Superpixel Correspondence for Non-parametric Scene Parsing of Natural Images

Veronica Naosekpam[1], Alexy Bhowmick[1(✉)], and Shyamanta M. Hazarika[2]

[1] School of Technology, Assam Don Bosco University, Guwahati, Assam, India
alexy.bhowmick@gmail.com
[2] Indian Institute of Technology, Guwahati, Assam, India

Abstract. Scene parsing refers to the task of labeling every pixel in an image with the class label it belongs to. In this paper, we propose a novel scalable non-parametric scene parsing system based on superpixels correspondence. The non-parametric approach requires almost no training and can scale up to datasets with thousands of labels. This involves retrieving a set of images similar to the query image, followed by superpixel matching of the query image with the retrieval set. Finally, our system warps the annotation results of superpixel matching, and integrates multiple cues in a Markov Random Field (MRF) to obtain an accurate segmentation of the query image. Our non-parametric scene parsing achieves promising results on the LabelMe Outdoor dataset. The system has limited parameters, and captures contextual information naturally in the retrieval and alignment procedure.

Keywords: Scene · Scene parsing · Non-parametric · Label transfer · Partial correspondence

1 Introduction

In the recent decade, scene parsing has become an active area of research. The idea behind scene parsing is to label each pixel in an image to the category or the class it belongs to. It can be said that scene parsing is one step ahead of the traditional image segmentation. Semantic labels can be *stuffs* such as - grass or sky, as well as *things*, such as - person or building [8]. Traditionally there two approaches of parsing an image: (1) Parametric and (2) Non-Parametric [13].

The traditional parametric approach consists of model training phase using the training set which is then tested using the test data [5]. Though this approach is effective and gives highly accurate result, it does not account for the dynamic nature of the world where the size of the data keep increasing dynamically. Parametric approach work with "closed universe" datasets where size of the data is fixed. Once the data size varies, the model has to be trained again which is an overhead. The non-parametric approach [2,8,13] provides several advantages

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 614–622, 2019.
https://doi.org/10.1007/978-3-030-34869-4_67

over the parametric approach. Instead of training the model, semantic labels are directly transfered to the query image from the retrieval set. It works with "Open universe" datasets whose size can vary at all time. Since it is a data-driven approach, it requires almost no training. In non-parametric scene parsing, the semantic label transfer can be done at pixel level (dense correspondence) [8] or the superpixel level (partial correspondence) [2,13].

A typical pipeline of non-parametric scene parsing is shown in Fig. 1. The image database comprises of training and test images and their corresponding annotations or the ground-truth labels. There are three major steps involved in non-parametric scene parsing: (1) *Scene Retrieval*: Given a query image, a set of similar images are retrieved from the training images; (2) *Scene Correspondence*: It is the alignment of the query image with the retrieved images (results from the scene retrieval); for correspondence of points to be used for labeling. And (3) *Markov Random Field* (MRF) *Inference*: It performs the task of spatial smoothing or aggregation of labels to obtain better segmentation results.

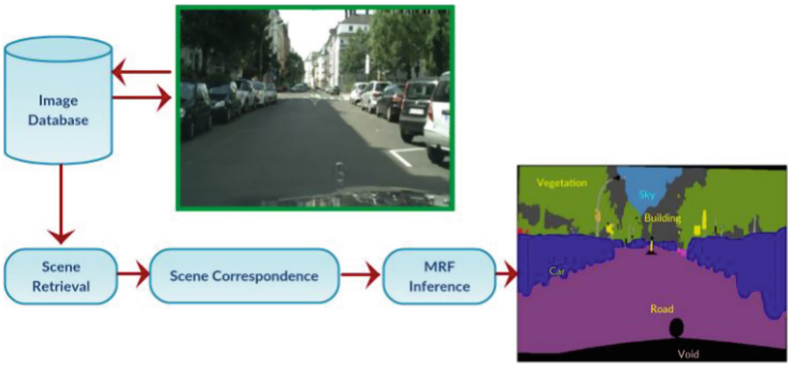

Fig. 1. A typical non-parametric scene parsing pipeline

In this work, our aim is to propose an improved context based non-parametric framework for the scene parsing problem. Our entire scene parsing pipeline heavily relies on the correct retrieval of the similar image set to achieve a correct labels of the query image. We aim to get a fairly similar set of images as a retrieval set of the query image in order to perform the parsing. Our scene retrieval module has been built by taking into account the spatial features of the query image. The label transfer is performed based on partial correspondence using super-pixels. An improved super-pixels features extraction techniques are used for correspondence in the partial scene correspondence module. Regarding the MRF inference module, off-the-shelf graph based MRF inferencing technique has been used. We demonstrate our system on SIFT Flow (LabelMe Outdoor) dataset. Experimental evaluation is done based on per-pixel and per-class accuracy.

