NONPARAMETRIC SCENE PARSING WITH DEEP CONVOLUTIONAL FEATURES AND DENSE ALIGNMENT

Chih-Hao Ma[1], Chiou-Ting Hsu[1], and Benoit Huet[2]

[1]Department of Computer Science, National Tsing Hua University, Taiwan
[2]Multimedia Communications Department, Eurecom, France

ABSTRACT

This paper addresses two key issues which concern the performance of nonparametric scene parsing: (1) the semantic quality of image retrieval; and (2) the accuracy in label transfer. First, because nonparametric methods annotate a query image through transferring labels from retrieved images, the task of image retrieval should find a set of "semantically similar" images to the query. Second, with the retrieval set, a good strategy should be developed to transfer semantic labels in pixel-level accuracy. In this paper, we focus on improving scene parsing accuracy in these two issues. We propose using the state-of-the-art deep convolutional features as image descriptors to improve the semantic quality of retrieved images. In addition, we include dense alignment into the Markov Random Field inference framework to transfer labels at pixel-level accuracy. Our experiments on the SIFT Flow dataset shows the improvement of the proposed approach over other nonparametric methods.

Index Terms **—** scene parsing; object window; deep convolutional network; SIFT flow

1. INTRODUCTION

Automatic image or scene understanding is essential to many computer vision tasks. The goal of scene parsing is to assign a semantic label to every pixel in a query image. Recently, two different mechanisms of scene parsing have been popularly discussed: parametric and nonparametric methods. Parametric methods [1][2] mainly rely on pre-trained models to classify each pixel or region to a certain class. Although these methods may achieve impressive results, the performance highly depends on the pre-trained classifiers. Moreover, as new classes as well as new training data are increasingly generated, it is impossible to predefine every potential class. Parametric methods are thus not easy to adapt to dynamically changing or the open universe datasets. On the other hand, nonparametric methods [3]-[6] do not pre-define the semantic labels and mainly rely on transferring labels from semantically-similar images. Given an input or query image, an efficient visual search is first executed to collect a set of similar images from the annotated training set. The retrieved images then define the semantic context for the query image. Different strategies have been developed to transfer the semantic labels from the retrieval set to the query image. The major advantage of nonparametric methods is that no batch training process is needed to adapt to open universe datasets.

Several promising nonparametric methods have been developed recently. Liu *et al*. [3] proposed to transfer labels at pixel level by SIFT flow matching. As one single pixel alone carries little semantic information, a superpixel-based method is proposed in [4] and is further extended in [2] to refine the parsing results by combining per-exemplar detectors. These per-exemplar detectors need additional off-line training and more computational cost in the query stage. In [5], a superpixel-based method is proposed but focuses on rare object classes by expanding the retrieval set for rare classes. In [6], considering that superpixels represent only fragmental objects and that these is no widely accepted descriptors for superpixels, Tung *et al.* proposed to transfer labels through content-adaptive windows. The windows are evaluated by "objectness" [7] and are expected to better capture objects than superpixels. A second iteration of semantic retrieval is further adopted in [6] to include similarly labeled images into the retrieval set. The unary potential for each query window is computed in terms of the similarity between the query and its matched windows. However, although the windows may capture similar objects and contextual information, they usually have different spatial layout. This spatial misalignment between matched windows may degrade the following label inference process.

In this paper, we focus on improving the quality of retrieved images/windows by using descriptors directly learned from raw images. We also propose using dense alignment to transfer accurate semantic context provided by the matched windows. Note that, the goal of this paper is to show that, once the two issues are carefully addressed, nonparametric methods alone can achieve comparable performance with parametric methods. The overall performance can be further improved once other parametric models (e.g., per-exemplar detectors [2] or rare-class

