Scene parsing by nonparametric label transfer of content-adaptive windows

Frederick Tung, James J. Little

Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, Canada

Abstract

Scene parsing is the task of labelling every pixel in an image with its semantic category. We present CollageParsing, a nonparametric scene parsing algorithm that performs label transfer by matching content-adaptive windows. Content-adaptive windows provide a higher level of perceptual organization than superpixels, and unlike superpixels are designed to preserve entire objects instead of fragmenting them. Performing label transfer using content-adaptive windows enables the construction of a more effective Markov random field unary potential than previous approaches. On a standard benchmark consisting of outdoor scenes from the LabelMe database, CollageParsing obtains state-of-the-art performance with 15 to 19% higher average per-class accuracy than recent nonparametric scene parsing algorithms.

Keywords: image parsing, semantic segmentation, scene understanding, objectness

1. Introduction

In computer vision, we can analyze an image at many levels of abstraction. We might assume that the image contains one dominant object, and develop algorithms to assign a semantic category to that object, such as ‘airplane’ or ‘motorbike’. Alternatively, we may remove the assumption of a single dominant object in the image, and attempt to detect and localize all instances...
of multiple object categories. Or instead of objects, we may be interested in identifying the type of scene depicted in the image, such as ‘forest’ or ‘city’. Instead of categorizing objects and scenes, we may wish to describe objects and scenes by associating them with high-level attributes, such as ‘shiny’ or ‘rugged’.

In this work, we are interested in assigning a semantic label to every pixel in an image. We would like to explain the entire image, including both foreground objects of interest and background elements. Foreground objects are typically compact objects with well-defined boundaries, sometimes referred to as ‘things’ in the literature [1]; examples include cars and people. Background elements are often amorphous in shape or defined by texture, sometimes referred to as ‘stuff’; examples include grass and mountain. The task of labeling an entire image by its foreground and background semantic categories is referred to as scene parsing.

The diversity of foreground and background categories makes the task of scene parsing challenging. Traditionally, object categorization and detection tasks are approached by using machine learning techniques to learn compact, parametric models for the categories of interest. We can learn parametric classifiers to recognize airplanes or motorbikes, for example. However, it is challenging to scale these model-based approaches to detect the wide variety of categories that may be present in an image. A natural question to ask is whether we might adopt a nonparametric, database-driven approach to scene parsing instead. A nonparametric approach to scene parsing attempts to label the query image by matching parts of the image to similar parts in a large database of labelled images. The category classifier learning is replaced by a Markov random field in which unary potentials are computed by nearest-neighbor retrieval. Intuitively, labels are ‘transferred’ from exemplars in the database to the query image.

Nonparametric label transfer offers a powerful scene parsing tool especially in the context of open-universe image databases. Open-universe databases are collections that are continually changing as users add new images and annotations; the LabelMe database [2] is a well-known example. Nonparametric approaches to scene parsing are particularly suitable for databases that are continually changing because there is no need to re-train category models as new data are added. Moreover, no special accommodation is required when the vocabulary of semantic category labels is expanded.

State-of-the-art nonparametric scene parsing algorithms perform label transfer at the level of pixels or superpixels [3, 4, 5, 6, 7]. Superpixels offer
a higher level of perceptual abstraction than pixels, enabling more efficient inference in a random field model. Large, cohesive groups of pixels can be labelled at once. However, while superpixel-based methods tend to perform well on large regions of background ('stuff') categories, they tend to perform less effectively on foreground object ('thing') categories. We argue that, though better than pixels, superpixels are still a relatively low-level unit for nonparametric label transfer. Superpixels tend to fragment objects. Is it possible to perform label transfer at a higher level of organization?

In our preliminary work [8], we showed that it is indeed possible to perform label transfer at a higher level of perceptual organization. The CollageParsing algorithm performs label transfer using content-adaptive windows in a Markov random field framework. Windows are content-adaptive in the sense that they are designed to enclose entire objects instead of fragmenting them. There are many ways to generate content-adaptive windows, and we adopted Alexe et al.'s objectness algorithm [9] in our earlier implementation. We showed that performing label transfer using content-adaptive windows allows for the construction of a more effective unary potential than previous approaches, leading to state-of-the-art results on the standard SIFT Flow benchmark [3] among nonparametric scene parsing methods.

