Partial similarity based nonparametric scene parsing in certain environment

Honghui Zhang∗ Tian Fang∗ Xiaowu Chen† Qinping Zhao† Long Quan∗
The Hong Kong University of Science and Technology∗
Beihang University†
{honghui,fangtian,quan}@cse.ust.hk chen@buaa.edu.cn zhaoqp@vrlab.buaa.edu.cn

Abstract

In this paper we propose a novel nonparametric image parsing method for the image parsing problems in certain environment. A novel and efficient nearest neighbor matching scheme, the ANN bilateral matching scheme, is proposed. Based on the proposed matching scheme, given a test image, we first retrieve some partially similar images from the training image database so that for each region in the test image, a similar region exists in one of the retrieved training images. With the retrieved training images, we match the test image to the retrieved training images with the ANN bilateral matching scheme, and parse the test image by integrating multiple cues in a Markov random field. Experiments on three datasets shows our method has achieved promising parsing accuracy and outperforms two state-of-the-art nonparametric image parsing methods that are proposed for general image parsing problem.

1. Introduction

Image parsing, or segmenting objects in images and identifying their categories is one of the core problems of computer vision. Traditional learning-based methods[19, 20, 6] typically work with a fixed-number of object categories and require a generative or discriminative model to be trained in advance. As an alternative, with the increasing availability of large image collections with annotation, the “label transfer” methods have shown the potential for solving the image parsing problem with a data driven way [13, 23], other applications including object and scene recognition [17], image geo-location [5]. Instead of training sophisticated parametric models, these methods try to reduce the inference problem for an test image to the problem of retrieving similar training images for the test image and transferring annotation information from the retrieved training images to the test image, also called nonparametric methods. As data driven methods, these nonparametric methods for image parsing [13, 23] which target at general image parsing problem, usually require a large training dataset. Though great effort has been devoted to create such large datasets, like the LabelMe dataset [18], it is still hard to say that they have been large enough, as the variability of the visual world is so vast.

In general, nonparametric image parsing involve the two key issues: (1) How to retrieve some similar images from a training database for a given test image; (2) How to parse the test image with the retrieved images by transferring the annotation associated with the retrieved images to the test image. The first issue is well studied in some previous work [15, 24], and different scene matching methods that measure the global similarity between images are proposed, such as the methods used in nonparametric image parsing [13, 23] and scene completion [4]. However, in the context of label transfer, the global similarity should not the only consideration. On one hand, the number of retrieved training images is limited for efficiency and performance consideration, and the global similarity is used to select those best candidates. On the other hand, the retrieved images should cover all categories that exist in the test image, as the test image is parsed by transferring the label from the retrieved
images to the test image. Otherwise, categories that exist in the test image, but are missed in the retrieved training images, can not be correctly parsed. These two requirements are difficult to be satisfied at the same time in the global similarity based retrieval. In a database of limited size, what typically happens during scene matching is that some parts of the query image are matched quite well, while other parts are not. For example, for a street scene, one matching image could have a building match very well, but getting road missing, while another matching image could get the road exactly right, but without a building.

Though it is usually difficult to find some highly similar images for a test image from a database of limited size, partial similarity is common between images, as Figure 1 shows. It has already been shown that a query image could be well explained by a spatial composite of different regions taken from several different images in previous work [16], and similar idea was exploited in the object category recognition [9]. This motivates us to exploit the partial similarity between images for the label transfer. Instead of retrieving global similar images from the training database, we retrieve some partially similar images so that for each region in the test image, a similar region exists in one of the retrieved training images. Based on this motivation, we proposed a novel nonparametric parsing method for images in certain environment based on partial similarity between images, including a novel method for retrieving partially similar training images and a parsing scheme with retrieved training images. In the following, we first briefly introduce some related work.

