Scene Parsing Using Scene Attributes As Global Features

Hang Su
Department of Computer Science
Brown University
suhangpro@gmail.com

May 13, 2013

Abstract

Data-driven methods have been proven very effective for the task of scene parsing. A crucial step in these methods is to retrieve a set of visually similar scenes from existing image collections for the query image according to certain global scene representations. In this work, we incorporate scene attributes into data-driven scene parsing systems as global scene features. We show that when used as global features, our compact attribute-based scene representation can compete with or improve on traditional low-level scene representations for the task of scene parsing and scene retrieval in general.

1 Introduction

Scene parsing is the task of segmenting all the objects in a natural image and identifying their categories. Categorical labels can be given to either each pixel or each region (e.g. superpixel) of the input image, giving a thorough interpretation of the scene content. Most methods proposed for this problem require a generative or discriminative model to be trained for each category, and thus only work with a handful of pre-defined categories [2, 3, 4, 5, 8, 11, 13, 14, 15]. The training process can be very time-consuming and must be done in advance. Even worse, the entire training has to be repeated whenever new training images or class labels are added to the dataset. Recently, several nonparametric, data-driven approaches have been proposed for the scene parsing problem [7, 16, 1]. These approaches require no training in advance. They can easily scale to hundreds of categories

Figure 1: Sample outputs of scene attribute detection. (from [10])

and have the potential to work with internet-scale, continuously growing datasets like LabelMe [12].

There are low-level representations and high-level representations (e.g. attributes or categories) of natural scenes. While much research has been done on various low-level representations, such as the gist descriptor [9] or spatial pyramid [6], less attention has been given to high-level scene representations and their applications for data-driven vision tasks. Compared with other high-level representations, scene attributes keep the benefit of being compact and carrying sementic meanings, while giving more flexible and comprehensive interpretations to natural scenes. We adopt scene attributes designed in [10], which have 102 discriminative attributes discovered and learned from crowdsourcing. Figure 1 shows some sample outputs of the attribute detector provided in [10].

In this paper we show how well we can improve nonparametric, data-driven scene parsing by adopting scene attributes. Tighe and Lazebnik investigate nonparametric, data-driven scene parsing and achieve state-of-the-art performance [16]. We follow their system pipeline (section 2) and show that by simply adding scene attributes as one of the features used for global scene representation we can achieve significant performance improvement (section 3).

2 System Pipeline

The following is a summary of the steps taken by the parsing system for every query image (Figure 2).

Retrieval Set. The first step in parsing a query image is to find a retrieval set of images similar to the query image. The purpose of finding such a subset of training images (there is actually no training process, though we

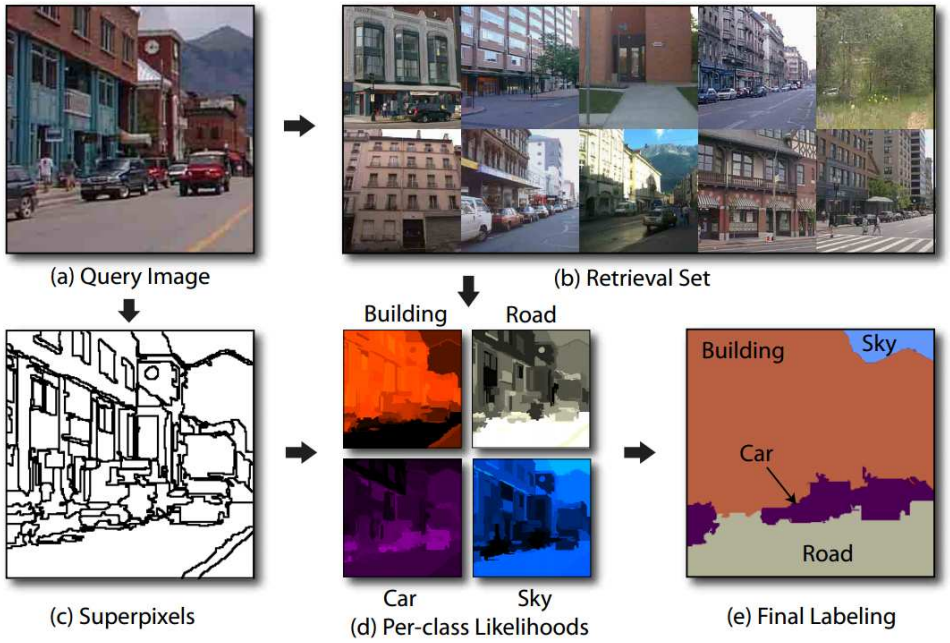

Figure 2: System pipeline of scene parsing. (from [16])

