Nonparametric scene parsing with adaptive feature relevance and semantic context

Gautam Singh Jana Košecká
George Mason University
Fairfax, VA
{gsinghc,kosecka}@cs.gmu.edu

Abstract

This paper presents a nonparametric approach to semantic parsing using small patches and simple gradient, color and location features. We learn the relevance of individual feature channels at test time using a locally adaptive distance metric. To further improve the accuracy of the nonparametric approach, we examine the importance of the retrieval set used to compute the nearest neighbours using a novel semantic descriptor to retrieve better candidates. The approach is validated by experiments on several datasets used for semantic parsing demonstrating the superiority of the method compared to the state of art approaches.

1. Introduction

The problem of semantic labelling, requires simultaneous segmentation of an image into regions and categorization of all the image pixels. The main ingredients of the problem are the choice of elementary regions (pixels, superpixels), types of features used to characterize them, methods for computing local label evidence and means of integrating the spatial information. Semantic segmentation has been particularly active in recent years, due to the development of methods for integration of object detection techniques, with various contextual cues and top down information as well as advancements in inference algorithms used to compute the optimal labelling.

With the increasing complexity and size of the datasets used for evaluation of semantic segmentation, nonparametric techniques [15, 26] combined with various context driven retrieval strategies have demonstrated notable improvement in the performance. These methods typically start with an oversegmentation of an image into superpixels followed by the computation of a rich set of features characterizing both appearance and local geometry at the superpixel level. Due to a large number of diverse features, distance learning techniques have been shown to be effective for retrieval of the closest neighbours.

In the proposed work, we follow a nonparametric approach and make the following contributions: (i) We forgo the use of large superpixels and complex features and tackle the problem of semantic segmentation using local patches characterized by gradient orientation, color and location features. The appeal of this representation is its simplicity and resemblance to local patch based methods used in the context of biologically inspired methods; (ii) We adopt an approach for learning the *relevance* of individual feature channels (gradient orientation, color and location) used in k-nearest neighbour (k-NN) retrieval and (iii) We demonstrate a novel approach for obtaining a retrieval set where the coarse semantic labelling is used to retrieve similar views and refine the likelihood estimates. The proposed approach is validated extensively on several semantic segmentation datasets consistently showing improved performance over the state of the art methods.

2. Related Work

In recent years, a large number of approaches for semantic segmentation have been proposed. Due to the complex nature of the problem, the existing approaches differ in the choice of elementary regions, choice of features to describe them, methods for modeling spatial relationships, means of incorporating of context and choice of optimization techniques for solving the optimal labelling problem. The most successful approaches typically use Conditional Random Field (CRF) models [7, 6, 11, 23, 13, 12]. Traditional CRF models [23] combine local appearance information with a smoothness prior that favours same labellings for neighbouring regions. Researchers in [11] proposed the use of higher order potentials in a hierarchical framework which allowed the integration of features at different levels (pixels and superpixels). Other works have looked at exploring object co-occurrence statistics [7, 12] and combining results from object detectors [13] .

With the increasing sizes of datasets and an increasing

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.405

number of labels, the use of nonparametric approaches have shown notable progress [15, 26, 4, 31]. They are appealing as they can utilize efficient approximate nearest neighbour search techniques e.g. k-d trees [19] and contextual cues. Context is often captured by a retrieval set of images similar to the query and methods developed for establishing matches between image regions (at pixel or superpixel level) for labelling the image. Using the method of SIFT Flow, pixel-wise correspondences are established between images for label transfer in [15]. Authors in [26] work at the superpixel-level and retrieve similar images using global image features which is followed by superpixel-level matching using local features and a Markov random field (MRF) to incorporate neighbourhood context. The work of [26] was extended by [4] by training per superpixel per feature weights and also by incorporating superpixel-level semantic context. A set of partially similar images is used in [31] by searching for matches for each region of the query image and then using the retrieval set for label transfer. A nonparametric method which avoids the construction of a retrieval set is [8] which instead addresses the problem of semantic labelling by building a graph of patch correspondences across image sets and transfers annotations to unlabeled images using the established correspondences. However the degree of the graph vertices is limited due to memory requirements for large datasets like SiftFlow [15].