In this paper, we explore CollageParsing in more depth and make two algorithmic improvements to our original work. First, we compute content-adaptive windows using the state-of-the-art Edge Boxes algorithm [10] instead of the objectness algorithm [9] (Section 3.2). Second, we construct a more powerful unary potential by matching content-adaptive windows using deep learning features instead of HOG features (Section 3.3). To our knowledge, this improvement is the first use of deep learning features in a nearest neighbor search for scene parsing. These extensions significantly improve the scene parsing accuracy of the original CollageParsing algorithm, and to our knowledge enable the best per-pixel and per-class results to date on the SIFT Flow and LM+SUN benchmarks (Section 4).

2. Related work

2.1. Markov random fields in scene parsing

Markov random fields (MRFs) have been applied in a broad range of computer vision tasks, and have seen widespread adoption among methods for scene parsing. Scene parsing lends itself naturally to formulation in an MRF framework, with random variables corresponding to image pixels or
superpixels, and assignments to those random variables producing an image labelling. Shotton et al.’s TextonBoost framework [11] consists of a conditional random field (CRF) with shape-texture, colour, and location unary potentials and contrast-sensitive pairwise potentials. The shape-texture unary potential uses a boosted classifier of rectangular filter responses on texton maps. Kohli et al. [12] introduced a higher-order potential that encourages pixels within a segment to share the same label. The authors showed that inference in a CRF augmented with their higher-order potential can be performed efficiently using graph cuts. Ladický et al. [13] proposed a hierarchical CRF model in which potentials at multiple spatial scales are combined; scales in the hierarchy correspond to pixels and superpixels at varying granularities. Later, Ladický et al. [14] also introduced a higher-order potential over label co-occurrences that can be efficiently minimized using graph cuts. Krähenbühl and Koltun [15] performed scene parsing using a fully connected pixelwise CRF. The dense connectivity enables long-range relationships to be encoded. A thorough treatment of MRFs in computer vision can be found in the recent survey of Wang et al. [16].

2.2. Nonparametric methods for scene parsing

The term ‘label transfer’ for scene parsing was coined by Liu et al. [3]. Given a query image, Liu et al.’s nonparametric label transfer algorithm finds the image’s nearest neighbors in the database, warps the neighbors to the query image using SIFT Flow [17], and transfers the annotations from the neighbors to the query image using an MRF defined over pixels to integrate multiple cues.

Tighe and Lazebnik’s landmark SuperParsing algorithm [4] builds an MRF over superpixels instead of pixels. Label transfer is performed by matching superpixels in the query image with similar superpixels in the query’s nearest neighbor images. The nonparametric label transfer can be combined with additional parametric geometry classification (sky, vertical, horizontal) to improve labelling consistency.

Since the introduction of SuperParsing, several other effective superpixel-based nonparametric scene parsing algorithms have been developed. Eigen and Fergus [5] learned weights for each descriptor in the database in a supervised manner to reduce the influence of distractor superpixels, and augmented the retrieved set of neighbor superpixels with examples from rare classes with similar local context. Myeong et al. [6] applied link prediction techniques to superpixels extracted from the query image and its nearest
neighbors to learn more effective pairwise potentials. Singh and Košecská [7] applied a locally adaptive nearest neighbor technique to retrieve similar superpixels, and refined 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 [18] builds offline a graph of patch correspondences across all database images. Patch correspondences are found using an extended version of the PatchMatch algorithm [19] with additional “move” types for directing the local search for correspondences.

2.3. Recent parametric methods for scene parsing

Farabet et al. [20] 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 [21] extended the SuperParsing algorithm with per-exemplar detectors (Exemplar-SVMs [22]). A superpixel-based data term 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 final unary potential over pixels is determined by another SVM trained on the two data terms. Yang et al. [23] also proposed a hybrid parametric superpixel-based algorithm. The retrieved set of neighbor superpixels is augmented with a fixed set of examples of rare classes. Superpixel unary potentials are determined by a combination of non-parametric matching and a parametric SVM classifier trained on the rare class examples. Predicted labels are used to refine the retrieval sets of query image neighbors and superpixel neighbors in a second pass through the labelling pipeline.

