CollageParsing: Nonparametric scene parsing by adaptive overlapping windows

Frederick Tung and James J. Little

Department of Computer Science, University of British Columbia
{ftung,little}@cs.ubc.ca

Abstract. Scene parsing is the problem of assigning a semantic label to every pixel in an image. Though an ambitious task, impressive advances have been made in recent years, in particular in scalable nonparametric techniques suitable for open-universe databases. This paper presents the CollageParsing algorithm for scalable nonparametric scene parsing. In contrast to common practice in recent nonparametric approaches, CollageParsing reasons about mid-level windows that are designed to capture entire objects, instead of low-level superpixels that tend to fragment objects. On a standard benchmark consisting of outdoor scenes from the LabelMe database, CollageParsing achieves state-of-the-art nonparametric scene parsing results with 7 to 11% higher average per-class accuracy than recent nonparametric approaches.

Keywords: image parsing, semantic segmentation, scene understanding

1 Introduction

Computer vision enables us to understand scenes at many different levels of abstraction. At the most abstract, we may be concerned with determining the general semantic category of a scene [19], [26], [32], [36], such as *forest* or *urban*. Alternatively, instead of assigning an abstract category label to the scene, we may be interested in describing the scene by its semantic attributes [10], [27], [28], such as *rugged*. This paper is concerned with understanding scenes at the pixel level. Scene parsing is the challenging problem of assigning a semantic label to every pixel in the image. Semantic labels can span both amorphous background categories such as *grass* or *sea* (sometimes referred to "stuff" in the literature [16]), as well as localized object categories such as *person* or *car* (sometimes referred to as "things").

In recent years, the growth of online image collections and the adoption of crowdsourcing methods for annotating datasets have led to an interest in developing scalable methods that are suitable for open-universe datasets [34]. An open-universe dataset is one that is continually changing as users contribute new images and annotations, such as LabelMe [29]. Nonparametric methods are particularly well suited to open-universe datasets since they are data driven and require no training. As the dataset expands, there is no need to continually retrain the category models. This paper describes the CollageParsing algorithm for scalable nonparametric scene parsing.

Current state-of-the-art nonparametric algorithms for scene parsing match superpixels in the query image with superpixels in contextually similar database images. An advantage of superpixel based parsing is the ability to label large, cohesive groups of pixels at once. However, while superpixel based techniques tend to effectively label large regions of background ("stuff") categories, they fare less well on object ("thing") categories. There are at least two reasons for this gap. First, superpixel features are not very discriminative for objects. State-of-the-art object recognition algorithms employ more discriminative HOG or SIFT based features. There is no widely accepted feature descriptor for superpixels; various low-level features are often combined heuristically. Second, superpixels tend to fragment objects. Conceptually, superpixel based techniques reason about pieces of objects and apply auxiliary techniques on top to combine these pieces in a principled way. For instance, semantic label co-occurrence probabilities are commonly incorporated via a Markov random field model [8], [21], [34].

CollageParsing addresses both of these issues through its use of mid-level, "content-adaptive" windows instead of low-level superpixels. Window selection is content-adaptive in the sense that it is designed to capture entire objects and not only fragments of objects (Section 3.2). Surrounding contextual information is also partially captured by the windows. To describe the content-adaptive windows, CollageParsing employs HOG features, which have been demonstrated to be effective for object recognition [6], [7], [12]. Figure 3, explained in more detail in Section 3.3, shows the intuition behind CollageParsing's window-based label transfer.

Parametric scene parsing methods have a small advantage in accuracy over nonparametric methods, however as a tradeoff they require large amounts of model training (for example, training just the per-exemplar detector component of the extended SuperParsing algorithm [33] on a dataset of 45,000 images requires four days on a 512-node cluster), making them less practical for open-universe datasets. As we show in the experiments, CollageParsing achieves state-of-the-art results among nonparametric scene parsing methods, and comparable performance with state-of-the-art parametric methods while not requiring expensive model training.

2 Related work

Analyzing a scene at the level of labelling individual pixels with their semantic category is an ambitious task, but recent years have seen impressive progress in this direction.

Heitz and Koller [16] developed a graphical model to improve the detection of objects ("things") by making use of local context. In this work, local context refers to "stuff" classes such as *road* or *sky*. The "Things and Stuff" graphical model comprises candidate detection windows, region (superpixel) features, and their spatial relationships. Approximate inference is performed using Gibbs sampling.

Liu et al. [21] proposed the nonparametric label transfer technique for scene parsing. Liu et al.'s approach takes as input a database of scenes annotated with semantic labels. Given a query image to be segmented and labelled, the algorithm finds the image's nearest neighbors in the database, warps the neighbors to the query image using SIFT Flow [22], and "transfers" the annotations from the neighbors to the query image using a Markov random field model to integrate multiple cues.