Our work is closely related to the work of [26, 4] in that we also pursue nonparametric approach, but differ in the choice of elementary regions, features, feature relevance learning and the method for computing the retrieval set for k-NN classification. In our case, the retrieval set is obtained in a feedback manner using a novel semantic label descriptor computed from the initial semantic segmentation. Similarly to [4], we follow the observation that a single global distance metric is often not sufficient for handling the large variations within a class and propose to compute weights for individual features channels. The weights in our case are computed at the test time to indicate the importance of color, gradient orientation vs location for individual regions. The computation of the feature relevance we adopt falls into a broad class of distance metric learning techniques which have been shown to be beneficial for many problems like image classification [5], object segmentation [17] and image annotation [9]. For a comprehensive survey on distance functions, we refer the reader to [22].

3. Approach

In this section, we will describe our baseline approach, followed by the method of weight computation in Section 4 and semantic contextual retrieval in Section 5.

3.1. Problem Formulation

We formulate the semantic segmentation of an image segmented into small superpixels. The output of the semantic segmentation is a labelling $\mathbf{L} = (l_1, l_2, \dots l_S)^\top$ with hidden variables assigning each superpixel s_i a unique label, $l_i \in \{1, 2, \dots, nL\}$, where nL and S is the total number of the semantic categories and superpixels respectively. The posterior probability of a labelling $\mathbf{L}$ given the observed appearance feature vectors $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_S]$ computed for each superpixel can be expressed as:

$$P(\mathbf{L}|\mathtt{A}) = \frac{P(\mathtt{A}|\mathbf{L})\, P(\mathbf{L})}{P(\mathtt{A})}. \tag{1}$$

We estimate the labelling $\mathbf{L}$ as a Maximum A Posteriori Probability (MAP),

$$\underset{\mathbf{L}}{\operatorname{argmax}}\, P(\mathbf{L}|\mathbf{A}) = \underset{\mathbf{L}}{\operatorname{argmax}}\, P(\mathbf{A}|\mathbf{L})\, P(\mathbf{L}). \tag{2}$$

The observation likelihood $P(\mathbf{A}|\mathbf{L})$ and the joint prior $P(\mathbf{L})$ are described in later subsections.

3.2. Superpixels and features

For an image, we extract superpixels utilizing a segmentation method [29] where superpixel boundaries are obtained as watersheds on a negative absolute Laplacian image with LoG extremas as seeds. These blob-based superpixels are efficient to compute and naturally consistent with the boundaries. Similarly to [18], for each superpixel, we compute a 133-dimensional feature vector $\mathbf{a}_i$ comprised of SIFT descriptor (128 dimensions), color mean over the pixels of an individual superpixel in Lab color space (3 dimensions) and the location of the superpixel centroid (2 dimensions). The SIFT descriptor for a superpixel is computed at a fixed scale and orientation using publicly available code [27].

3.3. Appearance Likelihood

In order to compute the appearance likelihood for the entire image, we approximate the Naive Bayes assumption yielding

$$P(\mathbf{A}|\mathbf{L}) \approx \prod_{i=1}^{S} P(\mathbf{a}_i|l_i). \tag{3}$$

Such an approximation assumes independence between appearance features of the superpixels given their labels.