2. Related work

The nonparametric image parsing has been studied in several previous work [13, 23, 1]. Liu et al. [13] addressed this problem as “label transfer” for the first time, and solved it by transferring label from a set of similar training images retrieved from a database to a given test image based on the improved SIFT-flow matching scheme. To retrieve proper training images for the label transfer, Gist matching [14] is first used to find some globally similar images for a given test image, and the retrieved images are re-ranked according to computed the SIFT flow energy from them to the test image, with the top re-ranked retrievals selected for later label transfer. The SIFT flow energy can be understood as a kind of global similarity measurement that overcomes the layout limitation in general global similarity. However, as mentioned in the introduction, finding highly similar training images for a given query image is difficult, especially when the size of training database is small. Without retrieved training images of high quality, it has been observed that the parsing performance will be severely affected [13]. At the same time, it has been found that some flawed matches are generated by this SIFT-flow based method [1], inconsistent with the quantitative energy term. Another problem of the SIFT-flow based nonparametric image parsing method is low efficiency, in terms of time cost [23]. Another more efficient nonparametric image parsing method is proposed [23], following the idea of label transfer. It works by scene-level matching with global image descriptors, followed by superpixel-level matching with twenty different histogram based superpixel features and efficient Markov random field (MRF) optimization for incorporating neighborhood context. Before the computation of superpixel features, multiple descriptor coding schemes are needed to be established as prerequisite. Similar to the trained models involved in learning based segmentation methods [19, 20, 6], descriptor coding schemes need to be updated with the updating of training data, such as the increase of training data. In addition, these two nonparametric parsing methods [13, 23] both target at general image parsing problem and require an existing large database. However, the size of available databases with annotation information is still relative small for general image parsing. To overcome this problem, different from the above two methods that target are general image parsing problem, the proposed method targets at a constrained image parsing problem, the image parsing in certain environment, like urban environment. It involves no any preprocessing steps, like that involved in the previous work [23].

In the following, the paper is structured as follows: The ANN(approximate nearest neighbor) bilateral matching scheme is introduced in section 3. Then based on this matching scheme, a method to retrieve proper partially similar images for a given test image is presented in section 4, and the way to parse the test image with the retrieved partially similar images is introduced in section 5. Then we evaluate the proposed parsing method in section 6 through experiments on three datasets, and conclude in section 7.

3. ANN bilateral matching

Nearest neighbor matching has been proven to be useful for many computer vision applications [21, 3, 2]. Various nearest neighbor algorithms for searching similar patches in images are compared [11]. The dense nearest neighbor searching can not be solved efficiently, even with the most efficient algorithm. However, in the task of label transfer, we do not pursuit exact nearest neighbor matching as some other applications, and the goal is to match pixels of the same category together. Therefore, based on the PatchMatch [2], we propose the ANN bilateral matching, approximate nearest neighbor bilateral matching. Both the retrieval of partially similar images and the successive parsing scheme are built upon the ANN bilateral matching, which can establish dense correspondence across images on pixel level efficiently and integrate prior knowledge naturally.
3.1. Image-to-Image ANN bilateral matching

Different from the PatchMatch algorithm, the ANN bilateral matching is a bilateral matching scheme. Given two images A and B to be matched, it does not only find a correspondence for each pixel in A from B, but also a correspondence for each pixel in B from A. We use two offset maps $f_A$ and $f_B$ with the same size as A and B to store the correspondence information, the offset vectors for correspondences from A to B and B to A respectively. For example, for a pixel $p = (p_x, p_y)$ in A which is matched to $q = (q_x, q_y)$ in B, an offset vector $(q_x - p_x, q_y - p_y)$ will be stored in the corresponding position $(p_x, p_y)$ of the offset map $f_A$. If an offset vector $v = (q_x - p_x, q_y - p_y)$ for a pixel $p = (p_x, p_y)$ in one image is valid, then the position of the matched pixel can be located in another image by $q = p + v = (q_x, q_y)$. If $(q_x, q_y)$ is outside of the corresponding image domain, the offset vector is invalid. In the following parts, we will use $f(x, y)$ to denote the offset vector stored in $(x,y)$ of offset map $f$.

First, $f_A$ and $f_B$ are initialized randomly with uniformly sampled offset vectors from the valid offset vector space. Then, $f_A$ and $f_B$ will be iteratively updated until the convergence condition is satisfied. In our implementation, when the average matching distance of matched pairs keeps unchanged, we consider the algorithm converges. Each iteration of the algorithm proceeds as follows: inner-propagation, random search and intra-propagation. An outline of the algorithm is illustrated in Figure 2, with different steps explained in Figure 3.

3.1.1 Inner-propagation

This step is similar to the propagation of the PatchMatch algorithm. It is based on the insight that when a good offset is found for a pixel, it would probably also be a good guess for adjacent pixels. Checking the adjacent offsets makes the searching cooperatively. Offset map $f$ will be updated as:

$$f(x, y) = \arg \min_v \{D(v)\} \quad (1)$$

$$v = \{f(x, y), f(x-1, y), f(x+1, y), f(x, y-1), f(x, y+1)\} \quad (2)$$

where $D(v)$ denotes the matching distance between the pixel at $(x, y)$ in one image and the matched pixel $(x, y) + f(x, y)$ in another image. The inner propagation will be performed twice in each iteration, from the left-top to the right-bottom and the the right-bottom to left-top respectively.