2.4. Segmentation methods with overlapping regions

Pantofaru et al. [24] combined multiple image segmentations to create a map of intersections-of-regions. Each intersection-of-regions is classified by averaging over the category predictions of a learned region-based classifier on its constituent regions. Li et al. [25] also formed intersections-of-regions (referred to as superpixels in their formulation) using multiple image segmentations. intersections-of-regions are aggregated into objects by maximizing a
composite likelihood of predicted object overlap with segments. In contrast, CollageParsing is formulated in an MRF framework and operates on content-adaptive windows instead of intersections of regions generated by low-level image segmentation. CollageParsing is also a nonparametric method, while Pantofaru et al. [24] and Li et al. [25] are parametric model-based methods.

Kuettel and Ferrari [26] proposed a figure-ground (foreground-background) segmentation algorithm that transfers foreground masks from the training images by matching objectness windows. Matching is independent of the global similarity between the test and training images. The transferred foreground masks are used to derive unary potentials in a CRF. CollageParsing performs multi-class scene parsing instead of figure-ground separation, and specifically matches windows from globally similar images, which form the retrieval set. The retrieval set adds context that is important for multi-class labelling and enables scalable label transfer in large databases. The label transfer in CollageParsing is nonparametric, while Kuettel and Ferrari [26] iteratively learns a mixture model of appearance to refine the figure-ground segmentation.

2.5. Methods for computing content-adaptive windows

A key component of the CollageParsing algorithm is label transfer at the level of content-adaptive windows – image windows that are designed to preserve entire objects instead of fragmenting them. In our preliminary work [8], we used the objectness algorithm of Alexe et al. [9] to compute content-adaptive windows. Alexe et al. [9] 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 were 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 [27]. 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.
Since the introduction of objectness, several other effective methods for generating object proposals have been proposed. Some approaches employ hierarchical segmentation [28, 29]. For example, van de Sande et al.’s selective search [28] begins with an oversegmentation of the image and greedily merges similar segments until a single segment remains. All generated segments, or their bounding boxes, become candidate windows. To increase recall, multiple hierarchical segmentations are performed in color spaces with different sensitivities to shadows, shading, and highlights. Arbeláez et al.’s multiscale combinatorial grouping algorithm [29] produces a large initial set of object proposals by considering singletons, pairs, triplets, and 4-tuples of regions generated by the multiscale hierarchical segmentation algorithm at three individual scales and the combined multiscale hierarchy. Candidates are further refined using a random forest regressor trained on size, location, shape, and contour strength features.

Other methods generate object proposals by solving multiple figure-ground (foreground-background) segmentation problems from different initial seed superpixels [30, 31, 32]. For example, Carreira and Sminchisescu [30] performed multiple foreground-background segmentations from initial seeds on a regular grid using color features. A random forest regressor trained on mid-level appearance features is applied to rank the resulting segments, and maximal marginal relevance selection is used to diversify the final pool of proposals. Endres and Hoiem [31] selected initial seeds from a hierarchical segmentation and, instead of color features, applied multiple learned classifiers to compute the superpixel affinities for foreground-background segmentation. A structural learning method is used to rank the object proposals based on overlap and appearance features. Krähenbühl and Kolton [32] computed geodesic distances from foreground seed superpixels to the background superpixels. Geodesic distances refer to shortest paths over a weighted graph of superpixels, in which edges are weighted by the probability of boundary between superpixels. Critical level sets in the geodesic distance map form candidate object proposals.

Segmentation-based object proposal methods depend on the quality of the superpixels, which are often sensitive to small perturbations in the image [33]. Zitnick and Dollár [10] recently proposed the Edge Boxes algorithm, which generates object proposals using edge cues alone. The objectness score of a window is determined by the number and strength of edges wholly contained within the window. A fast structured forest method [34] is used for edge detection. The improved CollageParsing algorithm presented in this paper
Figure 1: High-level overview of the CollageParsing scene parsing pipeline. CollageParsing constructs an MRF over image pixels and performs label transfer at the level of content-adaptive windows. Content-adaptive windows are designed to preserve objects instead of fragmenting them, and are computed using Edge Boxes [10] (our preliminary work [8] used the objectness algorithm [9]).

uses the Edge Boxes method to generate content-adaptive windows for label transfer.

