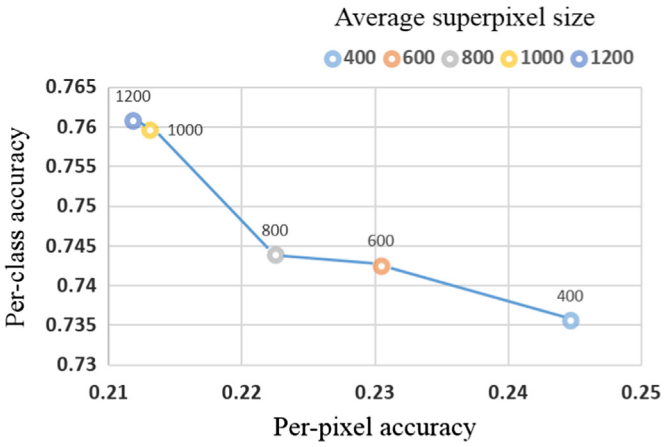


Fig. 6. Performance analysis under different settings of super-pixel size.

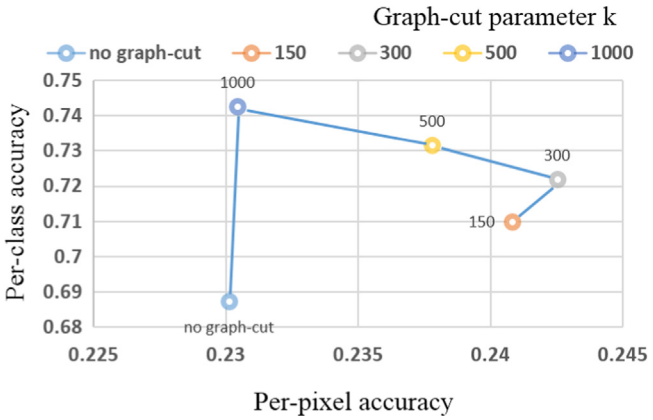


Fig. 7. Performance analysis of graph-cut under different settings of k value.

Table 2
The average image parsing accuracy comparison of different methods over SIFT Flow dataset.

Method	Per-pixel	Per-class
Gould and Zhang [45]	65.2	14.9
Liu et al. [42]	76.6	23.5
Tighe and Lazebnik [5]	77.0	30.1
Myeong et al. [44]	77.1	32.3
Farabet et al. [46]	74.2	46.0
OURS	77.4	20.6

Table 3
The category-wise average image parsing accuracy statistics of our method.

Class	Accuracy	The number of images	The number of features
sky	94.40	2080	46,929
building	91.41	991	33,676
road	80.19	701	12,021
mountain	78.92	849	21,908
tree	78.38	976	17,415
sea	72.85	344	9838
field	49.06	203	4860
grass	39.82	201	3346
car	37.52	334	1413
door	19.03	126	531
river	18.41	216	2414
sand	17.59	121	1881

k-means method, our selecting process can choose useful candidate words efficiently.

Super-pixel segmentation is an essential part of image parsing. Since we use SLIC super-pixel [35] as initial over-segmentation, we only have to decide the average size of super-pixels when performing segmentation. In order to combine both global and local features, here we use the hierarchical semantic voting method. By integrating different sizes of super-pixel's judgement, we can get more reasonable results. Here, we compare the accuracy using different sizes of super-pixels only. Fig. 6 shows the comparison of super-pixel sizes. From the accuracy, we can conclude that when the average size of super-pixels increases, the per-pixel accuracy increases, but at the expense of decrease on per-class accuracy. Larger-size super-pixels tend to recognize segments in a more global way, but give rise to ignoring small parts like people. Smaller-sized super-pixels tend to recognize small parts more accurately, but lose the overall appearance of objects.

Since our method is based on super-pixels which over-segment objects in images, and in view of super-pixels objects are considered to be combination of (many) fragments, which may mislead our overall and high-level judgement. So we add the process of graph-cut segmentation to coarsely create semantic segments. Fig. 7 displays the final accuracy of graph-cut segmentation of different parameters and the accuracy without graph-cut process. Here we vary k in the experiment (k is value for the threshold function), which causes a preference for large components. From the chart, we can find that larger k results in more accuracy in the per-pixel sense while less accurate in the per-class sense, due to the same reason as we had analyzed in the experiments about the super-pixel size. Also, when we remove the graph-cut process, the result indicates that both per-pixel and per-class accuracies decrease, which verifies the efficiency of such process.