Tighe and Lazebnik's SuperParsing algorithm [34] for scene parsing takes a similar nonparametric approach but operates on the level of superpixels. The query image's superpixels are labelled using a Markov random field model, based on similar superpixels in the query's nearest neighbor images in the database. The nonparametric semantic class labelling is combined with additional parametric geometry classification (sky, vertical, horizontal) to improve labelling consistency. Eigen and Fergus [8] proposed two extensions to SuperParsing. First, weights are learned for each descriptor in the database in a supervised manner to reduce the influence of distractor superpixels. Second, to improve the labelling of rare classes, the retrieved set of neighbor superpixels is augmented with superpixels from rare classes with similar local context. Myeong et al. [25] applied link prediction techniques to superpixels extracted from the query image and its nearest neighbors to learn the pairwise potentials for Markov random field based superpixel labeling, similar to SuperParsing. Singh and Košecká [31] proposed a nonparametric superpixel based method in which a locally adaptive nearest neighbor technique is used to obtain neighboring superpixels. The authors also proposed refining the retrieval set of query image neighbors by comparing spatial pyramids of predicted labels.

Instead of computing the set of nearest neighbor images at query time, the PatchMatchGraph method of Gould and Zhang [14] builds offline a graph of patch correspondences across all database images. Patch correspondences are found using an extended version of the PatchMatch algorithm [2] with additional "move" types for directing the local search for correspondences.

Farabet et al. [9] developed a parametric scene parsing algorithm combining several deep learning techniques. Dense multi-scale features are computed at each pixel and input to a trained neural network to obtain feature maps. Feature maps are aggregated over regions in a hierarchical segmentation tree. Regions are classified using a second neural network and pixels are finally labelled by the ancestor region with the highest purity score.

Tighe and Lazebnik [33] recently extended the SuperParsing algorithm with per-exemplar detectors (Exemplar-SVMs [23]). The data term based on superpixel matching is the same as in the SuperParsing algorithm. A detector based data term is obtained by running the per-exemplar detectors of class instances found in the retrieval set and accumulating a weighted sum of the detection masks. The two data terms are input to another trained SVM to obtain the class score for a pixel, and the final smooth class prediction is determined using a Markov random field.

Isola and Liu [18] proposed a "scene collage" model and explored applications in image editing, random scene synthesis, and image-to-anaglyph. In contrast to CollageParsing, the scene collage is an image representation: it represents an image by layers of warped segments from a dictionary.

3 Algorithm description

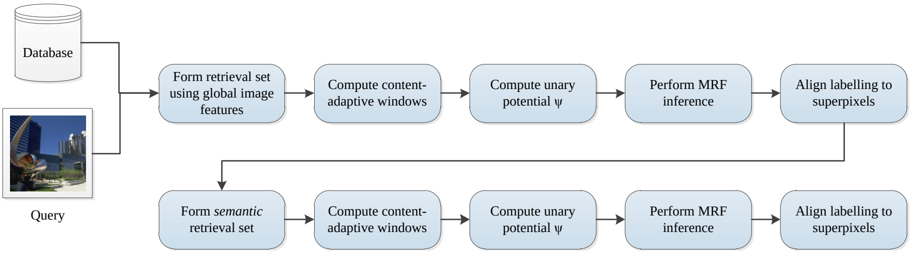
Database
Query
Form retrieval set using global image features
Compute content-adaptive windows
Compute unary potential ψ
Perform MRF inference
Align labelling to superpixels
Form *semantic* retrieval set
Compute content-adaptive windows
Compute unary potential ψ
Perform MRF inference
Align labelling to superpixels

Fig. 1. High-level overview of the CollageParsing pipeline. A query image is labelled in two iterations of the pipeline. The first iteration computes a retrieval set from global image features (Gist and HOG visual words) and outputs an initial image labelling that is then used to produce a semantic retrieval set for the second iteration.

Figure 1 shows a high-level overview of the CollageParsing pipeline. Given a database of images and a query image, the algorithm first finds the query image's nearest neighbors in the database according to a global image similarity measure (Section 3.1). The resulting short list of database images is referred to as the retrieval set. Content-adaptive windows are then extracted from the query image (Section 3.2). The query image's windows are matched with the content-adaptive windows in the retrieval set to compute a unary potential (or energy) for labelling each pixel with each semantic category (Section 3.3). The unary potential is combined with a pairwise potential in a Markov random field to obtain an initial labelling, which is refined by aligning the labelling to the query image's superpixels (Section 3.4). Finally, the previous steps are repeated with a *semantic* retrieval set consisting of similarly labelled images (Section 3.5).