3.1.2 Random search

In the $i$th iteration, the offset map $f$ will be updated by testing a sequence of candidate offsets which are generated by:

$$u_i = f(x, y) + ws\alpha^i R_i \quad (3)$$

$$f(x, y) = \arg \min_{u_i} D(u_i) \quad (4)$$

where $R_i$ is a uniform random in $[-1, 1] \times [-1, 1]$, $s$ is a large maximum search window that is clamped to the bounds of valid search space, $\alpha$ is the shrinking ratio of the search window size, $w = 1 - \exp(-D(f(x, y)))$ is a weight term that makes the updating process adaptive to the
matching distance. With larger matching distance, it generates longer candidate offset sequence. The updating process stops when the search window size $w_{\alpha}$ shrinks to less than one pixel.

### 3.1.3 Intra-propagation

In this step, offset maps exchange good offsets found in individual inner-propagation and random search. Offset maps $f_A(x, y)$ and $f_B(x, y)$ are updated as:

\[
\begin{align*}
    f_A(x, y) &= \arg\min_v \{ D(v) : v = \{ f_A(x, y), (x_0 - x, y_0 - y) \} \} \\
    f_B(x_0, y_0) + (x_0, y_0) &= (x, y) \\
    f_B(x, y) &= \arg\min_v \{ D(v) : v = \{ f_B(x, y), (x_0 - x, y_0 - y) \} \} \\
    f_A(x_0, y_0) + (x_0, y_0) &= (x, y)
\end{align*}
\]

The motivation behind the intra-propagation is similar to the inner-propagation: if pixel $q$ in $A$ is a good matching for pixel $p$ in $B$, then $p$ is probably also a good matching for $q$ among all pixels in $B$.

### 3.1.4 Mismatch rejection

Among the dense correspondences, mismatches are inevitable, given the two images matched only partially similar. To reduce these mismatches, we only keep those correspondences that satisfy the following condition:

\[
\begin{align*}
    \| f_A(x, y) + f_B(x_0, y_0) \| &\leq \varepsilon \\
    (x_0, y_0) &= f_A(x, y) + (x, y)
\end{align*}
\]

where $\varepsilon$ is a predefined small shift range. Suppose the probability of finding a semantically correct matching for one pixel in one image from another image is $\rho$, the probability that a correspondence satisfying the above condition is semantically correct is increased to $1 - (1 - \rho)^2 = (1 - \rho)\rho + \rho$, when $\varepsilon = 0$. For a rejected correspondence, the corresponding matching distance will be set as $\infty$.

### 3.1.5 Distance metric for ANN bilateral matching

We use the following distance metric to compute the matching distance of correspondences, which naturally integrates image level prior into the matching scheme. Given two images $A$ and $B$ to be matched, when searching a correspondence $q$ in $B$ for one pixel $p$ in $A$, the matching distance is computed as:

\[
D(p, q) = \begin{cases} 
    \| (D_p - D_q) \|^2 & \text{if } C_q \in R, \text{or } R = \emptyset \\
    +\infty & \text{else}
\end{cases}
\]

where $D_p$ and $D_q$ are the feature descriptor of $p$ and $q$. $C_q$ is the category label of $q$, obtained from the associated annotation of $B$, if it is available. $R$ is the image level prior about the categories contained in $A$. When no image level prior of $A$ or no annotation of $B$ is available, we set $R = \emptyset$. For the feature descriptor, the Texton feature [8] is used. In addition to the image level prior knowledge, we can also naturally integrate the position prior knowledge into the matching process by imposing constraint on the offset of two matched pixels.

### 3.2. Image-to-ImageSet ANN bilateral matching

For the image parsing by label transfer, we actually need to establish correspondence between the test image and multiple retrieved training images, which form an image set. Based on the proposed Image-to-Image ANN bilateral matching, we propose the Image-to-ImageSet ANN bilateral matching to match an image and an image set.

Given an test image $I$ and an image set $\{A_k\}_{k=1}^N$, we first perform the Image-to-Image ANN bilateral matching with $I$ and each image in $\{A_k\}_{k=1}^N$. For each pixel $p$ in $I$, we can find $N$ correspondences from $\{A_k\}_{k=1}^N$, and we denote them as $\{p, A_k, (x_p^k, y_p^k), D_p^k\}_{k=1}^N$. For the notations, $(x_p^k, y_p^k)$ denotes the position of the matched pixel in $A_k$, and $D_p^k$ denotes the corresponding matching distance. Then, we select one proper correspondence for $p$ from the $N$ candidates $\{p, A_k, (x_p^k, y_p^k), D_p^k\}_{k=1}^N$. We formulate this problem with a markov random field so that we can integrate spatial
smoothness preference into the selection, with the following energy function:

$$E(L) = \sum_{v \in I} \mu(L) + \alpha \sum_{e \in E} \phi(L_i, L_j)$$

(12)