Comparison with state-of-the-art methods: According to several well-recognized accuracy indicators used in image parsing, Table 2 lists the quantitative accuracy statistics of our method and several state-of-the-art methods, including per-pixel accuracy and per-class accuracy. Here, we can conclude that we achieve high accuracy in per-pixel accuracy, but our method falls behind in per-class accuracy when comparing with other methods. However, since high per-pixel accuracy reflects overall theme of the scene, it is more essential for parsing images.

Table 3 lists the category-wise average image parsing accuracy statistics of our method. When calculating the number of training images for certain category objects/things, it takes into account all the images containing corresponding objects/things in the training set. And the number of features takes into account all the representations extracted from the aforementioned training images that contain corresponding objects/things. It shows that our method can achieve good performance for the categories having enough annotated features. And thus our method can make full use of such comprehensive features to explain the corresponding entities in the new testing image. In the meanwhile, the categories which have small number of annotated images like bird behave worse in

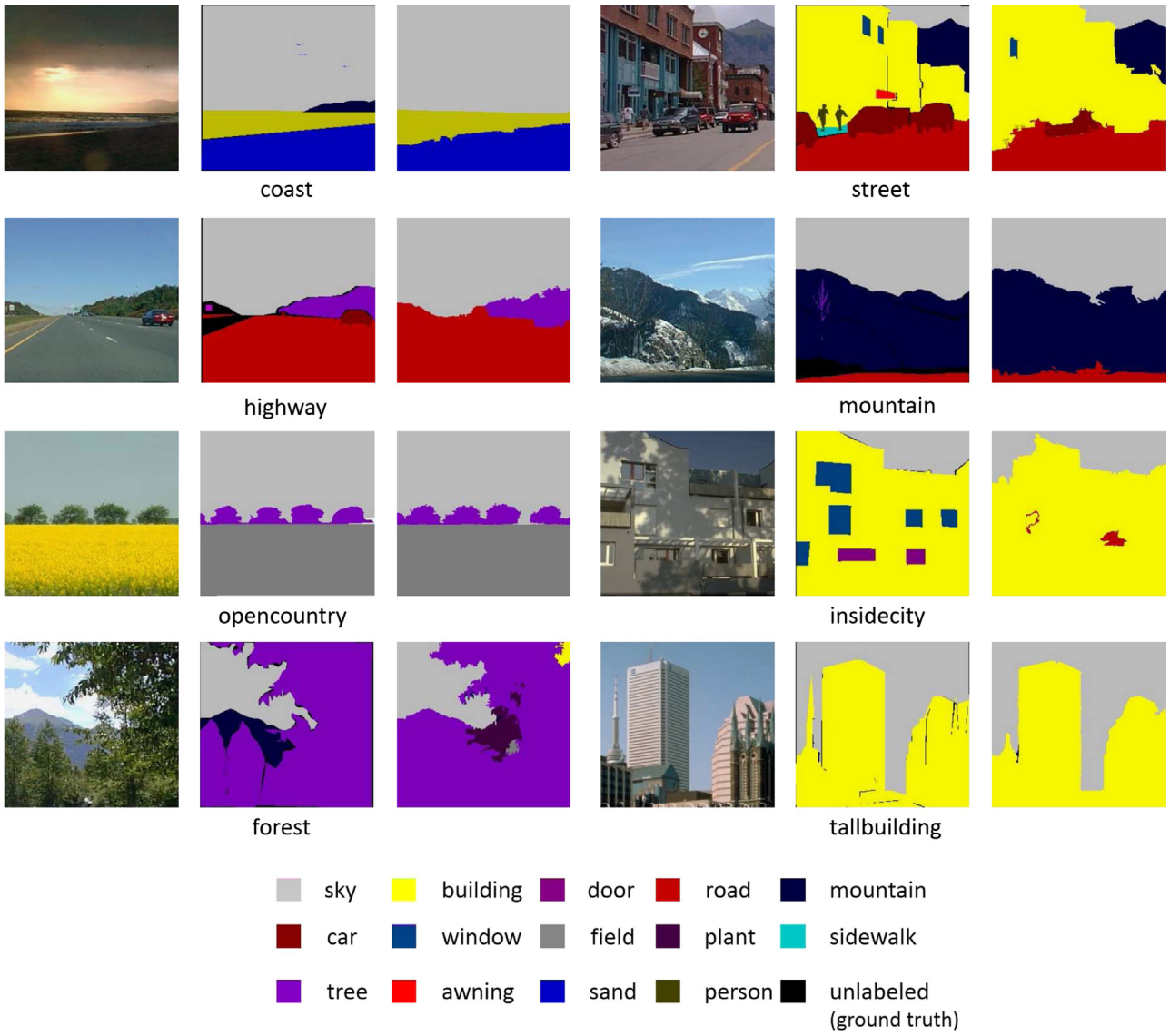


Fig. 8. Comparison of image parsing results between our method and the corresponding ground truth.

Table 4
Running time of each step.

Steps in our method	Running time (s)
Per-exemplar dictionary construction	2.21
Super-pixels segmentation	0.058
Descriptors computation	4.36
Reconstruction error computation	1.85
Error diffusion	0.016
Hierarchical voting	0.014
Graph-cut segmentation	0.051
Space filling curve construction	0.001
Dynamic programming comparison	0.008
Average time per image(s)	
Liu et al. [42]	3.97
Tighe et al. [5]	138.93
Ours	7.58

our method, which largely lower our final category related accuracy. The reason is that the corresponding category related dictionary does not include enough representation for parsing, which indicates that we should further expand our latent-learning training dataset to make the object classes distribute more uniformly. Fig. 8 shows our final parsing results over examples with different-situations.

