Semantic segmentation of street scenes by superpixel co-occurrence and 3D geometry*

Branislav Mičušík[1,2] Jana Košecká[2]
[1]AIT Austrian Institute of Technology, Video and Security Technology Unit, Vienna, Austria
[2]George Mason University, Computer Science Department, Fairfax, USA

Abstract

We present a novel approach for image semantic segmentation of street scenes into coherent regions, while simultaneously categorizing each region as one of the predefined categories representing commonly encountered object and background classes. We formulate the segmentation on small blob-based superpixels and exploit a visual vocabulary tree as an intermediate image representation. The main novelty of this generative approach is the introduction of an explicit model of spatial co-occurrence of visual words associated with super-pixels and utilization of appearance, geometry and contextual cues in a probabilistic framework. We demonstrate how individual cues contribute towards global segmentation accuracy and how their combination yields superior performance to the best known method on the challenging benchmark dataset which exhibits diversity of street scenes with varying viewpoints, large number of categories, captured in daylight and dusk.

1. Introduction

Combining object segmentation and recognition is one of the fundamental problems in computer vision. This area has been particularly active in recent years, due to the development of methods for integration of object specific techniques, with various contextual cues and top down information. Despite notable progress in this area, challenges remain, when large number of classes occur simultaneously, objects vary dramatically in size and shape, often comprising of small number of pixels, and scene varies dramatically in viewpoint. In this work we focus on the semantic segmentation of street scenes, which poses all the difficult characteristics mentioned above.

Recent development in large scale modeling of cities and urban areas has predominantly focused on the creation of 3D models. Attempts to detect objects in the images of urban areas mainly focused on cars, pedestrians, faces, car plates, and used standard window based object detector pipelines. In this paper we present a novel approach for semantic labeling of street scenes, with a goal of automatically annotating different regions by labels of commonly encountered object and background categories, see Fig. 1. The work presented here naturally extends multi-class segmentation methods, where one seeks to simultaneously segment and associate semantic labels with individual pixels. It is also closely related to approaches for scene analysis, where one aims at integration of local, global and contextual information across image regions of varying size to facilitate better object recognition.

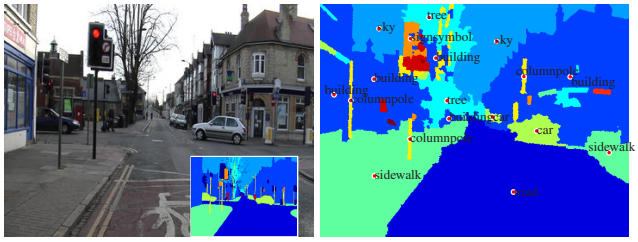


Figure 1. Semantic segmentation of a street sequence. Left: An input image with ground-truth in-set. Right: Output of the proposed method.

The main contribution of our approach is in *i)* exploitation of large visual vocabularies for representation and segmentation of object and background categories, *ii)* a novel representation of local contextual information using spatial co-occurrence of visual words, and *iii)* use of an image segmentation into small superpixels for selection of locations where the descriptors are computed. These ingredients are integrated in a probabilistic framework yielding a second-order Markov Random Field (MRF), where the final labeling is obtained as a MAP solution of the labels given an image. We show how the effort to capture the local spatial context of visual words is of fundamental importance, providing interesting insight into part based representation. The street scenes are particularly challenging and of interest

*This research received funding from the US National Science Foundation Grant No. IIS-0347774 and the Wiener Wissenschafts-, Forschungs- und Technologiefonds - WWTF, Project No. ICT08-030.

to a variety of applications which strive to associate meta-data with the street scene imagery. We show a substantial gain in the global segmentation accuracy compared to the best known method on the street benchmark dataset.

Related work Our work is related to several efforts in multi-class segmentation, visual representations of object/background categories, context modeling for scene analysis. The existing work on multi-class segmentation typically differs in the choice of elementary *regions* for which the labels are sought, the types of *features* which are used to characterize them, and means of integrating the *spatial* information. We will next elaborate on all of these aspects and point out the differences between previous work and the approach presented here.