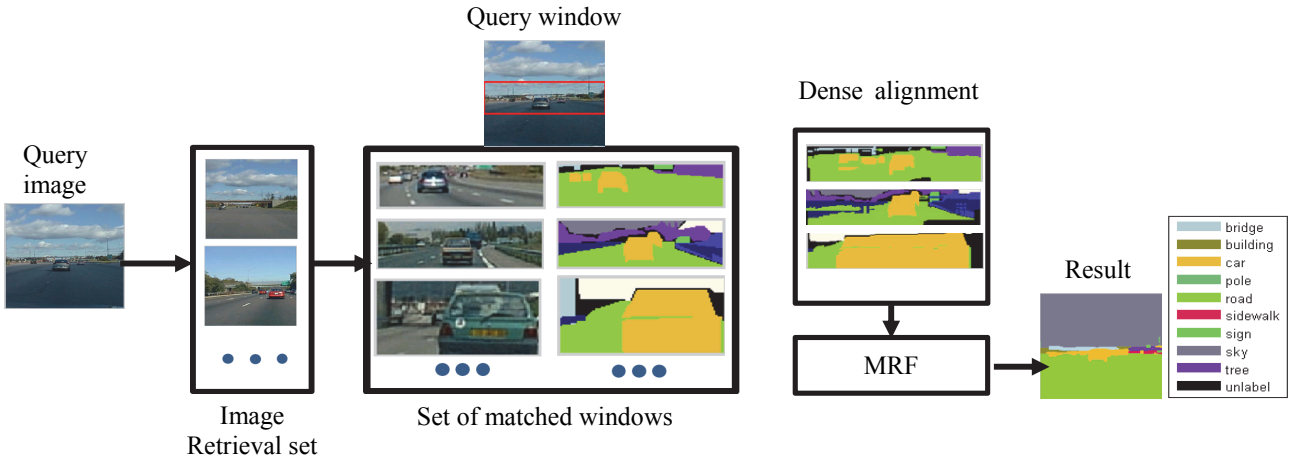


Figure 1. Overview of the proposed method.

exemplar [5]) are available or a second iteration of semantic retrieval [6] is included.

2. BACKGROUND AND MOTIVATION

We first summarize some observations that commonly concern the performance of nonparametric methods.

(1) The quality of image retrieval is critical to the whole parsing process. The image retrieval set should be "semantically similar" to the query image and define all the scene labels for the query.

(2) There are usually more than one objects in an image. Recent work represents the query and retrieved images using windows (or superpixels) to locate objects (or fragmental objects). Each query window is then matched with retrieval windows to determine the regional correspondence. This step can be seen as decomposing the original scene parsing problem into a set of subproblems; each subproblem deals with the label transfer between the query window and matched windows. Again, the quality of semantic similarity for each query window is essential to this stage; that is, the retrieved windows should now be more specific to capture the local structure at approximately the same scale.

(3) Once the semantic context per window is available, the next step is determining semantic label at pixel level. A typical approach is computing a likelihood score for each class and then a Markov Random Field (MRF) framework is followed to incorporating neighboring scores to compute the labeling of the query image.

From the above discussion, we explore several strategies to leverage nonparametric methods through improving the quality of retrieved images/matched windows, and enforcing dense alignment into the MRF framework. The overview of the proposed method is illustrated in Figure 1. Details of our nonparametric scene parsing approach are given in the following section.

3. PROPOSED METHOD

3.1. Image Retrieval Using Deep Convolutional Features

Recent work mostly forms the retrieval set by global search using the hand-crafted features (such as GIST [8] and color histogram used in [4], SIFT [9] used in [5], or GIST and HOG [10] visual words in [6]). Inspired by the success of convolutional networks (CNNs) [11] on many recognition tasks, we propose to adopt the deep convolutional features to form the retrieval set. In contrast to the hand-crafted features, CNN features are directly learned from raw inputs and have been shown to be the most accurate method for generic object classification.

Given a query image I_q, we first extract the 7-th layer output of Caffe [11] as the 4096-dimensional feature vector to find the top K similar images from the training database. Similar to previous work, we rank the training images in increasing order of Euclidean distance from the query.

Although there is no guarantee that the retrieved images are "semantically similar" to the query, with the CNN features, we can achieve similar performance as previous work with reduced size of retrieval set. For example, as shown in Table 1, we reduce the size of K from 200 [4] and 400 [6] to 40 in our experiment and can still achieve competitive parsing accuracy.