3. Algorithm description

CollageParsing follows a typical nonparametric scene parsing pipeline, in which an MRF is constructed and nonparametric label transfer is used to determine unary potentials. The CollageParsing pipeline is illustrated in Fig. 1. First, a short list of the query image’s nearest neighbors in the database is formed using global image features (Section 3.1). The short list is referred to as the retrieval set. Next, content-adaptive windows are extracted from the query image (Section 3.2). Unary potentials are then computed by matching these windows to similar content-adaptive windows in the retrieval set (Section 3.3). The unary potentials are combined with pairwise potentials, and the MRF is solved to obtain a pixelwise labelling of the image (Section 3.4).

3.1. Forming the retrieval set

The standard first step in the nonparametric scene parsing pipeline is the retrieval of database images that are contextually similar to the query image. 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. This coarse filtering step is typically performed using fast global image descriptors. CollageParsing
Figure 2: Examples of image windows with high objectness, as computed using Zitnick and Dollár’s Edge Boxes method [10]. In each image, windows with the top ten objectness scores are shown. Images are from the SIFT Flow dataset [3].

compares the query image to the database images using Gist [35] and HOG visual words [3, 36]. 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

CollageParsing performs label transfer at the level of content-adaptive windows. In our preliminary work [8], we generated content-adaptive windows using the objectness algorithm of Alexe et al. [9]. Since Alexe et al. [9], many other algorithms have been proposed for generating objectness windows or object proposals (Section 2.5). In this work, we adopt a state-of-the-art method called Edge Boxes.

Introduced by Zitnick and Dollár [10], the Edge Boxes algorithm generates object window proposals using edge cues alone. The objectness score of a window is determined by the number and strength of edges wholly contained within the window. A fast structured forest method [34] is used for edge detection. Candidate windows for score evaluation are generated by a sliding-window search over position, scale, and aspect ratio, with the step size parameterized by the target intersection over union (IoU). High-scoring windows are further refined by greedy iterative search. Finally, a non-maximal suppression step parameterized by the target IoU is performed.

We use the publicly available implementation of Edge Boxes with the default parameter values recommended in Zitnick and Dollár [10]. Fig. 2 shows examples of content-adaptive windows computed using the Edge Boxes algorithm.
3.3. Computing unary potentials

CollageParsing constructs an MRF over pixels and computes unary potentials by performing nearest-neighbor matching on content-adaptive windows and transferring the category labels. Compared with previous nonparametric scene parsing methods, label transfer is performed at a higher level of perceptual organization designed to preserve the integrity of objects.

To achieve this matching, we require a way to compare content-adaptive windows in the query image with content-adaptive windows in the retrieval set (Section 3.1). In our preliminary work [8], we compared content-adaptive windows using HOG features. The HOG features were augmented with the scaled, normalized spatial coordinates of the window centroid, following the common practice of spatial coding in object recognition [37, 38, 39], which encourages matches to come from spatially similar regions in the respective images.

This work improves the window matching step by replacing the HOG features with deep features. We use the CNN-M-1024 convolutional neural network from the recent experimental evaluation of Chatfield et al. [40], which was demonstrated to provide high-quality features for object classification. To obtain the feature descriptor for a content-adaptive window, we resize the cropped image window to the canonical size required by the convolutional neural network, run the network, and use the 1024-dimensional response as the feature descriptor. We use Chatfield et al.’s CNN-M-1024 network pre-trained on ImageNet as an off-the-shelf feature extractor, without any fine-tuning on our scene parsing images. We keep the spatial coding component of our preliminary work [8] as described above.

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. Since the content-adaptive windows overlap and vary in size, labels are transferred in a collage-like manner.