In [1, 16, 8, 9] authors used larger windows or superpixels, which are characterized by specific features such as color, texture moments or histograms, dominant orientations, shape *etc*., where the likelihoods of the observations are typically obtained in discriminative setting. Another direction is to use highly discriminative features defined on sampled isolated pixels in associated windows, obtained by training randomized trees [10, 19]. We instead choose as features the SIFT descriptors [11] augmented with color information. These features are computed at centers of small blob-based superpixels, which are obtained by watershed segmentation on LoG interest points as seeds. The descriptors are then organized hierarchically in a vocabulary tree. The choice of this generative model enables us to benefit from the extensive body of work [11, 21, 15] on invariance properties, quantization, and retrieval utilizing large visual vocabularies to further improve efficiency and scalability of the method.

While the SIFT features have been used previously in the context of semantic labeling, object detection and segmentation, their information was either integrated over large super-pixels [16] or large rectangular regions over a regular point grid [7]. Integration of the descriptors over large spatial regions has been shown to improve performance on the Graz-2, MSRC21 and PASCAL datasets, where the superpixels were represented by histogram of signatures of SIFT descriptors. The images in these data sets have mostly small number (2-5) of object/background classes in the image and there is typically one object in the center, which takes dominant portion of the image, see Fig. 2. In the presence of larger number of smaller objects, the approach of [7] is not suitable and the strategy of [16] critically depends on the initial segmentation as its success relies on superpixels having boundaries aligned with object/background. This is very challenging in street scenes, due to the presence of large number of small and narrow structures, such as column poles, signs *etc*.

At last, several means of integrating spatial relationships between elementary regions have been proposed in the past, *e.g*. correlograms [18], texture layout filters [20], CRF with relative location prior [8], enforcing full CRF connectivity between large superpixels [16], enforcing object appearance and position in a probabilistic graphical model [22], initial scene alignment by global image features [17], or utilization of higher order potentials [9]. The higher order potentials, motivated by overcoming the smoothing properties of the CRFs with pairwise potentials, has been used to integrate results from multiple segmentations, to obtain crisper boundaries, and to improve the error due to an incorrect initial segmentation. The majority of approaches for exploiting spatial relationships use either class co-occurrence information between regions [16, 8] or use large spatial support from neighboring regions as an additional evidence for the region label [18, 20]. These approaches for capturing global contextual information about spatial co-occurrence of different class label are meaningful when the number of classes per image and the change of the viewpoint are relatively small as in the MSRC21. There, the cars/cows typically appear next to road/grass and below the sky. In the street scenes with the larger number of object categories and larger changes in viewpoint, these types of contextual relationships are no longer so persistent. For example, cars can appear next to sky, building, tree, another car, pedestrian, all simultaneously at different locations in the images and are often of very small size. That makes also the location prior [8] less feasible. This can be seen in Fig. 2, where the spatial class co-occurrences of the two datasets are compared. It has also been demonstrated in [3] that performance of the TextonBoost [20] on the CamVid street dataset drops by 6% compared to the MSRC21. This indicates that the considered class of street scenes deserves special attention. In our approach we, instead of modeling co-occurrences or spatial locations of class labels, exploit spatial co-occurrences between visual words of neighboring superpixels.

Figure 2. Class-class co-occurrence matrices with example images for the 11-class CamVid and 21-class MSRC21 dataset. Rows and columns of the matrices correspond to class labels and the numbers stand for class-class co-occurrence frequency in the datasets. White color stands for zero occurrence. Notice the sparsity of the matrix for the MSRC dataset, meaning that usually only 2-5 objects are present in the image while in the CamVid usually all 11 objects appear simultaneously.

2. Semantic segmentation

We formulate the semantic segmentation on an image oversegmented into a disjoint set of small superpixels. Our elementary regions are computed by watershed segmentation on LoG interest points as seeds and can be seen in Figure 4. These elementary regions typically do not straddle boundaries between different classes and naturally correspond to semantically meaningful object (scene) primitives/parts. Furthermore, they dramatically reduce computational complexity of an MRF inference. Superpixels have been used in the past extensively as intermediate primitives in various formulations of image parsing and object recognition tasks.

The output of the semantic segmentation is a labeling vector $\mathbf{L} = (l_1, l_2, \dots l_S)^\top$ with hidden variables assigning each superpixel i one unique label, $l_i \in \{1, 2, \dots, L\}$, where L and S is total number of the labels/classes and superpixels respectively. The posterior probability of a labeling $\mathbf{L}$ given the observed appearance and geometry feature vectors $\mathtt{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_S]$, $\mathtt{G} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_S]$ for each superpixel can be expressed as

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

The appearance features in our case are SIFT descriptors computed at the centers of small superpixels and quantized using commonly used vocabulary tree. This yields a visual vocabulary with V words and we can associate with each descriptor $\mathbf{a}_i$ an index corresponding to the nearest visual word from the vocabulary. Hence in the subsequent discussion we approximate the matrix $\mathtt{A}$ by the vector of scalar indexes $\mathbf{V} = (v_1, v_2, \dots, v_S)^\top$. Details of the vocabulary tree construction are described in Sec. 3.2.