The label set \( \{L_k\}_{k=1}^N \) for each pixel in \( I \) consists of the index set \( \{1, 2, ..., N\} \), corresponding to its \( N \) candidate correspondences. The data term \( \mu(L) \) is set as the matching distance between pixels and their candidate correspondences measured by the distance metric \( (11) \). \( E \) contains all the 4-connected spatial neighborhood of pixels. For any two adjacent pixels \( s \) and \( t \) linked by an edge in \( E \), smooth term is defined as:

$$\phi(L_i, L_j) = \left\{ \begin{array}{ll}
0 & \text{if } (x_s^{L_i}, y_s^{L_i}) \text{ and } (x_t^{L_j}, y_t^{L_j}) \text{ adjacent;} \\
\lambda & \text{else.}
\end{array} \right.$$  

(13)

The alpha expansion algorithm [10] is used to obtain a dense matching configuration. An example of matching a test image and an image set with annotation is illustrated in Figure 4.

4. Retrieval of partially similar source images

To find some proper partially similar images for a given test image, we first retrieve \( N \) initial candidate images by Gist matching [15] from the training database, then we match the test image to the retrieved candidate image set \( \{A_k\}_{k=1}^N \) with the proposed matching scheme in section 3.2 to re-rank the candidate images according to their relevance to the test image. The relevance of each image \( A \) in \( \{A_k\}_{k=1}^N \) to the test image is measured by: \( r = M/S \), where \( M \) denotes the number of pixels in the test image are matched to \( A \), and \( S \) denotes the total number of pixels in the test image. With the matching map as Figure 4 shows, the relevance can be computed very easily. Meanwhile, with the obtained dense pixel-level matching, we can also predict the image level prior about the content contained in the test image. Through the parsing result obtained by parsing every pixel in the test image with the label of the matched pixel in the source image set, as shown in Figure 4, we compute the ratio of different categories in the parsing result and add categories with ratio larger than 0.01 to the image level prior set of the test image. With the re-ranked candidate images, the top \( K \) images with the largest positive relevance are selected as the source images for the successive parsing.

In this step, the matching scheme is performed on a coarse scale based on the following two considerations: (1) efficiency; (2) better image level prior prediction, as patches with the same size on more coarse scale can carry more semantic information. As all regions in the test image will be included in the relevance measurement, we can capture the partial similarity between the test image and training images.

5. Parsing with retrieved source images

After performing the retrieval scheme in section 4, we perform the matching scheme in section 3.2 for the test image and the retrieved source image set on the finest scale. In this step, the image level prior predicted in the retrieval process is used in the distance metric \( (11) \). With obtained dense pixel-level correspondences, a simple method to parse the test image is directly parsing each pixel in the test image with the label associated with the matched pixel in images from the source image set. However, the result by this pixel level parsing is usually quite noisy, as shown in Figure 4. To overcome this problem, we formulate the parsing problem as an optimization problem in a markov random field by introducing spatial regularity.

First, the test image \( I \) is decomposed into super pixels using the algorithm [12]. A graph \( G = (\mathcal{V}, \mathcal{E}) \) is defined, with each vertex \( T \in \mathcal{V} \) in the graph denotes one super pixel in \( I \), while the edges \( \mathcal{E} \) denotes the neighboring relationship between super pixels. Then we build a markov random field upon \( G \), with the energy function defined as:

$$E(C) = \sum_{T \in \mathcal{V}} \psi_t(C_t) + \kappa \sum_{e_{ij} \in \mathcal{E}} \delta_{ij}(C_i, C_j)$$

(14)

data term and smooth term are defined as:

$$\psi_t(C_t) = p(C_t|T) = h(C_t)/S_T$$

(15)

$$\delta_{ij}(C_i, C_j) = |C_i \neq C_j| S_T / (1 + \lambda ||D_i - D_j||^2)$$

(16)

where \( h(C_t) \) is the number of pixels in \( T \) whose matching correspondences are associated with label \( C_t \) and \( S_T \) is the total number of pixels in \( T \). \( D_i \) and \( D_j \) are the feature descriptor of two neighboring super pixels, we choose the average RGB color of super pixel \( T \) as the feature descriptor. The alpha expansion algorithm [10] is used to optimize the energy function and obtain an optimal label configuration.