Running time analysis: We detail the time cost of the different steps of our method separately in Table 4. Here, we only count the average running time of each step when calculating each layer. Besides, we compare the average per-image parsing time cost among Liu et al. [42], Tighe et al. [5] and our method, wherein the time computing takes into account descriptors computation process and parsing process. In addition, the dynamic programming comparison time represents only one image's string comparison time. We can conclude from Table 4 that the most time-consuming method is the computation process for descriptors, because our method uses several kinds of descriptors to make the representation more distinguishable to each other. Meanwhile, the per-exemplar dictionary construction and reconstruction error computation also cost more time than other processes. Such process's time is largely dependent on the retrieval set size when selecting similar images and the max words' number when constructing dictionary. Fig. 9 demonstrates the influences of the retrieval set size on dictionary construction and the max words' number on computation of reconstruction error. We can see that, when the

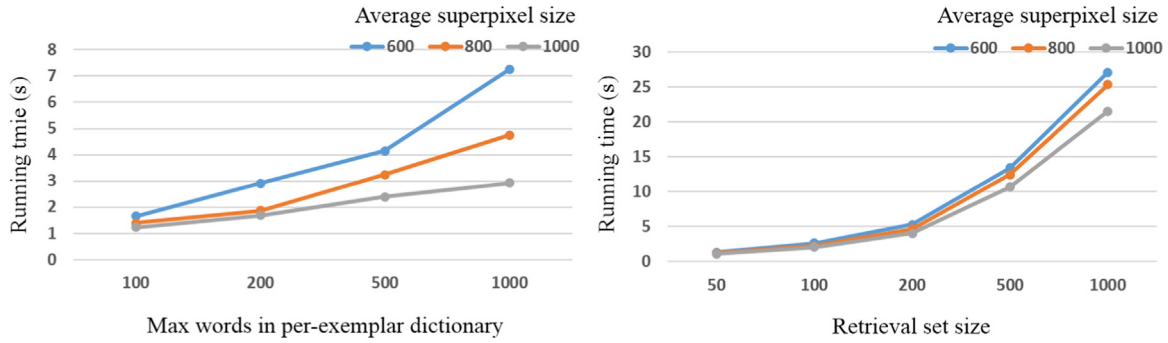


Fig. 9. Running time analysis under different parameter settings. The left chart is running time of reconstruction error computation of different max words in per-exemplar dictionary. The right chart is running time of dictionary construction with different sizes of retrieval set.

number of words increases, the corresponding time cost also increases.

7. Conclusion and discussion

In this paper, we have systematically presented a novel light-weight non-parametric method to address a suite of research challenges encountered in image parsing by integrating new modeling and inferring strategies into a flexible and expandable framework. The extensive experiments demonstrate the effectiveness of our method in handling natural images with high-variability and broad-coverage of things and/or entities. In particular, the critical and novel technical components of our approach include semi-local complementary feature integration based on bi-harmonic distance distribution, per-exemplar candidate-label detection based on sparse–dense reconstruction, anisotropic label-confidence propagation based on bi-harmonic distance distribution, hierarchical semantic voting, string encoding of global spatial–contextual cues based on Hilbert curve, relaxed semantic string matching, and its assistance in parsing refinement. All of these technical innovations contribute to automatic image parsing with state-of-the-art performance in accuracy, versatility, flexibility, and efficiency.