We estimate the labeling $\mathbf{L}$ as a Maximum Aposteriori Probability (MAP),

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

We assume independence between the appearance and geometry features factorizing the likelihood into two parts, the appearance $P(\mathbf{V}|\mathbf{L})$ and geometry $P(\mathtt{G}|\mathbf{L})$ likelihood. All terms, the observation likelihoods and joint prior, are described in the following subsections.

2.1. Appearance Likelihood

The observation appearance likelihood $P(\mathbf{V}|\mathbf{L})$ can be expressed using a chain rule as

$$P(v_1|l_1) P(v_2|v_1, l_1, l_2) \dots P(v_S|v_1, v_2, \dots, v_{S-1}, \mathbf{L}).$$

However, learning or just setting of the high-order conditional dependencies is intractable. A commonly utilized and also the simplest approximation is to make a Naive Bayes assumption, yielding

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

Such an approximation assumes independence between visual words (SIFT descriptors) of superpixels given their labels.

This assumption may be partially overcome by a proper design of the smoothness term $P(\mathbf{L})$ enforcing class co-occurrence prior. However, the smoothness term is typically pairwise because of commonly utilized second-order Markov Random Field capturing thus only local relations of neighboring superpixels.

An alternative strategy, investigated in the past when evaluating $P(v_i|l_i)$, is to gather measurements from superpixels farther from the investigated one. An example is a relative location prior [8] used in a CRF based framework where each superpixel gets votes from all the remaining. Another example is spatial layout filters capturing texton co-occurrence [20]. Alternative solution is to explicitly model the dependencies between the visual words which co-occur in the image, by treating them as random variables and trying to approximate the dependencies between them. A version of this alternative was to employ a tree based approximation of the distribution using *e.g.*, the Chow-Liu dependence tree [4]. The tree approximates a discrete distribution by the closest tree-structured Bayesian network and is obtained as a maximum spanning tree on a fully connected graph with visual words as vertices and mutual co-occurrence gathered from a training set as edge weights. With this type of approximation, the appearance observation likelihood would read as

$$P(\mathbf{V}|\mathbf{L}) \approx P(\mathbf{Z}|\mathbf{L}) = P(z_r|l_r) \prod_{i=2}^{V} P(z_i|z_{p_i}, l_i, l_{p_i}), \tag{4}$$

where $\mathbf{Z} = (z_1, \dots, z_V)^\top$ is a vector of binary variables indicating presence or absence of the i-th word of the vocabulary, z_r is a root, z_{p_i} is a parent of z_i in the Chow-Liu tree, and V is the total number of visual words. This strategy watches whether: "When a particular visual word is observed, is the most likely word from the learned tree observed too?" It has been shown in [4] in connection of location recognition that, if using a vocabulary tree and image signatures, such an approximation yields substantially better results than the Naive Bayes.

In their setting they do not perform any image segmentation and gather co-occurrences between visual words which appear simultaneously in the image without considering spatial relationships. Furthermore in the presence of multiple labels $\mathbf{L}$ per image as in our case, the learning and

keeping the CPD tables becomes intractable. Motivated by drawbacks of previously discussed methods, we propose a new way of representing the likelihood as

$$P(\mathbf{V}|\mathbf{L}) \approx \prod_{i=1}^{S} P(v_i|\mathcal{B}_i, l_i), \quad (5)$$

where $\mathcal{B}_i$ is a subset of most likely visual words associated with superpixels from a 2-depth neighborhood which appear together with the superpixel i. The 2-depth neighborhood of a superpixel i contains superpixels which are its direct neighbors and superpixels which are neighbors of the direct neighbors. Two superpixels are said to be neighbors if they share at least one pixel in an 8-point neighborhood connectivity. In particular for each class label we learn in the training stage a probability of co-occurrence of different visual words, which are encountered in the same spatial neighborhood. The details of this stage are given in Sec. 2.2.