3.1 Forming the retrieval set

The retrieval set aims to find a subset of database images that are contextually similar to the query image, and is a typical component of nonparametric scene analysis methods [15], [21], [34]. In addition to filtering out semantically irrelevant database images that are likely to be unhelpful, a small retrieval set makes nearest neighbor based label transfer practical on large datasets. To form the retrieval set, CollageParsing compares the query image to the database images using Gist [26] and HOG visual words [21], [37]. Specifically, the database images are sorted by similarity to the query image with respect to these two features, and the K best average ranks are selected as the retrieval set.

3.2 Computing content-adaptive windows

To implement content-adaptive windows, the current implementation of CollageParsing adopts the "objectness" algorithm of Alexe et al. [1].

Alexe et al. [1] defined the "objectness" of an image window as the likelihood that the window contains a foreground object of any kind instead of background texture such as grass, sky, or road. The authors observed that objects often have a closed boundary, a contrasting appearance from surroundings, and/or are unique in the image. Several cues are proposed to capture these generic properties: multiscale saliency, colour contrast, edge density, and superpixels straddling. The multiscale saliency cue is a multiscale adaptation of Hou and Zhang's visual saliency algorithm [17]. The color contrast cue measures the difference between the color histograms of the window and its surrounding rectangular ring. The edge density cue measures the proportion of pixels in the window's inner rectangular ring that are classified as edgels. The superpixels straddling cue measures the extent to which superpixels straddle the window (contain pixels both inside and outside the window). Windows that tightly bound an object are likely to have low straddling. Cues are combined in a Naive Bayes model. Our implementation of CollageParsing uses Alexe et al.'s publicly available implementation of objectness [1] with the default parameter values. Figure 2 shows a few examples of content-adaptive windows extracted using the objectness algorithm.

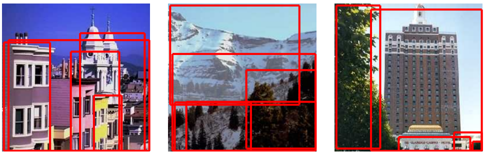

Fig. 2. Examples of image windows with high "objectness" [1]. In each image, windows with the top five objectness scores are shown. Images are from the SIFT Flow dataset [21].

Other algorithms for generating class-generic object window predictions, such as van de Sande's hierarchical segmentation based windows [30], can also be used in place of objectness at this stage in the pipeline.

Conceptually, CollageParsing performs nonparametric label transfer by matching content-adaptive windows in the query image with content-adaptive windows in the retrieval set. Each content-adaptive window is described using HOG features. The HOG features are dimension and scale adaptive: the algorithm sets

[1] v1.5, available at http://groups.inf.ed.ac.uk/calvin/objectness/

a target of six HOG cells along the longer dimension and allows the number of HOG cells along the shorter dimension to vary according to the window dimensions. When matching windows in the query image to windows in the retrieval set, only windows with the same HOG feature dimensions are compared. Each HOG feature vector is augmented with the scaled, normalized spatial coordinates of the window centroid, following the common practice of spatial coding in object recognition methods [3], [24], [35]. Spatial coding encourages matches to come from spatially similar regions in the respective images.

3.3 Computing unary potentials

Figure 3 shows a high-level visualization of the computation of the unary potential. Conceptually, the unary potential is computed by transferring the category labels from the most similar content-adaptive windows in the retrieval set, in a collage-like manner.

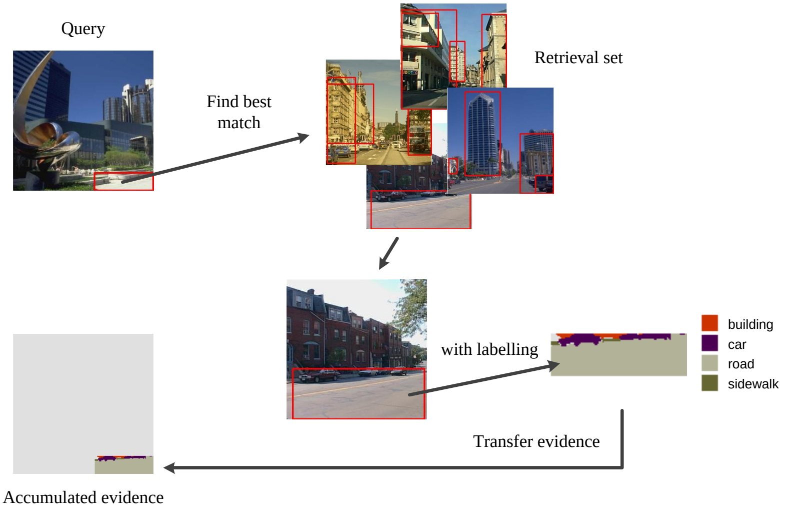

Fig. 3. Visualization of the computation of the unary potential ψ. The contribution of one window in the query image is depicted.