3.2. Detection and Matching of Object Windows

As discussed in [6], content-adaptive windows are advantageous over superpixels for locating objects. We thus propose to detect and transfer labels between matched windows. We adopt the selective search algorithm [12], whose results shows its superiority over the objectness [7],

(a) (b)

Figure 2. (a) The query image and its object windows located by the selective search algorithm [12]; and (b) the top 4 matched windows of the query window (marked by the red rectangle in (a)).

to locate a small set of data-driven and class-independent object windows for each image. Next, for each query window w_q, we search the top κ matched windows from retrieval images. Here, instead of using the hand-crafted features (e.g., HOG features are used in [6]), we again extract the 4096-dimensional Caffe feature [11] to measure the similarity between windows.

In addition, observing that the same object may be located by more than one windows, we allow no more than one window be selected from the same retrieval image. In other words, for each query window w_q, all the κ matched windows must come from κ different training images. An example is given in Figure 2.

3.3. Dense Alignment via SIFT Flow

Once the set of matched windows for each query window is determined, a standard MRF framework is usually defined over the field of labels to derive the parsing results. As we discussed in Sec. 2, because the spatial layout between matched windows can be very different (as shown in Fig. 2, where the location and size of the target object "car" are very different), additional alignment should be included to improve the dense correspondence between windows. In [6], the labeling is refined within a superpixel; that is, pixels within a superpixel are assigned the same label. However, this alignment heavily rely on the setting and implementation of superpixels.

In this work, we argue that a dense correspondence in pixel level between matched windows is essential to the MRF inference framework. Inspired by [3], where SIFT flow algorithm has been shown to be able to establish semantically meaningful correspondences among images, we proposed to follow the coarse-to-fine SIFT flow matching scheme [3] to establish dense correspondence

(a) (b) (c)

Figure 3. Visualization of the class label "car". (a) The ground truth; (b)(c) The results obtained without and with dense alignment, respectively.

between the query window and each of its matched windows.

For a query window w_q with its Caffe feature descriptor $\boldsymbol{s}_q$, we first resize all its matched windows W_{rs}^q to the dimensions of w_q. Let $\widetilde{W}_{rs}^q$ denote the resized set of W_{rs}^q. We then compute the SIFT flow from w_q to each window in $\widetilde{W}_{rs}^q$. For details of SIFT flow calculation, please refer to [3].

After the dense alignment, for a query window w_q, we have the matched set of windows $\widetilde{W}_{rs}^q = \{\boldsymbol{c}_i, \mathbf{f}_i | i = 1, \dots, \kappa\}$ where $\boldsymbol{c}_i$ and $\mathbf{f}_i$ are the class labels and SIFT flow fields of the ith matched window. Then, we compute the unary potential by transferring the labels from $\widetilde{W}_{rs}^q$ to w_q. Let $\psi(\mathrm{c}, \boldsymbol{p})$ denote the unary potential associated with a semantic label of class c to a pixel $\boldsymbol{p}$.

$$\psi(\mathrm{c}, \mathbf{p}) = -\sum_{w_q \in W_q} \sum_{w_i \in W_{rs}^q} \delta\left[L\left(\widetilde{w}_i(\boldsymbol{p} + \mathbf{f}_i), \boldsymbol{p} - \mathrm{offset}(w_q)\right) = \mathrm{c}\right] \cdot \phi_{size}(w_i) \cdot \phi_{idf}(\mathrm{c}), \quad (1)$$

where W_q is the set of windows in the query image, w_i is the matched window of w_q in the set W_{rs}^q, $\widetilde{w}_i$ is the resized version of w_i, $L(.,.)$ transfers the label from the window $\widetilde{w}_i$ to w_q. The term $\boldsymbol{p} - \mathrm{offset}(w_q)$ is defined as in [6] to give the window-centric coordinates of the pixel $\boldsymbol{p}$ in window w_q; thus $L\left(\widetilde{w}_i(\boldsymbol{p} + \mathbf{f}_i), \boldsymbol{p} - \mathrm{offset}(w_q)\right)$ gives the label of the SIFT-flow-aligned pixel of $\boldsymbol{p}$ in $\widetilde{w}_i$ to w_q.