The conditional probability $P(v_i|\mathcal{B}_i, l_i)$ is set as an average of the CPDs $P(v_i|l_i)$ of the investigated superpixel i and its most likely class neighbors from $\mathcal{B}_i$,

$$P(v_i|\mathcal{B}_i, l_i) = \frac{1}{|\mathcal{B}_i| + 1}\Big(P(v_i|l_i) + \sum_{j \in \mathcal{B}_i} P(v_j|l_i)\Big), \quad (6)$$

where the CPDs $P(v_i|l_i)$ are set according to the learned vocabulary tree, more in Sec. 3.2.

We show that this model with spatial co-occurrence statistics outperforms the Naive Bayes from Eq. (3). It is very important to notice that our formulation takes into account spatial co-occurrence of visual words as we collect statistics over a controlled neighborhood of the superpixel. In contrary to the Chow-Liu tree, when evaluating likelihood of a particular superpixel, we consider more superpixels, resp. visual words, than just one (node - parent relation only, see Eq. (4)). That gives us more robust measurement and still enables to encode spatial co-occurrence information.

2.2. Word co-occurrence matrix

Mutual spatial appearance of visual words in an image is not random but depends on the scene and especially on objects we observe. For example, a visual word corresponding to a car front light has a high probability to appear together with a car plate or a bumper. It logically follows that when inferring a class label for each superpixel, taking into account also neighboring superpixels can resolve many ambiguities. The question arises how to select the most informative neighbors when inferring a label for a given superpixel.

Such an observation is not surprising and many authors have proposed partial solutions. In [16] they propose to use large superpixels and create superpixel signatures from all the SIFTs points [11] detected in the superpixel. The problem is that on one side the superpixels must be large and contain enough SIFTs from one object class to get a meaningful signature, on the other hand, they must be small enough to capture small structures and to avoid oversegmentation. In the sequences we are interested in, the objects (cars, pedestrians, bicyclists) are small comprising of small number of pixels and therefore very small superpixels are needed to capture them. Another way to compute a signature is to encompass all points in a large fixed squared window around the point [7]. It is clear that such strategy fails when most of the square pixels are drawn from another object/class than the inferring pixel. This happens especially for pixels on the object boundaries.

We therefore propose a different strategy, accounting for the aforementioned drawbacks, by employing visual word co-occurrence statistics. When computing an observation likelihood of a particular label l_i of a superpixel we consider only those neighbors which are most likely to appear in the close proximity of the investigated superpixel given a particular class. To evaluate this, we learn in the training stage the statistics of words with same labels appearing together in some pre-defined neighborhood. This requires ground truth labels being associated with all regions of the training images.

We learn a visual word co-occurrence matrix $\mathtt{C}_d$ with a dimension $V_d \times V_d \times L$, where V_d stands for number of visual words at d-th depth of the vocabulary tree and L for number of classes/labels. Learning the co-occurrence at a single depth is sufficient as the information overlaps when traversing the vocabulary tree. The depth level selection is a trade-off between memory requirements and distinctiveness. Going down the tree, the matrix becomes larger and sparser; going up the tree, the matrix looses distinctiveness. Given the depth d the matrix $\mathtt{C}^d$ is constructed as follows.

1. Repeat steps 2, 3 for all training images and their superpixels having assigned a class label.
2. Consider a superpixel i and find all its N 2-depth neighbors $j = 1 \dots N$. Push the descriptors $\mathbf{a}_i$ and $\mathbf{a}_j$ down the vocabulary tree, yielding corresponding visual words v_i^d, v_j^d at the depth d.
3. Increment $\mathtt{C}^d[v_i^d, v_j^d, l_i]$ for all j which have $l_i = l_j$.
4. After building the $\mathtt{C}^d$, normalize it such that each row sums to one. Each row is then an approximation of the probability $P(v_j|v_i, l_i, l_i = l_j)$.

The co-occurrence matrix $\mathtt{C}^d$ is employed in the evaluation of Eq. (6) and is used to find the set $\mathcal{B}_i$ of most likely neighbors of a particular superpixel i when evaluating the observation likelihood given a label l_i. The set is obtained by taking the top B best candidates from the 2-depth neighborhood of superpixels j, in sense of the highest co-occurrence probability $\mathtt{C}_d[v_i^d, v_j^d, l_i]$.