More formally, let $\psi(c,p)$ denote the unary potential or energy associated with assigning a semantic label of category c to pixel p:

$$\psi(c,p) = -\sum_{w \in W_q} \delta[L(\tilde{w}', p - \text{offset}(w)) = c]\phi_{\text{sim}}(w,w')\phi_{\text{idf}}(c) \tag{1}$$

$$w' = \operatorname*{arg\,min}_{u \in W_{\text{rs}}} ||f(w) - f(u)||_2$$

$$\tilde{w}' \text{ is } w' \text{ resized to the dimensions of } w$$

where w is a window in the set of content-adaptive windows in the query image, denoted W_q; $f(\cdot)$ is the HOG-based feature descriptor as described in Section 3.2; w' is the nearest neighbor window of w in the set of content-adaptive windows in the retrieval set, denoted W_{rs}; $\tilde{w}'$ is a resized version of w' such that it matches the dimensions of w; and $L(\cdot, \cdot)$ maps a window and an offset to a category label, or null if the offset is outside the window bounds. The term $p - \mathrm{offset}(w)$ gives the window-centric coordinates of the pixel p in window w. Therefore, $L(\tilde{w}', p - \mathrm{offset}(w))$ gives the category label of the projection or image of p in the matched window w'.

The term $\phi_{\mathrm{sim}}(w, w')$ is a weight that is proportional to the similarity between w and w'. Intuitively, higher quality matches should have greater influence in the labelling. We define

$$\phi_{\mathrm{sim}}(w, w') = s_f(w, w') s_l(w, w') \tag{2}$$

where $s_f(w, w')$ is the similarity between the two HOG-based feature descriptors, and to include color information $s_l(w, w')$ is the similarity between the windows' RGB color histograms. The feature descriptor distance is already computed in Eq. 1 and we convert it to a similarity score by

$$s_f(w, w') = \exp(-\alpha ||f(w) - f(u)||_2) \tag{3}$$

where α controls the exponential falloff. For the similarity between the windows' color histograms we take the histogram intersection.

The term $\phi_{\mathrm{idf}}(c)$ is a weight that is inversely proportional to the frequency of category c in the retrieval set:

$$\phi_{\mathrm{idf}}(c) = \frac{1}{N(c)^{\gamma}} \tag{4}$$

where $N(c)$ denotes the number of pixels of category c in the retrieval set. Eq. 4 performs a softened IDF-style weighting to account for differences in the frequency of categories. The constant γ controls the strength of the penalty given to high frequency categories, and as we show later in the experiments, influences the tradeoff between overall per-pixel and average per-class accuracy.

After computing $\psi(c, p)$ for all categories c and pixels p in the query image, all values are rescaled to be between -1 and 0.

Figure 4 visualizes the unary potential for an example query image from the SIFT Flow dataset [21].

3.4 Performing MRF inference

The unary potential ψ is combined with a pairwise potential θ defined over pairs of adjacent pixels. For θ we adopt the same pairwise potential term as in SuperParsing, which is based on the co-occurrences of category labels [34]:

$$\theta(c_p, c_q) = -\log[(P(c_p|c_q) + P(c_q|c_p))/2]\delta[c_p \neq c_q] \tag{5}$$

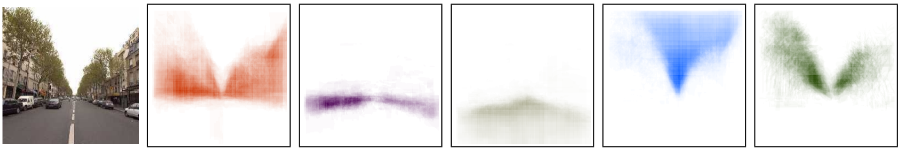

Fig. 4. Visualization of the unary potential for a sample query image (see also bottom row of Fig. 7). From left to right: building, car, road, sky, and tree categories.

where c_p and c_p are the category labels assigned to pixels p and q, which are adjacent. Intuitively, the pairwise term biases the labelling towards category transitions that are more frequently observed.

The global MRF energy function over the field of category labels $\mathbf{c} = \{c_p\}_{p \in I}$ is given by

$$E(\mathbf{c}) = \sum_{p \in I} \psi(c_p, p) + \lambda \sum_{(p,q) \in \varepsilon} \theta(c_p, c_q) \tag{6}$$

where ε is the set of pixel pairs (adjacent pixels) and λ is the MRF smoothing constant. The MRF energy is minimized using α/β-swap, a standard graph cuts technique [4], [5], [20].