6. Experiment

To evaluate the performance of our method, we tested and compared it with several state-of-the-art techniques on three datasets: Multiple-View dataset, Google Street View dataset and the polo dataset. As the evaluation criterion used in some previous work [20, 19], the category average accuracy (the average proportion of pixels correctly labeled in each category) and the global accuracy (total proportion of pixels correctly labeled) are used to evaluate the performance of our method.
Multiple-View dataset The Multiple-View dataset is composed by taking three shoe multiple view sequences, three motorbike multiple view sequences from the training dataset used in previous work [22]. This experiment is designed to evaluate the performance of the proposed ANN bilateral matching scheme, and we compared it with the coarse-to-fine SIFT-Flow based parsing method [13]. Each image is segmented into two categories: foreground and background. For each sequence, we randomly select two images with annotation as the training images, and left images are used as testing images. We test our method and the SIFT-Flow based parsing method by using the two selected training images as source images for label transfer directly, bypassing the retrieval of source images. Through this way, the two methods are tested with exactly the same data. Some parsing results by our method and the SIFT-Flow based parsing method are shown in Figure 5. The average parsing accuracy on the six sequences of our method is 90.6%/93.2% for foreground/background, much better than that of SIFT-Flow based parsing method, 67.9%/92.4% for foreground/background.

Google Street View dataset The Google Street View dataset consists of about 10,000 images captured in the downtown of Pittsburgh by Google Street View, used in previous work [25, 7]. We randomly selected 320 images from the entire dataset, and manually labeled them into five categories: building, ground, car, sky, tree. For quantitative evaluation, the labeled images are split into train/test set with ratio 50%/50% by random selection for the testing of our method and the SuperPixel parsing method [23]. For the initial candidate images of the relevance based retrieval in our method, we select the top one hundred matches of Gist matching and choose the top five candidate images with the largest relevance as the source images for the successive parsing. For the SuperPixel parsing method, we also set the size of retrieval set as one hundred. We compare the performance of our method with the SuperPixel parsing method [23] and the reported state-of-the-art performance obtained by several other methods on this dataset [7], shown in Table 1. In terms of global accuracy and category average accuracy, our method outperforms the other two nonparametric parsing methods [13, 23]. Compared with the supervised label transfer method [7], the global accuracy is improved, with the drop in category average accuracy. Some parsing result obtained by our method is shown in Figure 6.

To evaluate the improvement by the proposed relevance based retrieval for partially similar images in section 4, we also compare the performance of our method with different retrieval scheme: the proposed relevance based retrieval and Gist matching based retrieval. For both retrieval schemes, the top five matches are chosen for the successive parsing. The comparison is shown in Table 2. Through the comparison, we can find the improvement brought by the relevance based retrieval scheme is significant. The parsing accuracy of all categories is improved, with global accuracy and category average accuracy significant improved. With respect to the efficiency, average processing time to parse one image of size 320×453 is 75 second.

polo dataset The polo dataset consists of 320 images downloaded from the flickr with keyword polo, and most of them are related to the popular polo game. We labeled each image into six categories: sky, horse, person, ground, tree, grass. This dataset is split into train/test set by random selection, with ratio 25%/75%. We compare our method with the Textonboost method [20] and the Super-pixel pars-
Table 1. The performance of our method and several other methods on the Google Street View dataset.

<table>
<thead>
<tr>
<th>Retrieval scheme</th>
<th>building</th>
<th>car</th>
<th>ground</th>
<th>sky</th>
<th>tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gist Matching</td>
<td>0.952</td>
<td>0.366</td>
<td>0.879</td>
<td>0.829</td>
<td>0.295</td>
</tr>
<tr>
<td>Relevance</td>
<td>0.953</td>
<td>0.405</td>
<td>0.96</td>
<td>0.925</td>
<td>0.414</td>
</tr>
</tbody>
</table>

Table 2. The performance of our method tested with different retrieval scheme on the Google Street View dataset: Gist matching based retrieval 89.1%/66.4%(global accuracy/average category accuracy), relevance based retrieval 93.2%/73.1%(global accuracy/average category accuracy).

Table 3. The performance of our method tested with different retrieval scheme on the polo dataset: Gist matching retrieval 84%/70.7%(global accuracy/average category accuracy), relevance based retrieval 89.8%/82.5%(global accuracy/average category accuracy).

Table 4. The performance of our method tested with different retrieval scheme on the polo dataset: Gist matching retrieval 84%/70.7%(global accuracy/average category accuracy), relevance based retrieval 89.8%/82.5%(global accuracy/average category accuracy).

Figure 6. Some parsing result obtained by our method on the Google Street View dataset.

Figure 6. Some parsing result obtained by our method on the Google Street View dataset.
Acknowledgements This work was partially supported by the Hong Kong RGC GRF 618908 and 619409, and the National Natural Science Foundation of China (60933006).

References