In our simulation, we observe that larger object windows usually contain complex contextual background; whereas smaller windows better locate the object with tight bounding boxes. Therefore, we include a term $\phi_{size}(w_i)$ in (1) to penalize large retrieval windows:

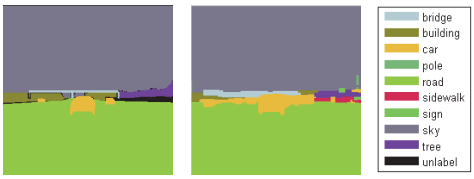
$$\phi_{size}(w_i) = \frac{1}{N(w_i)}, \quad (2)$$

where $N(w_i)$ denotes the number of pixels in the retrieval window w_i.

In (1), the term $\phi_{idf}(\mathrm{c})$ is similarly defined as in [6] to reflect the rareness of class label c in the retrieval set:

$$\phi_{idf}(\mathrm{c}) = \frac{1}{N(\mathrm{c})^{\gamma}}, \quad (3)$$

where $N(\mathrm{c})$ denotes the number of pixels of class c in the image retrieval set, and γ is a constant used to control the



(a) (b)

Figure 4. (a) Ground truth of the query image in Fig. 2 (a); and (b) the predicted labels of our approach.

strength of this term. With $\phi_{idf}(\mathrm{c})$, a large weight is assigned to the less frequent classes.

3.4. MRF Labeling

Next, we build a Markov Random Field model to combine the unary potential with a pairwise potential to derive the final parsing result. The MRF energy function over the field of labels $\boldsymbol{c} = \{c(\boldsymbol{p})|\boldsymbol{p} \in I_q\}$ is defined as

$$E(\mathbf{c}) = -\sum_{\mathbf{p}\in I_q}\psi(\mathbf{c},\mathbf{p}) + \lambda\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon}\theta(c(\mathbf{p}),c(\mathbf{q})), \quad (4)$$

where ε defines the set of adjacent pixels and λ is a smoothing constant. The pairwise potential term θ is similarly defined as in [4][6] according to the probabilities of label co-occurrences in the training set:

$$\theta\big(c(\mathbf{p}),c(\mathbf{q})\big) = -\log[(P\big(c(\mathbf{p})\big|c(\mathbf{q})\big) + (P\big(c(\mathbf{q})\big|c(\mathbf{p})\big)/2)] \cdot \delta[c(\mathbf{p}) \neq c(\mathbf{q})], \quad (5)$$

where $P\big(c(\mathbf{p})\big|c(\mathbf{q})\big)$ is the conditional probability of the class label $c(\mathbf{p})$ of the pixel $\mathbf{p}$ given that its adjacent pixel $\mathbf{q}$ has label $c(\mathbf{q})$.

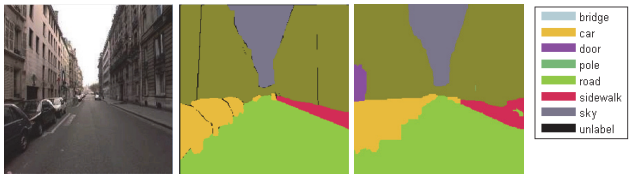
Finally, we use the alpha-beta swap algorithm [13] to minimize the MRF energy function and obtain the final parsing result.

4. EXPERIMENTAL RESULTS

We conduct experiments on the SIFT Flow dataset [3] and compare with existing methods [2]-[6]. The dataset contains 2488 training images and 200 testing images; all the images are of size 256×256 and the pixels are labeled with 33 semantic classes. In all our experiments, we set the number of retrieved image $K = 40$, the number of matched windows $\kappa = 10$, and the parameter $\gamma = 0.38$.

Figure 3 shows an example that, when the dense alignment is included to the MRF framework, the inferred result (Fig. 3(c)) is greatly improved over the non-aligned case (Fig. 3(b)). With the dense alignment, the labeling result in Fig. 4 (b) preserve labeling accuracy in pixel level. Another example is given in Figure 5.