To improve the alignment of the labelling with the query image structure, the labelling is then refined so that superpixels in the query image share the same label. All pixels within a superpixel are assigned the most common (mode) label in the superpixel. Superpixels are extracted using the graph-based method of Felzenszwalb and Huttenlocher [11], following Tighe and Lazebnik [34].

3.5 Retrieving semantic neighbors

Recall that the original retrieval set consisted of database images with similar Gist and HOG visual words to the query image (Section 3.1). Ideally, a retrieval set should consist of *semantically* similar database images. A retrieval set constructed from global image features provides a good first approximation. As a second approximation, the query image labelling is used to retrieve similarly labelled database images. Specifically, the database images are ranked by the pixel-by-pixel labelling correspondence with the query image labelling (in practice, the label fields may need to be resized). The K top ranked database images form the semantic retrieval set, and the CollageParsing pipeline is executed a second time with this retrieval set to obtain the final image labelling.

4 Experiments

We performed experiments on the SIFT Flow dataset [21], which consists of 200 query images and 2,488 database images from LabelMe. Images span a range of outdoor scene types, from natural to urban. Pixels are labelled with one of 33 semantic categories. The label frequencies are shown in Figure 5.

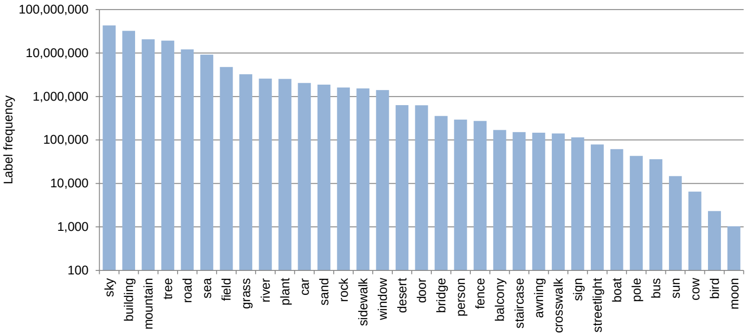

Fig. 5. Frequency counts of the semantic categories in the SIFT Flow dataset

Table 1 shows the experimental results and a comparison with the state-of-the-art nonparametric and parametric approaches on this dataset. We set the retrieval set size $K = 400$, $\alpha = 0.1$, $\gamma = 0.38$, and the MRF smoothing parameter λ to 0.01. We investigate the effect of these parameters on the algorithm performance later in this section. Both the overall per-pixel accuracy and the average per-class accuracy are reported. Average per-class accuracy is a more reliable measure of how well the algorithm performs across different categories, not just on the most commonly occurring ones.

Table 1. Per-pixel and per-class labelling accuracy on the SIFT Flow dataset [21]

	Per-pixel	Per-class
State-of-the-art nonparametric		
Liu et al. [21]	76.7	-
Gould and Zhang [14]	65.2	14.9
Tighe and Lazebnik [34]	77.0	30.1
Myeong et al. [25]	77.1	32.3
Eigen and Fergus [8]	77.1	32.5
Singh and Košecká [31]	79.2	33.8
CollageParsing	77.1	41.1
State-of-the-art parametric		
Farabet et al. [9], “natural”	78.5	29.6
Farabet et al. [9], “balanced”	74.2	46.0
Tighe and Lazebnik [33]	78.6	39.2

CollageParsing obtains higher per-class accuracy than all state-of-the-art nonparametric alternatives, by a wide margin, demonstrating its effectiveness across different categories and not only common “stuff” categories such as sky or grass. In particular, gains of 7 to 11% in per-class accuracy are obtained over state-of-the-art superpixel based approaches [8], [25], [31], [34], confirming our intuition described earlier that reasoning about mid-level, content-adaptive windows can be more productive than reasoning about low-level fragments.

CollageParsing's performance is also comparable to state-of-the-art parametric approaches. Compared with Tighe and Lazebnik's extended SuperParsing algorithm with per-exemplar detectors and a combination SVM [33], CollageParsing obtains 1.9% higher per-class accuracy and 1.5% lower per-pixel accuracy. Compared with Farabet et al.'s system [9] with "natural" training, at a tradeoff of 1.4% lower per-pixel accuracy an 11.5% per-class accuracy improvement is obtained. Farabet et al.'s system with "balanced" training achieves higher per-class accuracy, however at this setting the per-pixel accuracy falls below almost all nonparametric approaches in Table 1.