The individual label likelihood $P(\mathbf{a}_i|l_j)$ for a superpixel s_i is obtained using a k-NN method. Since a superpixel is uniquely represented by its feature vector, we use the symbols s_i and $\mathbf{a_i}$ interchangeably. For each class l_j and every superpixel s_i of the query image, we compute a label likelihood score:

$$L(\mathbf{a}_i, l_j) = \frac{n(l_j, N_{ik})/n(l_j, G)}{n(\bar{l_j}, N_{ik})/n(\bar{l_j}, G)} \tag{4}$$

where

- $\bar{l_j} = L \setminus l_j$ is the set of all labels excluding l_j;
- N_{ik} is a neighbourhood around $\mathbf{a}_i$ with exactly k points in it;
- $n(l_j, N_{ik})$ is the number of superpixels of class l_j inside N_{ik};
- $n(l_j, G)$ is the number of superpixels of class l_j in the set G (described later in Section 3.5).

We compute the normalized label likelihood score using the individual label likelihood:

$$P(\mathbf{a}_i|l_j) = \frac{L(\mathbf{a}_i, l_j)}{\sum_{l_k=1}^{nL} L(\mathbf{a}_i, l_k)} \tag{5}$$

A straightforward way to compute the neighbourhood N_{ik} is to use the concatenated feature $\mathbf{a}_i$ (Section 3.2) and retrieve the k nearest points by computing distance to superpixels in G. Such a retrieval can be efficiently performed by the use of approximate nearest neighbour methods like k-d trees [19].

3.4. Inference

For the joint prior $P(\mathbf{L})$, we adapt the approach of [18] which used as its smoothness term E_{smooth}, a combination of the Potts model (using constant penalty δ) and a color difference based term. The maximization in Eq. (2) can be rewritten in log-space and the optimal labelling $\mathbf{L}^*$ achieved as

$$\underset{\mathbf{L}}{\operatorname{argmin}} \Big(\sum_{i=1}^{S} E_{app} + \lambda \sum_{(i,j)\in\mathcal{E}} E_{smooth} \Big), \tag{6}$$

where $E_{app} = -\log P(\mathbf{a}_i|l_j)$ from Eq. (5) and the set $\mathcal{E}$ contains all neighbouring superpixel pairs. The scalar λ is the weight for the smoothness term. We perform the inference in the MRF, i.e. a search for a MAP assignment, using an efficient and fast publicly available MAX-SUM solver [28].

3.5. Retrieval Set

The computation of the appearance likelihood in Section3.3 uses images from the training set. Instead of using the entire training set in the k-NN method, it is more useful to utilize a subset of images which are similar to the query image. For example, when trying to label a seaside image, it is more helpful if we search for the nearest neighbours in images of beaches and discard views from street scenes. We use overall scene appearance to find a relatively smaller set of training images instead of using the entire training set. It helps discard images which are dissimilar to the query image and provides a scene-level context which can help improve the labelling performance. The retrieval subset will serve as the source of image annotations which will be used to label the query image. We compute three global image features for the dataset, namely: (i) GIST [21], (ii) spatial pyramid [14] of quantized SIFT [16] and (iii) rgb-color histograms with 8 bins per color channel. All the images in the training set T are ranked for each individual global image feature in ascending order of the Euclidean distance from the query image. We then add the individual feature ranks and re-rank the images of the training set based on the aggregate rank. Finally, we select a subset of images T_g from the training set T as the retrieval set. The superpixels of the images in set T_g compose the set of training instances G in Eq. (5).

This constitutes our baseline approach and is denoted UKNN-MRF in the experiments for the uniformly weighted k-NN. Its distinguishing characteristics are the use of small patch-like superpixels, simple features and approximate nearest neighbour methods in the context of k-NN classification. In the next two sections, we describe in detail the two contributions of this work: a method for weighting different feature channels and the strategy for improving the retrieval set.