More formally, let $\psi(c, p)$ denote the unary potential or energy associated with assigning a semantic label of category $c$ to pixel $p$:
Figure 3: Visualization of the computation of the unary potential \( \psi \) by label transfer. The contribution of one content-adaptive window in the query image is depicted.

\[
\psi(c, p) = - \sum_{w \in W_q} \delta[L(\tilde{w}', p - \text{offset}(w)) = c] \phi(c) \\
\]

\[
w' = \arg \min_{u \in W_{rs}} ||f(w) - f(u)||_2 \\
\]

\( \tilde{w}' \) is \( w' \) 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 window feature descriptor as described above; \( w' \) 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 \( w \); \( \delta[\cdot] \) is the indicator function; 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 \( p - \text{offset}(w) \) gives the window-centric coordinates of the pixel \( p \) in window \( w \). Therefore, \( L(\tilde{w}', p - \text{offset}(w)) \) gives the category label of the projection or image of \( p \) in the matched window \( w' \).

The term \( \phi(c) \) is a weight that is inversely proportional to the frequency of category \( c \) in the retrieval set:
Figure 4: Visualization of the unary potential for three sample query images (see also three bottom rows of Fig. 7). Top row, from left to right: building, car, plant, sidewalk, and sky categories. Middle row, from left to right: building, car, fence, person, and road categories. Bottom row, from left to right: building, car, road, sky, and tree categories.

\[
\phi(c) = \frac{1}{N(c)^\gamma}
\]  

where \(N(c)\) denotes the number of pixels of category \(c\) in the retrieval set. Eq. 2 performs a softened IDF-style weighting to account for differences in the frequency of categories. The constant \(\gamma\) 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 three example query images from the SIFT Flow dataset [3].

3.4. Performing MRF inference

The unary potential \(\psi\) is combined with a pairwise potential \(\theta\) defined over pairs of adjacent pixels. For \(\theta\) we adopt the same pairwise potential term as in SuperParsing, which is based on the co-occurrences of category labels [4]:
where $c_p$ and $c_p$ are the category labels assigned to pixels $p$ and $q$. 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 $c = \{c_p\}_{p \in I}$ is given by

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

where $\varepsilon$ is the set of pixel pairs (adjacent pixels) and $\lambda$ is the MRF smoothing constant. We minimize the MRF energy using $\alpha/\beta$-swap, a standard graph cuts technique [41, 42, 43].

The labelling is spatially refined in a post-processing step by aligning the labels to the query image’s superpixels. All pixels within a superpixel are assigned the most common (mode) label in the superpixel. Following SuperParsing [4], superpixels are extracted using the graph-based method of Felzenszwalb and Huttenlocher [44].

4. Experiments

We performed experiments on the SIFT Flow dataset [3], a scene parsing benchmark consisting of 200 query images and 2,488 database images from
Table 1: Per-pixel and per-class labelling accuracy on the SIFT Flow dataset [3]

<table>
<thead>
<tr>
<th></th>
<th>Per-pixel</th>
<th>Per-class</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State-of-the-art nonparametric methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu et al. [3]</td>
<td>76.7</td>
<td>-</td>
</tr>
<tr>
<td>Gould and Zhang [18]</td>
<td>65.2</td>
<td>14.9</td>
</tr>
<tr>
<td>Tighe and Lazebnik [4]</td>
<td>77.0</td>
<td>30.1</td>
</tr>
<tr>
<td>Myeong et al. [6]</td>
<td>77.1</td>
<td>32.3</td>
</tr>
<tr>
<td>Eigen and Fergus [5]</td>
<td>77.1</td>
<td>32.5</td>
</tr>
<tr>
<td>Singh and Košcká [7]</td>
<td>79.2</td>
<td>33.8</td>
</tr>
<tr>
<td>CollageParsing (our preliminary work [8])</td>
<td>77.1</td>
<td>41.1</td>
</tr>
<tr>
<td><strong>CollageParsing (this paper)</strong></td>
<td><strong>79.9</strong></td>
<td><strong>49.3</strong></td>
</tr>
<tr>
<td><strong>State-of-the-art parametric methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe and Lazebnik [21]</td>
<td>78.6</td>
<td>39.2</td>
</tr>
<tr>
<td>Farabet et al. [20], “natural”</td>
<td>78.5</td>
<td>29.6</td>
</tr>
<tr>
<td>Farabet et al. [20], “balanced”</td>
<td>74.2</td>
<td>46.0</td>
</tr>
<tr>
<td>Yang et al. [23]</td>
<td>79.8</td>
<td>48.7</td>
</tr>
</tbody>
</table>

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.