The remainder of the paper is as follows: Sect. 2 starts with the summarization of related works. It is followed by a presentation of the improved framework

and experimental evaluation in Sects. 3 and 4. Finally, conclusions are drawn and outlook on future work is presented in Sect. 5.

2 Related Works

Various literature on non-parametric scene parsing have been studied [2,8,13]. For label transfer from the retrieved images to the query images, the existing state-of-the-art approaches differ in the methods of retrieval of similar images, the nature of correspondence algorithms, and the choice of the features used.

The earliest works on non-parametric scene parsing are based on dense correspondence [8]. Dense correspondence is a pixel-by-pixel alignment of the query image with the retrieval set using dense local features such as SIFT. The per-pixel local information are transferred from an image in the retrieval set to the query image. Liu et al. [8] introduced the concept of "label transfer" based on local dense SIFT Flow matching algorithm [9]. ANN (Approximate Nearest Neighbour) bilateral matching transfers label at pixel level by integrating prior knowledge based on local partial similarity between images [17]. Collageparsing [15] is an approach that extracts mid-level content adaptive windows from the retrieved images and the query image. The label transfer is performed by matching the query image's content adaptive windows with retrieval set images' content-adaptive windows. As research on dense scene correspondence continued, it is found that dense correspondence is complex and computationally expensive and hence, the recent research approaches on non-parametric scene segmentation are based on partial correspondence which is super-pixel or patch based approach.

In the non-parametric scene parsing based on partial correspondence, the query image is aligned to the images in the retrieval set by considering a group of pixels at a time. Partial correspondence labels cohesive groups of pixels together without loss of geometrical information of the image and in turn significantly reduces the number of elements to process. Label transfer based on partial correspondence was first introduced by Tighe and Lazebnik [13]. The superpixels of the query image are labelled using an MRF model, based on similar superpixels in the query's nearest neighbor. It also performs simultaneous labeling of geometric and semantic classes of the query image. Their work is extended by Eigen et al. [2] where per-descriptor weights for each superpixels' segment are learnt in order to minimize classification error which in turn helps to nullify the biasing of the semantic classes towards common classes.

In order to detect interesting objects in a scene, Tighe and Lazebnik [14] proposed to augment their SuperParsing work with pre-trained exemplar SVMs [10]. Although the overall per-pixel accuracy improved, it failed to detect rare classes. Yang et al. [16] added rare classes in the retrieval set for a more balanced super-pixels classification in a feedback manner in order to refine the matching at the super-pixel level. The super-pixels in the rare classes are populated using exemplars. Various classifiers (ensemble approach) are combined to perform superpixels based scene correspondence in [4]. The likelihood scores of

various classifiers are combined in order to label each superpixel of the query image with a balanced score. Label transfer via efficient filtering [11] performed parsing by first sampling of superpixels based on some similarity score and transfering labels via filtering using a Gaussian kernel which encodes how similar two superpixels are in terms of their feature vector.

3 Non-parametric Scene Parsing

3.1 Scene Retrieval

The scene retrieval component aims to find a subset of the training set similar to the query image. It is a critical step in the pipeline as the overall scene parsing performance is highly dependent on correct scene retrieval. There is no chance of recovery in later stages in case of an incorrect retrieval. We used a combination of local and global features for image retrieval. The global feature used for obtaining the retrieval set is Gist. The local features used to obtain the retrieval set are VLAD (Vector of Locally Aggregated Descriptors) [6] and SPM (Spatial Pyramid Matching) [7]. Gist is low dimensional global feature which gives the summary of an image. VLAD is an extension of Bag of Visual Words which gives a more discriminative representation. SPM incorporates the spatial information of an image. The approximate nearest neighbor images for the retrieval set are obtained using KD-tree indexing. Our improved scene retrieval module performs better retrieval of the similar images. For each of the three mentioned features, the first three most similar images are selected as the content of the retrieval set. Hence, the total number of images in the retrieval set is $k = 9$ for each query image.

3.2 Partial Scene Correspondence

The approach followed for scene alignment is the partial scene correspondence using superpixels. It labels large, cohesive groups of pixels at once, decreases the number of elements to process and at the same time, preserves the geometric information. Superpixels are segmented from the query image and the retrieval set using fast graph based segmentation [3]. The advantage of using this approach is that the it segments a scene or an image into a combination of coarse and fine regions. The scale of segmentation for the query as well as for the retrieval set is set to 200.