4. Weighted k-NN

The baseline k-NN approach uses Euclidean distance to compute the neighbourhood around the point. We propose to use a weighted k-NN method to compute the neighbourhood of a query point. To compute a weighted distance between two superpixels $\mathbf{a}_i$ and $\mathbf{a}_j$, we split the feature vector into three feature channels of gradient orientation, color and location and first compute distances in individual feature spaces:

$$d_f^{ij} = [d_c^{ij}, d_s^{ij}, d_l^{ij}]^\top \tag{7}$$

where $d_c^{ij}, d_s^{ij}, d_l^{ij}$ are the Euclidean distances between the color, SIFT and location channels of the feature vectors $\mathbf{a}_i$ and $\mathbf{a}_j$ of the two superpixels respectively. We now define a weighted distance between the two superpixels as

$$d_w^{ij} = w^\top d_f^{ij} \tag{8}$$

where $w = [w_1, w_2, w_3] \in \Re^3$ defines the weights for the individual feature distances. Using the weighted distance from Eq. (8), we can now obtain the neighbourhood N_{ik} around a superpixel by applying it to the feature distance vector d_f^{ij} between $\mathbf{a}_i$ and $\mathbf{a}_j \in G$ to compute the label likelihood scores in Eq. (4). We now describe an approach to compute these weights.

Weight computation With the varying nature of the retrieval set for individual query images, we use the locally adaptive metric approach of [3] for the weight computation. It is a query-based technique which uses a global metric to select neighbours for a test point which are then used to

refine the feature weights. In our setting, the test points are the individual superpixels of the query image.

The goal is to estimate the relevance of a feature channel i by evaluating its ability to predict class posterior probabilites locally at a query point. This is done by computing the expectation of the posterior $P(l_j|\mathbf{x})$ conditioned at a test point $\mathbf{x_0}$ along feature channel i. The ability of feature channel i to predict $P(l_j|\mathbf{z})$ at $x_i = z_i$ is defined as

$$r_i(\mathbf{z}) = \sum_{l_j=1}^{nL} \frac{(P(l_j|\mathbf{z}) - \bar{P}(l_j|x_i = z_i))^2}{\bar{P}(l_j|x_i = z_i)} \tag{9}$$

Intuitively, the smaller the difference between $P(l_j|\mathbf{z})$ and $\bar{P}(l_j|x_i = z_i)$, the more information feature channel i provides for predicting the class posterior probabilities locally at $\mathbf{z}$. For the query point $\mathbf{x_0}$, the relevance for feature i can be computed by averaging the $r_i(\mathbf{z})$'s in its neighbourhood

$$\bar{r}_i(\mathbf{x_0}) = \frac{1}{|N(\mathbf{x_0})|} \sum_{\mathbf{z} \in N(\mathbf{x_0})} r_i(\mathbf{z}) \tag{10}$$

where $N(\mathbf{x_0})$ denotes a neighbourhood centered at $\mathbf{x_0}$ (using the current feature weights) with K_0 points in it. The relative relevance can then be computed as

$$w_i(\mathbf{x_0}) = \frac{exp\left(cR_i(\mathbf{x_0})\right)}{\sum_{p=1}^{m} exp\left(cR_p(\mathbf{x_0})\right.} \tag{11}$$

where m is the number of individual feature channels (three in our case), c is a parameter which determines the influence of $\bar{r}_i$ (at $c = 0$, all three feature channels have equal weights) and $R_i(\mathbf{x_0}) = \max_{p=1}^{m}\{\bar{r}_p(\mathbf{x_0})\} - \bar{r}_i(\mathbf{x_0})$. The quantities $P(l_j|\mathbf{z})$ and $\bar{P}(l_j|x_i = z_i)$ in Eq. (9) are estimated by considering neighbourhoods centered at $\mathbf{z}$ described in detail by [3]. In the experiments section, this method evaluates the effect of the weight learning on the final classification and is denoted WKNN-MRF for the weighted k-NN.

5. Semantic Contextual Retrieval

The semantic labelling of an image, even if inaccurate provides a strong cue about the presence and absence of different categories in the image. While the idea of using context to improve the labelling has been explored in the past for image superpixels [20, 4], here we examine the effectiveness of this idea in the stage of improving the entire retrieval set. In order to do so, we propose a global descriptor derived from the intial labelling of the image which will be used to improve the retrieval set.