In contrast to state-of-the-art parametric approaches, as a nonparametric approach CollageParsing does not require expensive model training. As discussed in Section 1, this characteristic makes CollageParsing (and other nonparametric approaches) particularly well suited for open-universe datasets since no model retraining is required as the dataset expands. On a dataset of 45,000 images with 232 semantic labels, just the per-exemplar detector component of Tighe and Lazebnik [33] requires four days on a 512-node cluster to train. On a dataset of 715 images with eight semantic labels [13], Farabet et al. [9] requires "48h on a regular server" to train. At query time, our current implementation of CollageParsing requires approximately two minutes on a desktop to label an image through two passes of the pipeline (Figure 1). Our current implementation contains Matlab components that are not yet optimized for speed, and further improvements in the labelling time may be possible in future. Labelling time can also be reduced by stopping after a single pass of the pipeline, skipping the second pass with a semantic retrieval set. A modest cost in labelling accuracy is incurred. On the SIFT Flow dataset, the second pass with a semantic retrieval set improves per-pixel accuracy by 2.0% and per-class accuracy by 1.4%.

Figure 6 shows the effect of varying α, γ, the retrieval set size K, and the MRF smoothing parameter λ on the overall per-pixel and average per-class accuracy. Similar to Tighe and Lazebnik [34], we observed that the overall per-pixel accuracy drops when the retrieval set is too small for sufficient matches, but also when the retrieval set becomes large, confirming that the retrieval set performs a filtering role in matching query windows to semantically relevant database images. Interestingly, the average per-class accuracy drops off later than the overall per-pixel accuracy, suggesting a tradeoff between having a compact retrieval set and a retrieval set with enough representation to effectively match rarer categories. The constant γ controls the strength of the penalty given to high frequency categories. As γ increases, evidence for high frequency categories is discounted more heavily, and labelling is biased towards rarer categories. As reflected in the figure, this tends to increase the average per-class accuracy but at the expense of overall per-pixel accuracy.

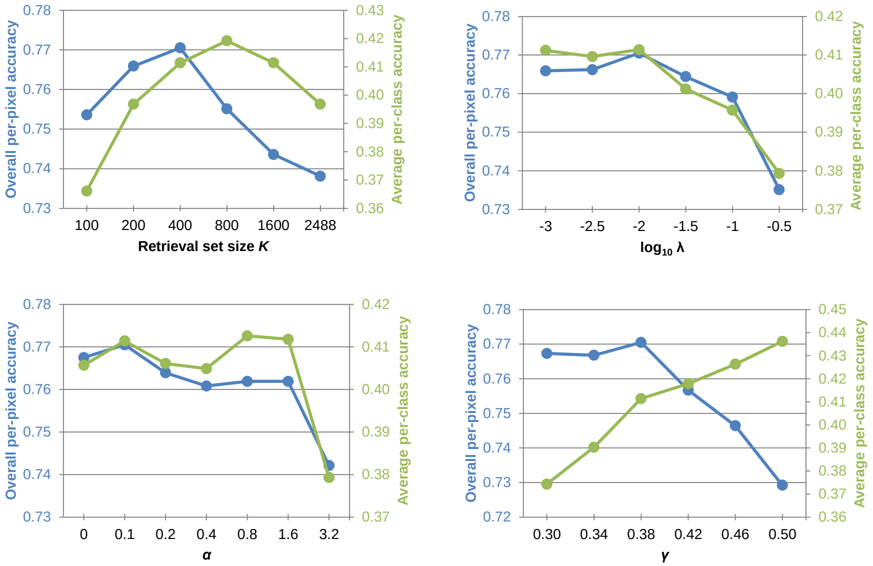

Fig. 6. Effects of varying the retrieval set size K, the MRF smoothing parameter λ, α, and γ on overall per-pixel accuracy and average per-class accuracy

Figure 7 shows some qualitative scene parsing results, including query images, system predicted labellings for both SuperParsing [34] and CollageParsing, and ground truth labellings. We found CollageParsing to perform robustly on a wide range of outdoor scenes, from natural (top) to urban (bottom) environments. Figure 8 shows two failure examples. The second example shows a case in which CollageParsing fails to predict the ground truth label for a large region but still provides a semantically reasonable alternative (*grass* instead of *field*).