still call the images from which we try to learn the "training images") is to expedite the parsing process and at the same time throw away irrelevant information which otherwise can be confusing. In [16], three types of global image features are used in this step: gist, spatial pyramid, and color histogram. For each feature type, Tighe and Lazebnik sort all the training images in increasing order of Euclidean distance from the query image. They take the minimum rank accross all feature types for each training image and then sort the minimum ranks in increasing order to get a ranking among the training images for the query image. The top ranking K images are used as the retrieval set.

Local Superpixel Labeling. After building the retrieval set, the query image and the images in retrieval set are segmented into superpixels. Each superpixel is then described using 20 different features. A detailed list of these features can be found in Table 1 in [16]. For each superpixel in the query image, nearest-neighbor superpixels in the retrieval set are found according to the 20 features for that superpixel. A likelihood score is then computed for each class based on the nearest-neighbor matches.

Classification. In the last step, we can simply assign the class with

the highest likelihood score to each superpixel in the query image, or use Markov Random Field (MRF) framework to further incorporate pairwise co-occurrence information learned from training dataset. As in [1], we report the performance without using the MRF layer in this paper so differences in local classification performance can be observed more clearly.

3 Scene Attributes As Global Features

Our main goal in investigating scene parsing is to see how well our scene attributes work as a scene representation. Thus, we keep most parts of the system in [16] unchanged but use the scene attributes as the global feature or one of the global features in addition to other low-level features for finding retrieval sets.

The dataset we use for this experiment is the SIFT-Flow dataset [7]. It is composed of 2,688 annotated images from LabelMe and has 33 semantic labels. Since the class frequencies are highly unbalanced, we report both per-pixel classification rate and per-class rate, which is the average of the per-pixel rates over all classes. We also report the performance of an "optimal retrieval set", which uses ground-truth class labels instead of global features to find similar scenes for the query image. This retrieval set is called Maximum Histogram Intersection. It is found by ranking training images according to the class histogram intersections they have with the query image:

$$\cap(Target, Query) = \frac{\sum_{j=1}^{33} min(H_T[j], H_Q[j])}{\sum_{j=1}^{33} H_Q[j]}$$

where H_T and H_Q are the histograms of target image and query image respectively.

This optimal retrieval set is meant to be a performance upper bound and should provide an insight into how much room for improvement there still is in the image retrieval step. In [16], Tighe and Lazebnik proposed a different type of "optimal" retrieval set by ranking training images in terms of the number of pixels their ground truth label maps share with the label map of the query. Our experiment shows ours is uaually better in terms of both per-pixel rates and per-class rates.

Figure 3 and Figure 4 show the performance comparison among different global features. As we can see from the result, using only scene attributes as global features we get higher per-pixel rates than [16], which uses three global features (G+SP+CH), while getting similar per-class rates.