To summarize the semantic label information of a labeled image, we introduce the *semantic label descriptor* for a labelled image. This descriptor captures the basic underlying structure of the image and can help divide images into sets of semantically similar images. For example, streets inside a city have high rise buildings on the side while highways generally have trees and plants besides the roadside. Our proposed descriptor helps encode the positional information of each category in the image and can be used for semantic contextual retrieval.

Given an image which has been labelled using the WKNN-MRF method, we consider a spatial pyramid of n levels over the labelled image. At level i in the pyramid, we divide I into a uniform grid of $d \times d$ cells where $d = 2^{i-1}$. Within each grid cell, we compute the distribution for each of the nL classes using the number of individual pixels in that grid cell which have been assigned that class. This results in a nL-bin histogram for a single grid cell. The class distribution values for each cell are normalized so that they sum to one. The histograms for all the grid cells in the spatial pyramid are concatenated together resulting in a image feature f_{seman} of length $nL \times C$ where $C = \sum_{i=1}^{n} 4^{i-1}$ is the total number of cells in the spatial pyramid.

A higher value for n will capture the details of the layout more precisely but be more prone to classification errors while a lower value for n would be less sensitive to errors in the labelling but does not encode the spatial position of the semantic categories as well. This approach of computing a semantic label-based descriptor is similar to [10]. However our method differs in the fact that we use a spatial pyramid over the labelled image instead of a single grid to encode the semantic label information and we do not include additional appearance information in the descriptor, because it has already been captured through other global image features (Section 3.5). Our method also differs from [4] who compute a superpixel-level semantic context descriptor as a normalized label histogram of neighbouring regions.

5.1. Semantic Retrieval Set

Global image features (GIST, color histograms and spatial pyramid over SIFT) were used to build retrieval set T_g in Section 3.5. We now use the semantic label descriptor f_{seman} introduced above to help us refine the quality of the retrieval set by exploiting the semantic context.

For each image I_k in the training set, we perform leave-one-out-classification on the image using the WKNN-MRF approach. Using the resultant semantic image labelling, we generate its corresponding semantic label descriptor f^k_{seman}. Similarily, for the query view I_q, we label it using WKNN-MRF method and compute the corresponding semantic label descriptor. We generate a new set of ranking for the images in training set T based on the distance between their semantic label descriptor and that of the query image. The ranking is computed in an ascending order of the semantic label descriptor distances. We can now use

this ranking in isolation or combine it with the rankings for other global image feature types as was done in Section 3.5 to obtain the semantic retrieval set T_s. Using the new retrieval set T_s, we once again perform semantic labelling on the image by the process described in Section 3.3- 3.4. This method is denoted as WLKNN-MRF in our experimental results. The WLKNN refers to a weighted k-NN using a retrieval set built using the label descriptor only. We also experiment with using the semantic layout descriptor with all the other three global image features for the building of the retrieval set and denote this method WAKNN-MRF.

6. Experiments

For evaluating the performance of our method, we tested and compared it with several state-of-the-art techniques on four different datasets: SiftFlow [15], SUN09 [1], Google Street View [30] and Stanford Background [6]. The evaluation criterion for the methods is the per pixel accuracy (percentage of pixels correctly labelled) and per class accuracy (the average of semantic category accuracies).