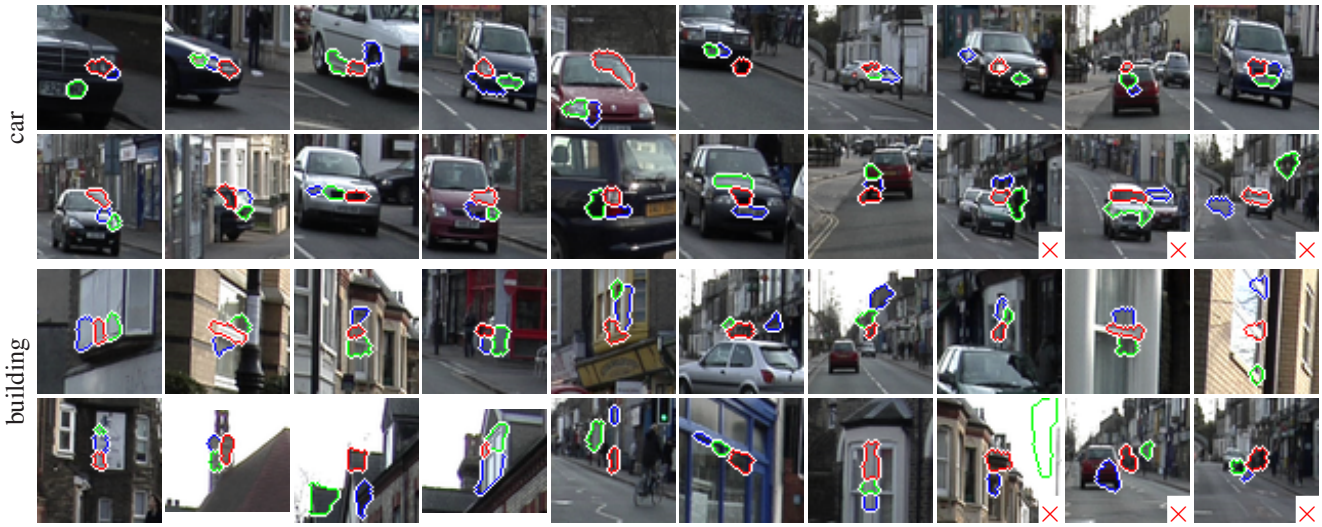


Figure 3. Superpixel co-occurrence. When inferring superpixels in query images, shown in red and centered in the cutouts, for a given class (car, building) the neighboring superpixels being most likely the same class are found by the learned co-occurrence matrix and further utilized in setting of the superpixel likelihood. The top 2 most likely superpixels are shown in green and blue. Notice the semantic dependency captured by the co-occurrence matrix. The last three images in each set show incorrect neighboring superpixels coming from different class than the inferring superpixel.

To demonstrate the co-occurrence dependencies, Fig. 3 shows the top $B = 2$ neighbors for some superpixels from test images corresponding to a car and a building class. One can see a clearly structural dependency of the parts captured by the learned co-occurrence matrix. For example, visual words corresponding to car elements like lights, bumpers, windows, plates, tires are learned to appear together. The same is observed for the building class and dependencies between facade elements, window inner structures, roof shapes. In all our experiments we computed the visual word co-occurrence matrix at one before the last depth level d of the tree with 1k leaves, thus $V_d = 1000$. We experimented with one higher and one lower level, however, both got worse results.

2.3. Geometry Likelihood

Since we are interested in sequences we can gain from additional temporal information and use more than just appearance cues from signle images. Similarly to [3] we propose to employ 3D information from SfM estimation to improve the recognition accuracy. We adopted one most discriminative feature from [3], the height of reconstructed 3D points above a camera center, to show the benefit of 3D geometric information.

Let us assume that we know 3D positions of some image points expressed in a common coordinate system along the sequence and that we can compute the height of these points above the camera center. Since 3D information is recovered at a sparse set of points, each superpixel has a binary geometry observation variable o, indicating whether there was any point projected in the superpixel. The geometry likelihood reads as

$$P(\mathtt{G}|\mathbf{L}) = \prod_{i=1}^{S} P(g_i|l_i, o_i), \tag{7}$$

where $P(g|l, o = 1)$ is learned from the training set from reconstructed 3D points with known labels. The probability distribution is estimated as a normalized histogram on the 3D point heights for each class l separately utilizing Parzen windows with the Gaussian kernel. If there are more reconstructed points in the superpixel, we consider the average height $\bar{g}_i$ of all of the points and evaluate $P(\bar{g}_i|l_i, o_i)$. If there are no reconstructed 3D points in the superpixel, indicated by $o = 0$, then $P(g_i|l, o = 0)$ is set to uniform prior $1/L$ for all l.