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 $K = 400$, $\gamma = 0.4$, and the MRF smoothing parameter $\lambda$ 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 and not just on the most commonly occurring ones.

By performing label transfer at the level of content-adaptive windows, CollageParsing obtains state-of-the-art labelling results on the benchmark. Our preliminary work [8] using traditional HOG features and objectness already obtained a 7 to 11% improvement in average per-class accuracy over state-of-the-art superpixel based nonparametric approaches [4, 5, 6, 7]. By using off-the-shelf deep features and the recently proposed Edge Boxes
method, we further extend CollageParsing in this paper to realize gains of 15 to 19% average per-class accuracy in comparison to state-of-the-art non-parametric approaches.

Labelling performance is also comparable to or better than state-of-the-art parametric methods, while not requiring expensive model training. For example, on the LM+SUN dataset (45,000 images with 232 semantic labels), just the per-exemplar detector component of Tighe and Lazebnik [21] requires four days on a 512-node cluster to train. On a dataset of 715 images with eight semantic labels [45], Farabet et al. [20] requires “48h on a regular server” to train. The bulk of our computation time comes from feature extraction. Once features are computed, matching content-adaptive windows and solving the Markov random field by graph cuts takes approximately 10 seconds in Matlab on a desktop. On average, 855 content-adaptive windows are generated per database image using Edge Boxes, and the nearest-neighbor window matching and label transfer account for 8 of those 10 seconds.

Results are comparable to the hybrid parametric approach recently proposed by Yang et al. [23], which achieves a large increase in average per-class accuracy (nearly 18% compared to the authors’ baseline system) by augmenting the retrieval set with examples from rare classes to improve rare class labelling. Interestingly, this idea of ‘rare class expansion’ is complementary to our work and could be an interesting direction for future investigation. However, from a nonparametric design perspective, ‘rare class expansion’ would make CollageParsing more difficult to scale in open-universe environments because as new images or categories are added to the database, the auxiliary set needs to be re-composed – both the selection of ‘rare’ categories and the examples themselves.

Figure 6 shows the effect of varying $\gamma$, the retrieval set size $K$, and the MRF smoothing parameter $\lambda$ on the overall per-pixel and average per-class accuracy. Similar to Tighe and Lazebnik [4], 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. The constant $\gamma$ controls the strength of the penalty given to high frequency categories. As $\gamma$ 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.

Figure 7 shows some qualitative scene parsing results, including query im-
Figure 6: Effects of varying the retrieval set size $K$, the MRF smoothing parameter $\lambda$, and $\gamma$ on overall per-pixel accuracy and average per-class accuracy on the SIFT Flow dataset.

ages, system predicted labellings for both SuperParsing [4] and Collage Parsing, and ground truth labellings. We found CollageParsing to perform well in a wide range of outdoor scenes, from natural (top) to urban (bottom) environments. The post-processing step of spatial refinement using superpixels produces more visually pleasing results and has only a small effect on accuracy: about 1% overall per-pixel accuracy and 0.1% average per-class accuracy. Figure 8 shows some examples of scene parsing results with and without superpixel post-processing.

We performed further experiments on a larger dataset containing both indoor and outdoor scenes. The LM+SUN dataset [4] consists of 500 query images and 45,176 database images from the LabelMe and SUN datasets [36]. Indoor and outdoor scenes are represented approximately evenly. LM+SUN is annotated with a broad vocabulary of 232 semantic categories.

Table 2 shows evaluation results and a comparison with the state-of-the-art nonparametric and parametric approaches on LM+SUN. Since LM+SUN contains a larger number of images and categories, we increased the retrieval set size $K$ to 1600. The trade-off of $\gamma$ is shown in Figure 9. LM+SUN is a very challenging benchmark but CollageParsing obtains encouraging results.
Figure 7: Examples of scene parsing results on the SIFT Flow dataset. From left to right: query image, ground truth labelling, SuperParsing [4] predicted labelling, CollageParsing predicted labelling.
Figure 8: Examples of scene parsing results with and without the spatial refinement post-processing step. From left to right: query image, labelling without post-processing, labelling with post-processing.