For Stanford Background and Google Street View datasets, we selected 10% of the training images as the size of our retrieval set. In case of the other two datasets, we used a retrieval set of 75 images. For all our experiments, we set $k = 9$ in Eq. (4) and $\lambda = 0.4$ in Eq. (6). We obtained these parameters by selecting a small subset of the training images as a validation set. Computation of the feature weights required an average of four minutes for a single query image. To help speed up the computation of the weights, we approximate the neighbourhood construction of [3] through k-d trees [19]. For the query view, we index the individual features from the retrieval set in a k-d tree, constructing one k-d tree per feature channel. The neighbourhood computation is then approximated using the set union of the k-NN from different feature channels. We carry out 5 iterations of the weight computation step in Eq. (11) adaptively changing the nearest neighbours in the weighted neighbourhood space. While this approximates the weight computation, it affected our performance only slightly (a maximum decrease of 0.4% in per-pixel accuracy across the three datasets) and helped reduce the time for weight computation for an image to 20 seconds. For an image, feature computation, k-NN likelihood computation and MRF inference took 1 second, 13 seconds and 0.5 second respectively. When reporting the performance, we used the following variants of our approach:

- UKNN-MRF: uniform weights for the features with retrieval set obtained by global image features
- WKNN-MRF: computed weights for the features with retrieval set obtained by global image features
- WLKNN-MRF: computed weights with retrieval set built using the semantic layout descriptor only
- WAKNN-MRF: computed weights with retrieval set built using a union of semantic layout descriptor and the three other global image features.

SiftFlow SiftFlow is a large dataset of 2688 images with 33 semantic categories. [15] split the dataset into 2488 training images and 200 test images. Table 1 reports our performance on this dataset. Our weighted k-NN MRF performs on a comparable level on the per-pixel accuracy with the top methods. However it still trails [4] for the per-class accuracy. When we incorporate semantic context to obtain a refined retrieval set, our system achieves the best performance for both per-pixel and per-class accuracies. The categories which saw an increase of more than 10% after the use of semantic context include *field, car, river, plant, sidewalk, bridge, door, crosswalk*. These are categories which do not occur very frequently but achieved improved labelling with the context. For example, identifying road and highways helps label cars, sidewalk and crosswalk.

System	Per-Pixel	Per-Class
Liu et al. [15]	76.7	-
Tighe et al. [26]	76.9	29.4
Eigen et al. [4]	77.1	32.5
UKNN-MRF	75.6	27.9
WKNN-MRF	77.2	29.3
WLKNN-MRF	78.5	32.0
WAKNN-MRF	79.2	33.8
WKNN-MRF (with HOG)	76.7	27.4

Table 1. Semantic labelling performance on the SiftFlow dataset

We also experimented with replacing the SIFT feature for the superpixel with a HOG feature [2]. This feature was computed by using a 4×4 spatial grid of 4-pixel HOG cells with the grid centered at the superpixel's center. The individual HOG cell descriptors were averaged to compute the superpixel feature. The last row in Table 1 contains the performance for this method. Classes which significantly improved with the use of HOG instead of SIFT include *tree, mountain, car* while the accuracy dropped for *road, sea, grass, sidewalk*.

SUN09 SUN09 dataset [1] has fully labelled per-pixel ground truth for a set of 107 semantic categories. In the experiments, the dataset was split into 4352 training images and 4310 test images. Table 2 reports the performance of our method on this dataset. Using the semantic context helped obtain an improvement of 3.6% compared to the WKNN-MRF method. In comparison to [25], we perform better on per-pixel accuracy but trail on per-class accuracy. It was observed that the per-pixel labelling accuracy of outdoor scenes was more than 11% better than indoor scenes highlighting the challenge of labelling indoor views.

Google-StreetView The Google Street View dataset contains 320 images selected from a set of 10,000 images