Table 1 summarizes the labeling accuracy of the proposed method and the state-of-the-art methods [2]-[6]. Note that, the two methods [2] and [5] are marked as parametric methods because they need additional training process for per-exemplar detector and rare-class exemplar, respectively. To have a fair comparison with nonparametric methods, we also include the experimental report of the baseline method of [5] in Table 1.



(a) (b) (c)

Figure 5. (a) Query image; (b) ground truth; and (c) the predicted labels of our approach.

From Table 1, even when we include no dense alignment, the proposed method already outperforms the other nonparametric methods in terms of per-pixel accuracy. This demonstrates the effectiveness of using deep convolutional features as visual scene descriptor. When we further include the dense alignment, the per-class accuracy improves and outperforms other nonparametric methods while moderately reducing the per-pixel accuracy. We believe that, the overall accuracy can be further improved once other parametric models (e.g., per-exemplar detectors [2] or rare-class exemplar [5]) are available or a second iteration of semantic retrieval [6] is included.

Table 1. Labeling accuracy on the SIFT Flow dataset.

	K	per-pixel	per-class
Parametric methods			
Tighe and Lazebnik [2]	200	78.6	39.2
Yang et al. [5]	40	**79.8**	**48.7**
Nonparametric methods			
Liu et al. [3]	85	76.7	-
Tighe and Lazebnik [4]	200	77	30.1
Yang et al. – Baseline [5]	40	78	27.5
CollageParsing [6]	400	77.1	41.1
Proposed method – w/o dense alignment	40	**78.5**	40.8
Proposed method	40	78.3	**46.1**

5. CONCLUSION

In this paper, we focus on improving both the semantic context and label transfer accuracy in nonparametric scene parsing methods. Unlike previous methods, without resorting to additional semantic retrieval or predefined models, we propose to use the state-of-the-art deep convolutional features as descriptors and include dense alignment to infer semantic label in pixel-level accuracy. In the experiments, we compare our results with existing methods and show improved performance over existing nonparametric methods while performing very competitively against parametric ones.

6. REFERENCES

[1] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Trans. PAMI*, vol. 35, no. 8, Aug. 2013, pp.1915-1929.

[2] J. Tighe and S. Lazebnik, "Finding Things: Image Parsing with Regions and Per-Exemplar Detectors," In *Proc. CVPR*, 2013.

[3] C. Liu, J. Yuen, and A. Torralba, "Nonparametric Scene Parsing via Label Transfer," *IEEE Trans. PAMI*, vol. 33, no. 12, Dec. 2011, pp. 2368-2382.

[4] J. Tighe and S. Lazebnik, "SuperParsing: Scalable Nonparametric Image Parsing with Superpixels," *Internationl Journal of Computer Vision*, vol. 101, no. 2, Jan. 2013, pp. 329-349.

[5] J. Yang, B. Price, S. Cohen and M.Yang. "Context Driven Scene Parsing with Attention to Rare Classes," In *Proc. CVPR*, 2014.

[6] F. Tung and J. J. Little, "CollageParsing: Nonparametric Scene Parsing by Adaptive Overlapping Windows." In *Proc. ECCV*, 2014.

[7] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the Objectness of Image Windows," *IEEE Trans. PAMI*, vol. 34, no. 11, Nov. 2012, pp. 2189-2202.

[8] A. Oliva, and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *International Journal of Computer Vision*, vol. 42, no. 3, Jan. 2001, pp. 145-175.

[9] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, Jan. 2004, pp. 91-110.

[10] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," In *Proc. CVPR*, 2005.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," http://caffe.berkeleyvision.org/, 2014.

[12] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers and A.W.M. Smeulders, "Selective Search for Object Recognition," *Internationl Journal of Computer Vision*, vol. 104, no. 2, April 2013, pp. 154-171.

[13] Y. Boykov, O. Veksler, and R. Zabin, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. PAMI*, vol. 23, no. 11, Nov. 2001, pp. 1222-1239.