5 Conclusion

In scene parsing we are interested in understanding an image at the pixel level of detail. This paper has described a novel algorithm for scalable nonparametric scene parsing that reasons about mid-level, content-adaptive windows, in contrast to recent state-of-the-art methods that focus on lower level superpixels. The CollageParsing pipeline consists of forming a retrieval set of similar database images, computing content-adaptive windows using the objectness technique [1], matching content-adaptive windows to accumulate a unary potential or energy for labelling each pixel with each semantic category label, and combining the unary potential with a co-occurrence based pairwise potential in a Markov random field framework. The initial labelling from Markov random field inference is refined by aligning the labelling with the query image superpixels. Finally, a second pass through the pipeline is taken with a semantic retrieval set of simi-

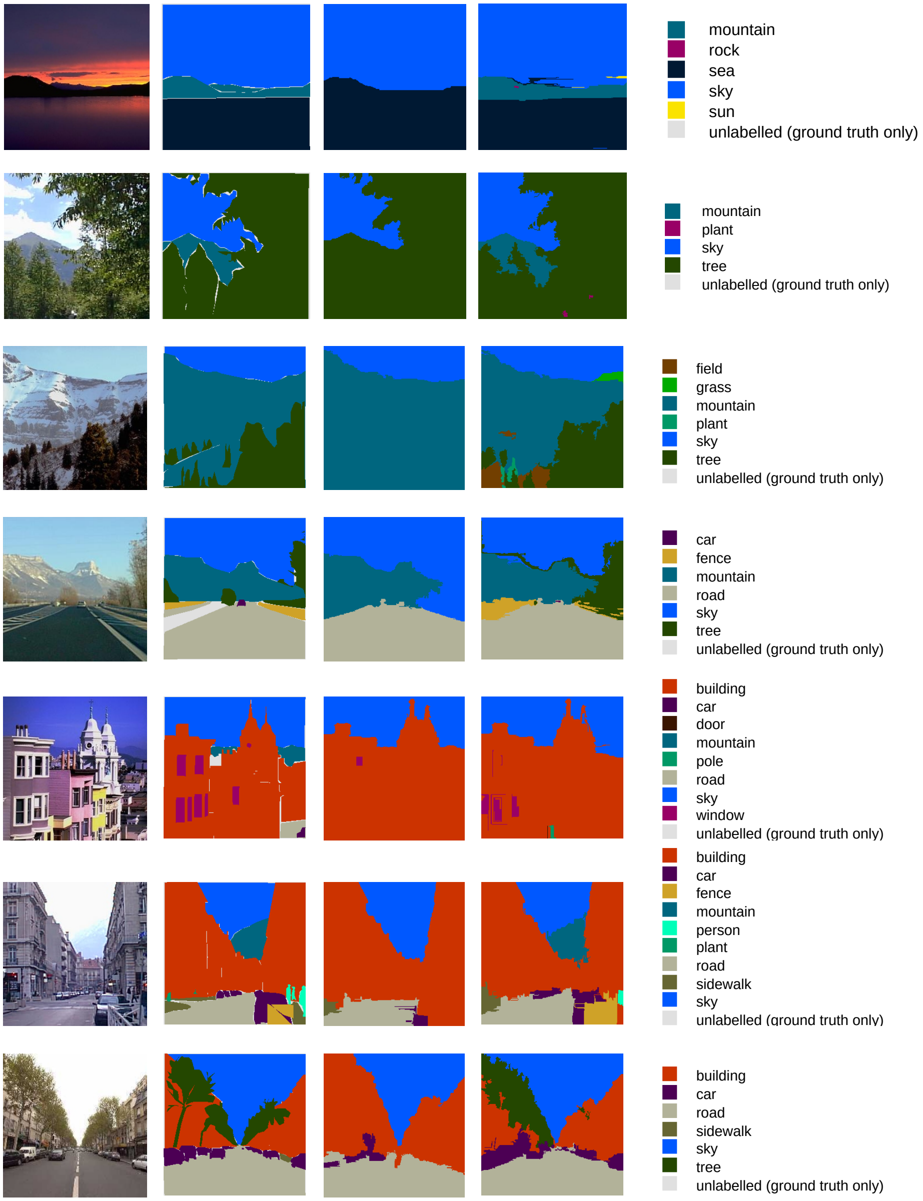

Fig. 7. Examples of scene parsing results on the SIFT Flow dataset. From left to right: query image, ground truth labelling, SuperParsing [34] predicted labelling, CollageParsing predicted labelling.