2.4. Joint Prior

The joint prior $P(\mathbf{L})$, or the smoothness term, is approximated by pairwise potentials as

$$P(\mathbf{L}) \approx \exp\Big(\sum_{(i,j)\in\mathcal{E}} g(i,j)\Big), \tag{8}$$

where the pairwise affinity function g is defined as

$$g(i,j) = \begin{cases} 1-e, & \text{iff } l_i = l_j \\ \delta + e, & \text{otherwise,} \end{cases} \tag{9}$$

with $e = \exp(-\|\mathbf{c}_i - \mathbf{c}_j\|^2/2\sigma^2)$, where $\mathbf{c}_i$ and $\mathbf{c}_j$ are 3-element vectors of mean colors expressed in the Lab color

space for i-th and j-th superpixel, respectively, and σ is a parameter set to 0.1. The set $\mathcal{E}$ contains all neighboring superpixel pairs.

The smoothness term is a combination of the Potts model penalizing different pairwise labels by the parameter δ and a color similarity based term. The aim is on one side to keep the same labels for neighboring superpixels, and on the other, to penalize same labels if they have different color. We have set δ to 0.8 in our experiments.

2.5. Inference

We formulated both, the observation likelihood and joint prior from Eq. (2), as unary and binary functions used in a second-order MRF framework. The maximization in Eq. (2) can be re-written in a log-space and the optimal labeling $\mathbf{L}^*$ achieved as

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

where $E_{app} = -\log P(v_i|\mathcal{B}_i, l_i)$ from Eq. (5), $E_{geom} = -\log P(g_i|l_i, o_i)$ from Eq. (7), and $E_{smooth} = g(i,j)$ from Eq. (8). The scalars λ_g, λ_s are the weighting constants of importance of the terms (set to 1 and 0.2 in our experiments).

We perform the inference in the MRF, *i.e.* a search for a MAP assignment, by efficient and fast publicly available MAX-SUM solver [24] based on linear programming relaxation and its Lagrangian dual. Although, finding a global optimum of the equation above is not guaranteed, as the problem is NP-hard, it has been shown that the provided solution is often a strong optimum.

3. Image representation

3.1. Superpixels

In a search for good superpixels, we were motivated by a success and popularity of SIFT features built on LoG (approximated by DoG) extrema points [11] in bag-of-feature based image retrieval systems [21]. In semantic segmentation we face slightly different problem. We need to assign a class label to each pixel and not only to DoG extrema points. We therefore utilize a segmentation method [5, 25] where a superpixel boundaries are obtained as watersheds on negative absolute Laplacian image with LoG extremas as seeds. Watershed transformation has been employed in the successful MSER detector [12], widely used to complement DoG extrema points for SIFT feature computation. Such blob-based superpixels are effective to compute, are regularly shaped and follow image edges, see Fig. 4. As we will show those superpixels are superior than regular or widely used Felzenszwalb's [6] superpixels.

Figure 4. The watershed LoG based superpixels shown on a part of the image from Fig. 1.

Each superpixel is assigned a 131-dimensional feature vector $\mathbf{a}_i$ consisting of 128 dimensional SIFT [11] and of 3 mean color components over the superpixel pixels expressed in the Lab color space to preserve color information. The SIFTs are computed on superpixel centroids at a fixed scale (support region of 12×12 pixels) and orientation using implementation from [23]. The absence of non-rotation invariance is a desired property as it increases distinctiveness, however, requiring a rich training set with large variations in a viewpoint.

3.2. Vocabulary tree

Recent work in an object based image retrieval has shown significant progress by utilizing a bag-of-feature model based on quantization of high-dimensional region descriptors into visual words [21, 14, 15]. The model exhibits high discriminative power, scalability to large datasets and computational efficiency in the inference stage. The same bag-of-feature model has been adapted recently for semantic segmentation [16, 7] and has been shown as a promising way of object representation.

We adapted the bag-of-feature model based on hierarchical K-means proposed by [14] for representing superpixels by their visual word indeces v_i. A vocabulary tree is created from large representative set of superpixels by K-mean clustering of region descriptors $\mathbf{a}_i$ into K clusters at each level of the tree; giving K^d nodes at the depth level d. In our experiments, we use max depth $D = 4$ with 10k leaf nodes, using implementation of [23]. To reduce high computational burden of K-mean clustering in the tree building stage we use only superpixels from 10% of all training images. Then, we push all the superpixel feature vectors from the training set with known class labels down the tree and count the occurrence of labels at each tree leaf. To get the probability $P(v|l)$, stored as a $L \times V$ matrix $\mathtt{M}$, one needs to normalize the leaf-class occurrences over rows such that $\forall l = \{1, \ldots, L\} : \sum_{v=1}^{V} P(v|l) = 1$.