Five types of features are used for representing each super-pixel, (1) SIFT histogram (1024D), RGB color histogram (128D), HSV color histogram (128D), Location histogram (36D) and PHOG histogram of the superpixel's bounding box (168D). For SIFT histogram, SIFT features are extracted and encoded by 5 words from a vocabulary of size 1024 using LLC algorithm. 128-D color histogram are obtained by quantizing color features from a vocabulary of 128 words. 36-D location histogram is obtained by quantizing (x, y) location into a 6×6 grid. In order to include contextual information, the superpixels are masked by

20 pixels and the same type of features are used for representing the dilated superpixels. So, the total number of features for each superpixel is $(1024+128+128+36+168)\times 2 = 2968$. Then, similar to [16], the classification cost of each input superpixel $s_i \in Q$ with reference to the K-Nearest Neighbour ($N_k(i)$) in Retrieval Set R (s_j, x_j, y_j) using Kernel $K(x_i, x_j)$ is given by:

$$P(y_i = c|s_i) = 1 - \frac{\sum_{j\in N_k(i), y_j} K(x_i, x_j)}{\sum_{j\in N_k(i)} K(x_i, x_j)} \tag{1}$$

Kernel functions are real valued functions that quantifies the similarity between two feature spaces. They map data into higher dimensional space. $K(x_i, x_j) > 0$ is the similarity of x_i and $x_j \in X$. To avoid operating in the high dimensional space, a feature space is chosen in which the dot product can be computed directly using a nonlinear function in the input space $K(x_i, x_j) = <\theta(x_i), \theta(x_j)>$ (called the kernel trick).

The kernel function is the Chi-square kernel. It is a kernel based on Chi-square distribution.

$$K(x_i, x_j) = 1 - \sum_{i,j=1}^{N} \frac{(x_i - x_j)}{\frac{1}{2}(x_i + x_j)} \tag{2}$$

3.3 MRF Inference

In order to fuse global semantic constraints, a random field model is used. For this, a four connected MRF (Markov Random Field) is chosen whose energy function is given by the equation:

$$E(Y) = \sum_{p} E_{data}(y_p) + \lambda \sum_{pq} E_{smooth}(y_p, y_q) \tag{3}$$

where p and q are pixels, λ is the pairwise energy weight. The data term of one pixel is given by the result of Eq. 1 and the smoothness cost is given by label variant cost as $E_{smooth}(y_p, y_q) = d(p, q).\mu(c, c')$. d(p,q) is the color dissimilarity between two neighbouring pixels and $\mu(c, c')$ is the penalty of assigning c and c' to adjacent pixels. The objective is to find the labeling that optimizes the energy function.

For semantic labeling, inferencing on E(Y) is performed via Alpha-Beta Swap [1]. It is a graph-cut algorithm that minimizes or optimizes the E(Y) by exchanging labels between an arbitrary set of pixels labeled α and another arbitrary set of pixels labeled β.

4 Experiments

Here we evaluate and compare our method with state-of-the-art approaches on SIFT Flow (LabelMe Outdoor) dataset. It consists of 2688 outdoor scene images

with 2488 images as training set and 200 images as test set as in [8]. The images are 256 × 256 pixels with 33 labels. The metric used for evaluation are: (1) *Per-pixel accuracy*: The percent of correctly labeled pixels in total and (2) *Per-class accuracy*: The average percent of correctly labeled pixels in each class.

The average per-pixel recognition rate r_p is calculated as

$$r_p = \frac{1}{\Sigma_i m_i} \Sigma_i \Sigma_{p \in \Lambda_i} 1(o(p) = a(p), a(p) > 0), \tag{4}$$

where, for a pixel p in image i, the ground-truth label is a(p) and system output is o(p); for unlabeled pixels, a(p) = 0. The symbol Λ_i represents the image lattice for test image i and m = $\Sigma_{p \in \Lambda_i} 1(a(p) > 0)$ is the number of labeled pixels for image i. It is to be noted that some pixels may remain unlabeled.

The average per-class recognition rate r_c is

$$r_c = \frac{\Sigma_i \Sigma_{p \in \Lambda_i 1(o(p)=a(p), a(p)=l)}}{\Sigma_i \Sigma_{p \in \Lambda_i 1(a(p)=l)}}, l = 1, 2, 3, ...L \tag{5}$$

For each query, the number of retrieval set is set to 9. Parameters for superpixels segmentation are set to: k = 200, sigma = 0.8 and minimum-area = 25. The nearest neighbour (K) superpixels for label transfer is set to 100. The MRF parameters are set to: $\lambda = 6$ and $\alpha = 0.7$. It has been identified that 5 of the 33 labels are common classes while the rest of the labels are rare classes. Our result is compared with recent works in Table 1.