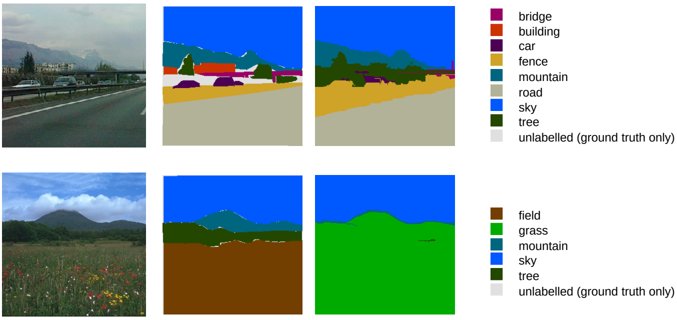

Fig. 8. Examples of failures on the SIFT Flow dataset. From left to right: query image, ground truth labelling, predicted labelling.

larly labelled database images. Experiments on the SIFT Flow benchmark [21] demonstrate the viability of CollageParsing, which obtains 7 to 11% higher average per-class accuracy than state-of-the-art nonparametric methods [8], [25], [31], [34], and comparable accuracy with state-of-the-art parametric methods [9], [33] while not requiring expensive model training.

As future work we plan to investigate whether other relevant image inference can be incorporated into the CollageParsing pipeline to further improve performance, such as the geometric inference that complements the semantic labelling in SuperParsing [34]. We would also like to assess the feasibility of using approximate nearest neighbor search methods to speed up window matching at query time.

Acknowledgements

This work was funded in part by the Natural Sciences and Engineering Research Council of Canada.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2189–2202 (2012)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. In: Proc. ACM SIGGRAPH (2009)
3. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2008)
4. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1124–1137 (2004)

5. Boykov, Y., Veksler, O., Zabih, R.: Efficient approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(12), 1222–1239 (2001)
6. Chen, X., Shrivastava, A., Gupta, A.: NEIL: extracting visual knowledge from web data. In: Proc. IEEE International Conference on Computer Vision (2013)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition. vol. 1, pp. 886–893 (2005)
8. Eigen, D., Fergus, R.: Nonparametric image parsing using adaptive neighbor sets. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2799–2806 (2012)
9. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Scene parsing with multiscale feature learning, purity trees, and optimal covers. In: Proc. International Conference on Machine Learning (2012)
10. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1785 (2009)
11. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
12. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
13. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: Proc. IEEE International Conference on Computer Vision (2009)
14. Gould, S., Zhang, Y.: PatchMatchGraph: Building a graph of dense patch correspondences for label transfer. In: Proc. European Conference on Computer Vision (2012)
15. Hays, J., Efros, A.A.: Scene completion using millions of photographs. In: Proc. ACM SIGGRAPH (2007)
16. Heitz, G., Koller, D.: Learning spatial context: using stuff to find things. In: Proc. European Conference on Computer Vision (2008)
17. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2007)
18. Isola, P., Liu, C.: Scene collaging: analysis and synthesis of natural images with semantic layers. In: Proc. IEEE International Conference on Computer Vision (2013)
19. Juneja, M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: Blocks that shout: distinctive parts for scene classification. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (2013)
20. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26(2), 147–159 (2004)
21. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing via label transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(12), 2368–2382 (2011)
22. Liu, C., Yuen, J., Torralba, A.: SIFT Flow: dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5), 978–994 (2011)
23. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of Exemplar-SVMs for object detection and beyond. In: Proc. IEEE International Conference on Computer Vision. pp. 89–96 (2011)

24. McCann, S., Lowe, D.G.: Spatially local coding for object recognition. In: Asian Conference on Computer Vision (2012)
25. Myeong, H., Chang, J.Y., Lee, K.M.: Learning object relationships via graph-based context model. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2727–2734 (2012)
26. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
27. Parikh, D., Grauman, K.: Relative attributes. In: Proc. IEEE International Conference on Computer Vision. pp. 503–510 (2011)
28. Patterson, G., Hays, J.: SUN Attribute database: discovering, annotating, and recognizing scene attributes. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 2751–2758 (2012)
29. Russell, B.C., Torralba, A., Murphy, K., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision 77(1-3), 157–173 (2008)
30. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In: Proc. IEEE International Conference on Computer Vision. pp. 1879 – 1886 (2011)
31. Singh, G., Košecká, J.: Nonparametric scene parsing with adaptive feature relevance and semantic context. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3151–3157 (2013)
32. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Proc. European Conference on Computer Vision (2012)
33. Tighe, J., Lazebnik, S.: Finding things: image parsing with regions and per-exemplar detectors. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3001–3008 (2013)
34. Tighe, J., Lazebnik, S.: Superparsing: scalable nonparametric image parsing with superpixels. International Journal of Computer Vision 101(2), 329–349 (2013)
35. Tuytelaars, T., Fritz, M., Saenko, K., Darrell, T.: The NBNN kernel. In: Proc. IEEE International Conference on Computer Vision. pp. 1824–1831 (2011)
36. Wu, J., Rehg, J.M.: CENTRIST: a visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1489–1501 (2011)
37. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (2010)