Table 2: Per-pixel and per-class labelling accuracy on the LM+SUN dataset [4]

<table>
<thead>
<tr>
<th></th>
<th>Per-pixel</th>
<th>Per-class</th>
</tr>
</thead>
<tbody>
<tr>
<td>State-of-the-art nonparametric methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe and Lazebnik [4]</td>
<td>54.9</td>
<td>7.1</td>
</tr>
<tr>
<td><strong>CollageParsing</strong></td>
<td><strong>60.8</strong></td>
<td><strong>19.3</strong></td>
</tr>
<tr>
<td>State-of-the-art parametric methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tighe and Lazebnik [21]</td>
<td><strong>61.4</strong></td>
<td>15.2</td>
</tr>
<tr>
<td>Yang et al. [23]</td>
<td>60.6</td>
<td>18.0</td>
</tr>
</tbody>
</table>
Figure 9: Varying $\gamma$ to trade off overall per-pixel accuracy and average per-class accuracy on the LM+SUN dataset

An improvement of 6% per-pixel and 12% average per-class accuracy is obtained relative to SuperParsing. Compared with state-of-the-art parametric methods, a modest gain of 1% average per-class accuracy is achieved (or 7% relative). An interesting direction for future investigation to further improve accuracy may be the integration of depth information from RGB-D cameras in indoor scenes or stereo pairs in outdoor scenes [46].

Inference on LM+SUN is more time consuming than on SIFT Flow because of the increased retrieval set size and number of categories. Once features are computed, matching and inference take approximately 3 minutes in Matlab on a desktop, which is about half the runtime of Yang et al. [23] on this dataset. No model training is required.

Figure 10 shows additional qualitative results on indoor scenes, which are not represented in the SIFT Flow dataset.

5. Conclusion

We have presented a nonparametric scene parsing algorithm that performs label transfer by matching content-adaptive windows instead of low-level superpixels. On the SIFT Flow scene parsing benchmark, CollageParsing obtains improvements of 15 to 19% in average per-class accuracy compared to state-of-the-art superpixel-based nonparametric approaches. In addition, since it is database-driven and requires no training of category models, CollageParsing also scales naturally if new images or categories are added to the database.
Figure 10: Examples of indoor scene parsing results on the LM+SUN dataset. Top left: query image; top right: ground truth labelling; bottom left: SuperParsing [4] predicted labelling; bottom right: CollageParsing predicted labelling.
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 [4]. We would also like to assess the feasibility of using hashing techniques to speed up window matching and reduce database storage requirements.

Acknowledgements

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


[18] S. Gould, Y. Zhang, PatchMatchGraph: Building a graph of dense patch
correspondences for label transfer, in: Proc. European Conference on

a randomized correspondence algorithm for structural image editing, in:

[20] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Scene parsing with mul-
tiscale feature learning, purity trees, and optimal covers, in: Proc. In-

[21] J. Tighe, S. Lazebnik, Finding things: image parsing with regions and
per-exemplar detectors, in: Proc. IEEE Conference on Computer Vision
and Pattern Recognition, 2013, pp. 3001–3008.

[22] T. Malisiewicz, A. Gupta, A. A. Efros, Ensemble of Exemplar-SVMs for
object detection and beyond, in: Proc. IEEE International Conference

with attention to rare classes, in: Proc. IEEE Conference on Computer
Vision and Pattern Recognition, 2014.

[24] C. Pantofaru, C. Schmid, M. Hebert, Object recognition by integrat-
ing multiple image segmentations, in: Proc. European Conference on

[25] F. Li, J. Carreira, G. Lebanon, C. Sminchisescu, Composite statisti-
cal inference for semantic segmentation, in: Proc. IEEE Conference on
Computer Vision and Pattern Recognition, 2013.

[26] D. Kuettel, V. Ferrari, Figure-ground segmentation by transferring win-
Recognition, 2012.

[27] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in:
Proc. IEEE Conference on Computer Vision and Pattern Recognition,
2007.