Table 1. Comparison with the state-of-the-art systems on LMO dataset

	Per-pixel	Per-class
Liu et al. [8]	74.75%	N/A
Tighe et al. [13]	73.2%	29.1%
Gould et al. [5]	65.7%	14.2%
Razzaghi et al. [12]	75.84%	31.3%
Eigen et al. [2]	75.3%	39.2%
Our approach	73.5%	21.719%

The class distribution graph of the labels is shown in Fig. 2. From the plot, it is been shown that building, sky, road, mountain, sea and tree are the most common class distribution. It also shows that the system failed to detect rare classes such as desert, bus, balcony, bird, cow, pole, moon and sun. Some of our labeling results along with the ground truth is shown in Fig. 3.

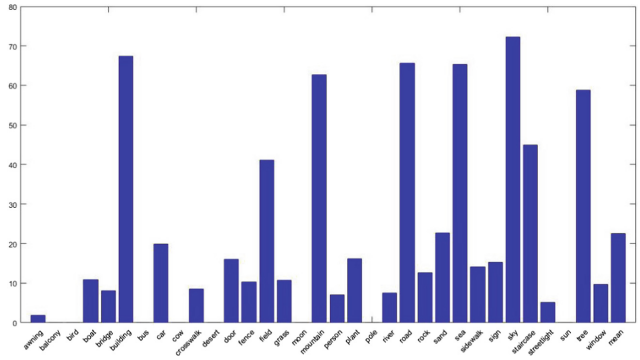

Fig. 2. Classes of objects in the dataset and their frequency.

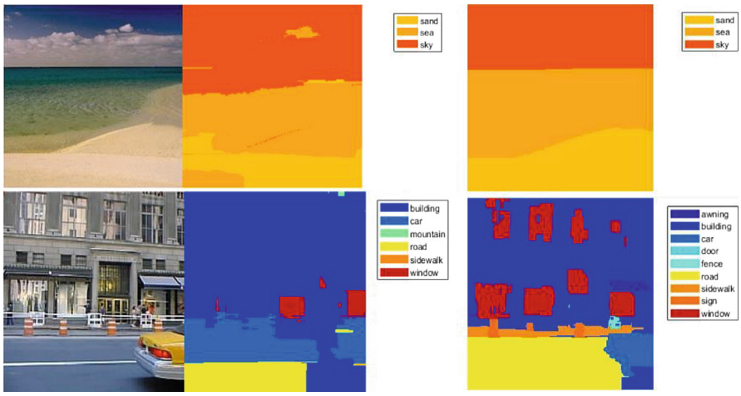

Fig. 3. Some results on SIFT flow dataset. On the left are query images, and results of scene parsing. On the right are ground truth images for comparison.

5 Conclusion

In this work, we present a novel non-parametric scene parsing framework whose overall per-pixel labeling accuracy is improved by accurate retrieval and partial correspondence. By combining various features for similar image retrieval, we have boosted the strength of the partial correspondence matching. Evaluation results have shown that our system obtains compatible performance on SIFT Flow dataset. Dividing the query image and retrieval set into non-uniform grids based on salient region detection and performing partial correspondence on corresponding grids will be a good research direction.

References

1. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, pp. 377–384. IEEE (1999)
2. Eigen, D., Fergus, R.: Nonparametric image parsing using adaptive neighbor sets. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2799–2806. IEEE (2012)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vis. **59**(2), 167–181 (2004)
4. George, M.: Image parsing with a wide range of classes and scene-level context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3622–3630 (2015)
5. Gould, S., Zhang, Y.: PatchMatchGraph: building a graph of dense patch correspondences for label transfer. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 439–452. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_32
6. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: CVPR 2010–23rd IEEE Conference on Computer Vision and Pattern Recognition, pp. 3304–3311. IEEE Computer Society (2010)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 2169–2178. IEEE (2006)
8. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing via label transfer. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2368–2382 (2011)
9. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT flow: dense correspondence across different scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_3
10. Malisiewicz, T., Gupta, A., Efros, A.A., et al.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV, vol. 1, p. 6. Citeseer (2011)
11. Najafi, M., Taghavi Namin, S., Salzmann, M., Petersson, L.: Sample and filter: nonparametric scene parsing via efficient filtering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 607–615 (2016)
12. Razzaghi, P., Samavi, S.: A new fast approach to nonparametric scene parsing. Pattern Recogn. Lett. **42**, 56–64 (2014)
13. Tighe, J., Lazebnik, S.: SuperParsing: scalable nonparametric image parsing with superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_26
14. Tighe, J., Lazebnik, S.: Finding things: Image parsing with regions and per-exemplar detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3001–3008 (2013)
15. Tung, F., Little, J.J.: CollageParsing: nonparametric scene parsing by adaptive overlapping windows. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 511–525. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_33

16. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3294–3301 (2014)
17. Zhang, H., Fang, T., Chen, X., Zhao, Q., Quan, L.: Partial similarity based non-parametric scene parsing in certain environment. In: CVPR 2011, pp. 2241–2248. IEEE (2011)