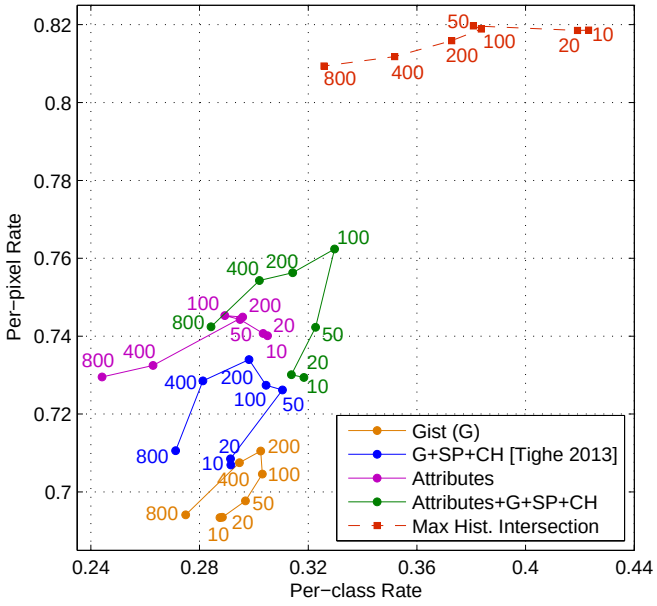

Figure 3: **Evaluation of using our scene attributes as a global feature for scene parsing on the SIFT-Flow dataset.** x-axis represents mean per-class classification rate, y-axis represents per-pixel classification rate. The best performance sits on the top-right corner of the space. The plots also show the impact of changing retrieval set size K. The blue plot shows the result of using gist (G), spatial pyramid (SP), and color histogram (CH) together as scene descriptors for finding retrieval sets [16]. Using scene attributes itself improves the per-pixel rates while the per-class rates are close. Using scene attributes together with the previous three features increases both the per-pixel rates and the per-class rates. "Maximum Histogram Intersection" is the upper bound we get by finding retrieval set using ground-truth labels of the query image.

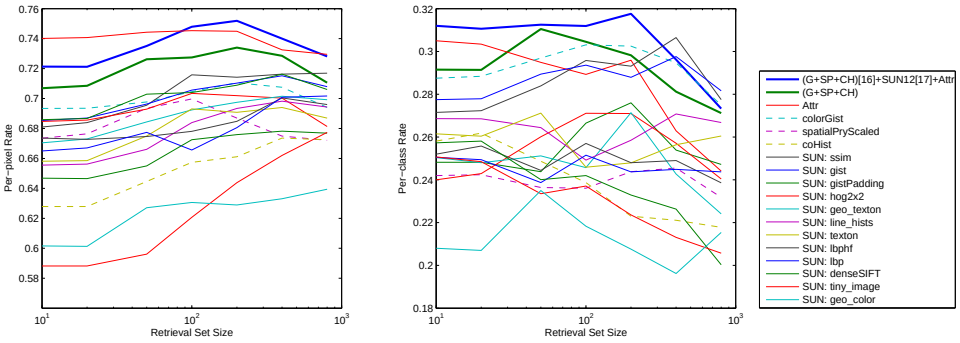

Figure 4: **Comparision of using various global features for scene parsing.** The left figure shows per-pixel rates and the right one shows per-class rates. Both the twelve features described in [17] and the three features used in [16] (G, SP, CH) are tried seperatly, as well as our scene attributes. We also report the performance of using all features together (G+SP+CH+SUN12+Attr) and using the three features in [16] (G+SP+CH).

When combining our scene attributes with those three global features (Attributes+G+SP+CH), both the per-pixel rates and the per-class rates increase significantly (73.4%, 29.8% ($K = 200$) vs. 76.2%, 33.0% ($K = 100$)). Considering the compact size of our scene attributes, 102 dimensions compared with the 5184-dimension G+SP+CH, this result demonstrates the scene attributes' strong ability for high-level scene representation. It is also worth noting that adding more features beyond this point does not necessarily improve the performance. For instance, by using all the 12 features described in [17] together with the scene attributes, the per-pixel rate and the per-class rate drop to 74.6% and 30.4% respectively ($K = 100$).

4 Conclusion

Scene parsing provides much deeper understanding of scenes than traditional category-based recognition. We investigated the use of attribute-based representation as global features for scene parsing. These experiments show its capability as a compact yet rich representation, and suggest the possible uses of scene attributes for future data-driven computer vision tasks.

References

[1] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2799 –2806, june 2012.

[2] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1 –8, 29 2009-oct. 2 2009.

[3] Xuming He, R.S. Zemel, and M.A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–695 – II–702 Vol.2, june-2 july 2004.

[4] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75(1):151–172, October 2007.

[5] L'ubor Ladický, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip H. S. Torr. What, where and how many? combining object detectors and crfs. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 424–437, Berlin, Heidelberg, 2010. Springer-Verlag.

[6] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[7] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(12):2368 –2382, dec. 2011.

[8] T. Malisiewicz and A.A. Efros. Recognition by association via learning per-exemplar distances. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[9] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[10] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[12] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.

[13] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, june 2008.

[14] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the 9th European conference on Computer Vision - Volume Part I*, ECCV'06, pages 1–15, Berlin, Heidelberg, 2006. Springer-Verlag.

[15] Richard Socher, Cliff C Lin, Andrew Y Ng, and Christopher D Manning. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, volume 2, page 7, 2011.

[16] Joseph Tighe and Svetlana Lazebnik. Superparsing. *International Journal of Computer Vision*, 101:329–349, 2013.

[17] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.