Nevertheless, if the objective is to perform absolutely fair comparisons with other state-of-the-art methods, our method may still have tremendous room to improve, and this is because, currently our latent-learning dictionaries are mainly obtained from the readily available SIFT Flow dataset, which only has 2488 annotated images. Specifically, the involved annotated object classes are extremely non-uniform, which lack enough annotated local regions for many categories such as awning, balcony and so on. Therefore, in our method the region-level reconstruction error will be larger when handling the testing images with object instances of such categories, which greatly restrain the performance potential of our method. In the future work, we plan to build complete annotated dataset for more comprehensive latent learning to further improve the parsing accuracy of our method. Meanwhile, extending our key ideas to other image applications, such as image retrieval and image classification, also deserves our immediate research endeavor in the near future.

Acknowledgments

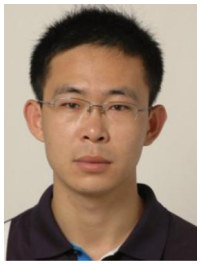
This research is supported by National Natural Science Foundation of China (No. 61190120, No. 61190125, No. 61190124, 61300067, and 61532002), and Beijing Key Laboratory (NO: BZ0211).

References

[1] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, arXiv preprint arxiv:1502.00717.
[2] J. Dong, Q. Chen, W. Xia, Z. Huang, S. Yan, A deformable mixture parsing model with parselets, in: IEEE International Conference on Computer Vision, 2013, pp. 3408–3415.
[3] A. Sharma, O. Tuzel, D.W. Jacobs, Deep hierarchical parsing for semantic segmentation, arXiv preprint arxiv:1503.02725.
[4] P.H. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene parsing, arXiv preprint arxiv:1306.2795.
[5] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: European Conference on Computer Vision, September 5-11, 2010, Heraklion, Crete, Greece, Springer, 2010, pp. 352–365.
[6] L. Zhu, Y. Chen, Y. Lin, C. Lin, A. Yuille, Recursive segmentation and recognition templates for image parsing, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2012) 359–371.
[7] B. Zhao, E.P. Xing, Hierarchical feature hashing for fast dimensionality reduction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2051–2058.
[8] J. Dai, Y. Hong, W. Hu, S.-C. Zhu, Y.N. Wu, Unsupervised learning of dictionaries of hierarchical compositional models, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2505–2512.
[9] C. Zhou, C. Liu, Semantic image segmentation using low-level features and contextual cues, Comput. Electr. Eng. 40 (3) (2014) 844–857.
[10] W. Xie, Y. Peng, J. Xiao, Semantic graph construction for weakly-supervised image parsing, in: AAAI Conference on Artificial Intelligence, 2014, pp. 2853–2859.
[11] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, Int. J. Comput. Vis. 81 (1) (2009) 2–23.
[12] X. Chen, A. Jain, L.S. Davis, Object co-labeling in multiple images, in: IEEE Winter Conference on Applications of Computer Vision, 2014, pp. 721–728.
[13] D. Kuettel, M. Guillaumin, V. Ferrari, Segmentation propagation in imagenet, in: European Conference on Computer Vision, 7-13 October 2012, Florence, Italy, vol. 7578, Springer, 2012, pp. 459–473.
[14] D. Kuettel, V. Ferrari, Figure-ground segmentation by transferring window masks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 558–565.
[15] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, S. Yan, Fashion parsing with weak color-category labels, IEEE Trans. Multimedia 16 (1) (2014) 253–265.
[16] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: IEEE International Conference on Computer Vision, 2009, pp. 1–8.
[17] J. Yuan, J. Li, B. Zhang, Scene understanding with discriminative structured prediction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[18] Z. Tu, X. Chen, A.L. Yuille, S.-C. Zhu, Image parsing: unifying segmentation, detection, and recognition, Int. J. Comput. Vis. 63 (2) (2005) 113–140.
[19] D. Modolo, A. Vezhnevets, V. Ferrari, Context forest for efficient object detection with large mixture models, arXiv preprint arxiv:1503.00787.
[20] A. Humayun, F. Li, J.M. Rehg, Rigor: reusing inference in graph cuts for generating object regions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 336–343.
[21] B. Hariharan, P. Arbelez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: European Conference on Computer Vision, September 6-12, 2014, Zurich, Switzerland, Springer, 2014, pp. 297–312.
[22] X. Wang, L. Zhang, L. Lin, Z. Liang, W. Zuo, Deep joint task learning for generic object extraction, in: Advances in Neural Information Processing Systems, 2014, pp. 523–531.
[23] C. Liu, J. Yuen, A. Torralba, Sift flow: dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 978–994.
[24] D. Eigen, R. Fergus, Nonparametric image parsing using adaptive neighbor sets, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2799–2806.