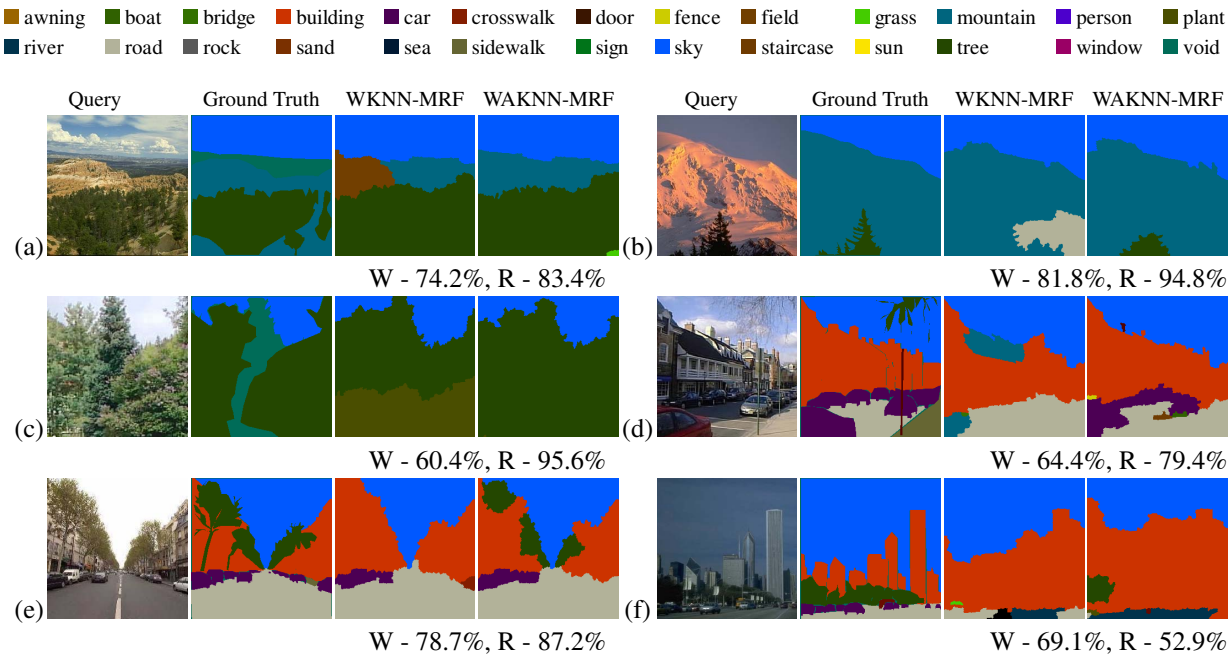

Figure 1. Example results from the SiftFlow test set (best viewed in color). W and R are per-pixel accuracies for WKNN-MRF and WAKNN-MRF respectively. Examples (a)-(c) are instances of semantic context improving the labelling as trees and mountains are predicted in the initial labelling. (d)-(e) show improvement in streets as road and buildings are predicted. Example (f) is an instance of degradation in performance as the initial prediction suggested higher percentage of buildings in the scene.

captured in Pittsburgh. The labelled data is equally split into a train and test set of 160 images. Table 2 contains the performance of our WAKNN-MRF approach against other methods. The individual class accuracies are: *building* 94%, *ground* 96.7%, *tree* 31%, *sky* 93.9% and *car* 66.2%. In comparison to the other methods, our performance was in the top-two for the per-pixel accuracy and for two semantic categories.

Stanford-Background This dataset contains 715 images with two separate label sets; semantic and geometric. We conducted our experiments for predicting the semantic category only. The semantic classes include seven background classes and a generic foreground class. Experiments on this dataset are conducted over five different splits with each split containing 572 training images and 143 test images. Table 2 summarizes our performance on this dataset. The use of semantic context leads to an improvement of only 0.5% and our results trail other state of the art methods. The lack of significant improvement with the use of semantic context here can be explained by the nature of the dataset as more than 90% of the images contain 4 or more of the 8 semantic categories.