[25] J. Tighe, S. Lazebnik, Finding things: image parsing with regions and per-exemplar detectors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3001–3008.
[26] M.-M. Cheng, S. Zheng, W.-Y. Lin, V. Vineet, P. Sturgess, N. Crook, N.J. Mitra, P. Torr, Imagespirit: verbal guided image parsing, ACM Trans. Graph. 34 (1) (2014) 3.
[27] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, H. Lu, Weakly supervised graph propagation towards collective image parsing, IEEE Trans. Multimedia 14 (2) (2012) 361–373.
[28] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, S. Yan, Matching-CNN meets kNN: quasi-parametric human parsing, arXiv preprint arxiv:1504.01220.
[29] J.L. Long, N. Zhang, T. Darrell, Do convnets learn correspondence, in: Advances in Neural Information Processing Systems, 2014, pp. 1601–1609.
[30] X. He, R.S. Zemel, M.A. Carreira-Perpindn, Multiscale conditional random fields for image labeling, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 695–702.
[31] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[32] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[33] T.L. Vu, C.H. Lee, Robust method to compute mutual-spatial feature for image parsing problem, J. Image Graph. 2 (1) (2014) 54–58.
[34] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.
[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Suesstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[36] Y. Lipman, R.M. Rustamov, T.A. Funkhouser, Biharmonic distance, ACM Trans. Graph. 29 (3) (2010) 27.
[37] X. Li, H. Lu, L. Zhang, X. Ruan, M.-H. Yang, Saliency detection via dense and sparse reconstruction, in: IEEE International Conference on Computer Vision, 2013, pp. 2976–2983.
[38] B. Efron, T. Hastie, T. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 494–499.
[39] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289–1306.
[40] G. Navarro, A guided tour to approximate string matching, ACM Comput. Surv. 33 (1) (2001) 31–88.
[41] T. Qi, J. Xiao, Y. Zhuang, H. Zhang, X. Yang, J. Zhang, Y. Feng, Real-time motion data annotation via action string, Comput. Animat. Virtual Worlds 25 (3–4) (2014) 291–300.
[42] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2368–2382.
[43] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[44] H. Myeong, J.Y. Chang, K.M. Lee, Learning object relationships via graph-based context model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2727–2734.
[45] S. Gould, Y. Zhang, Patchmatchgraph: building a graph of dense patch correspondences for label transfer, in: European Conference on Computer Vision, 2012, pp. 439–452.
[46] C. Farabet, C. Couprie, L. Najman, Y. Lecun, Scene parsing with multiscale feature learning, purity trees, and optimal covers, arXiv preprint arxiv:1202.2160.

Xinyi An received her B.Sc. degree in Computer Science, Beihang University, China. Now she is a Master Student in Beihang University. From 2013, she is studying in the State Key Laboratory of Virtual Reality Technology and Systems in China. Her research interests include computer vision and image processing.



Shuai Li received the Ph.D. degree in Computer Science from Beihang University. He is currently an Associate Professor at the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His research interests include computer graphics, physics-based modeling and simulation, virtual surgery simulation, computer vision, and medical image processing.

Hong Qin received the B.S. and M.S. degrees in Computer Science from Peking University. He received the Ph.D. degree in Computer Science from the University of Toronto. He is a Professor of Computer Science in the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing. He is a Senior Member of the IEEE.

Aimin Hao is a Professor in Computer Science School and the Associate Director of State Key Laboratory of Virtual Reality Technology and Systems at Beihang University. He received his B.S., M.S., and Ph.D. in Computer Science at Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, geometric modeling, image processing, and computer vision.