Dataset and System	Per-Pixel	Per-Class
SUN09		
Choi et al. [1]	33.0	10.6
[25] CascALE Expert	49.3	16.7
[25] CascALE Sharing	52.8	15.2
WKNN-MRF	49.5	8.7
WAKNN-MRF	53.1	12.1
Google Streetview		
Zhang et al. [32]	88.4	80.4
Zhang et al. [31]	93.2	73.1
Singh et al. [24]	94.4	81
WAKNN-MRF	93.7	76.4
Stanford background		
[6] Pixel CRF	74.3	66.6
[6] Region Energy	76.4	65.5
[20] Leaf Level	72.8	58.0
[20] Hierarchy	76.9	66.2
WKNN-MRF	73.6	61.2
WAKNN-MRF	74.1	62.2

Table 2. Performance on the SUN09, Google-Streetview and Stanford background datasets.

7. Conclusions

We have presented an approach for nonparametric scene parsing using a k-NN method. We formulate our approach over small patches characterized by simple features which draws inspiration from biological vision. A locally adaptive distance metric is learned at query time to compute the relevance of individual feature channels. Using the initial

labelling as a contextual cue for presence or absence of objects in the scene, we proposed a semantic context descriptor which helped refine the quality of the retrieval set which is a key component of nonparametric methods. The approach was validated by experiments on several datasets and the performance was equal to or achieved an improvement over state of the art techniques. The nonparametric nature of our approach enables it to be flexible as more training data is available as no retraining is required. For future work, we would like to explore better methods for incorporating spatial information at the patch level and also explore learning semantic concepts for scene understanding.

Acknowledgements We are grateful for the comments from the anonymous reviewers. We also thank Paul Sturgess for sharing the pixel level annotations for the SUN09 dataset. This work was supported by Army Research Office Grant W911NF-1110476.

References

[1] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, pages 129–136, 2010.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[3] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. *PAMI*, 24(9):1281–1285, 2002.

[4] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *CVPR*, pages 2799–2806, 2012.

[5] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*, pages 417–424, 2006.

[6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8, 2009.

[7] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 80(3):300–316, 2008.

[8] S. Gould and Y. Zhang. PatchMatchGraph: Building a graph of dense patch correspondences for label transfer. In *ECCV (5)*, pages 439–452, 2012.

[9] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009.

[10] Q. Huang, M. Han, B. Wu, and S. Ioffe. A hierarchical conditional random field model for labeling and segmenting images of street scenes. In *CVPR*, pages 1953–1960, 2011.

[11] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, pages 739–746, 2009.

[12] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV (5)*, pages 239–253, 2010.

[13] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and CRFs. In *ECCV (4)*, pages 424–437, 2010.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[15] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 33(12):2368–2382, 2011.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[17] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, pages 1–8, 2008.

[18] B. Micusik, J. Košecká, and G. Singh. Semantic parsing of street scenes from video. *International Journal of Robotics Research*, 31(4):484–497, 2012.

[19] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP (1)*, pages 331–340, 2009.

[20] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *ECCV (6)*, pages 57–70, 2010.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[22] D. Ramanan and S. Baker. Local distance functions: A taxonomy, new algorithms, and an evaluation. *PAMI*, 33(4):794–806, 2011.

[23] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

[24] G. Singh and J. Košecká. Acquiring semantics induced topology in urban environments. In *ICRA*, pages 3509–3514, 2012.

[25] P. Sturgess, L. Ladicky, N. Crook, and P. Torr. Scalable cascade inference for semantic image segmentation. In *BMVC*, pages 62.1–62.10, 2012.

[26] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *ECCV (5)*, pages 352–365, 2010.

[27] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[28] T. Werner. A linear programming approach to max-sum problem: A review. *PAMI*, 29(7):1165–1179, 2007.

[29] H. Wildenauer, B. Micusík, and M. Vincze. Efficient texture representation using multi-scale regions. In *ACCV (1)*, pages 65–74, 2007.

[30] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *ICCV*, pages 686–693, 2009.

[31] H. Zhang, T. Fang, X. Chen, Q. Zhao, and L. Quan. Partial similarity based nonparametric scene parsing in certain environment. In *CVPR*, pages 2241–2248, 2011.

[32] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *ECCV (5)*, pages 561–574, 2010.