In the inference stage, to find an approximate nearest leaf cluster to a given superpixel i, we let the feature vector descend the tree and consider the approached leaf as the corresponding visual word v_i. Then, the probability $P(v_i|l_i)$, utilized in Eq. (6), is a number corresponding to the column v_i and the row l_i in the matrix $\mathtt{M}$. The quantization and ap-

proximation effects of the vocabulary trees in connection to image retrieval have been studied in [15, 13] with some partial solutions which potentially offers a room for a further improvement.

4. Experiments

We evaluated our semantic segmentation on a challenging new publicly available database of complex driving scenes, the Cambridge-driving Labeled Video Database (CamVid) introduced by [2]. This database is the first collection of videos with object class semantic labels and SfM data obtained by tracking. The database provides ground truth labels that associate each pixel with one of 32 semantic classes which we grouped into 11 larger ones for our experiments to better reflect the statistically significant classes and to be consistent with the results published in [3]. The CamVid dataset was captured from the perspective of a driving automobile in daylight and dusk. The driving scenario increases the number and heterogeneity of the observed object classes. Over 10 min of 30 Hz footage is being provided, with corresponding semantically labeled images at 1 Hz and in part, 15 Hz. In our experiments, we downsampled the images to 320x240 pixels. We speculate that full resolution would improve scores for some of the smaller classes.

We trained our method on 305 day and 62 dusk images, and tested on 171 day and 62 dusk images, same setup as has been presented in [3]. Qualitative and quantitative results are shown in Fig. 5 and in Tab. 1, respectively. Fig. 5 shows the same images as in [3] to demonstrate visually more plausible segmentations, reader is referred to the results in that paper. The accuracy in Tab. 1 is computed by comparing the ground truth pixels to the automatically obtained segmentations. We report per-class accuracy as the normalized diagonal of the pixel-wise confusion matrix, the class average accuracy, and the global segmentation accuracy.

We compare our results to the state-of-the-art method reported in [3] where they utilize an appearance model based on the TextonBoost [20] and five geometry features. Employing our appearance model and only *one* geometry feature we are better in 6 classes and get an important 8% gain in the global accuracy, see Tab. 1. The reason why we achieve only the same average accuracy is due to weak performance in two classes, a fence, a sign-symbol. The average accuracy measure applies equal importance to all classes and is thus more strict than the global accuracy considering class prevalence. We presume that the weakness of recognizing those classes is their very small proportion in the training set, where each class is captured by only 1% of all pixels in the set. Since our approach falls into the category of generative techniques, the use of a vocabulary tree as an approximation of the distribution might underperform for classes with small number of training examples despite the proper normalization.

Tab. 1 shows an important contribution of this paper how employing the co-occurrence matrix significantly helps to increase accuracy of most classes, and obtain higher average and global accuracy. The row "Our, co-occ. & wshedLoG splxs" represents full model with co-occurrence statistics involved, whereas "Our, no co-occ. & wshedLoG splxs" stands for a Naive Bayes model. We experimented with value of B in Eq. (6), *i.e*. the number of considered most likely neighboring superpixels, and choice of $B = 5$ for all our tests gave us the best performance in sense of the highest average accuracy.

Furthermore, we experimented with different superpixels. First, with the regular ones obtained by splitting the image into 10×10 squares, motivated by the regular point grid for SIFT computation utilized in the semantic segmentation in [7]. Second, with widely used Felzenswalb's [6] superpixels based on minimum spanning tree on color differences with tuned parameters to get, in average, the same area of superpixels as the watershed based superpixels. The results shows that the watershed LoG based superpixels outperforms the others. At last, bottom of the Tab. 1 shows comparison when using only the appearance model. Compared to the state-of-the art appearance model, the TextonBoost [20], we are better in 6 classes, by 8% in global segmentation accuracy. However, because of the weak fence and sign-symbol classes, slightly worse in the average accuracy.

5. Conclusion

Capturing the co-occurrence statistics of visual words has been shown here to be an important cue towards improving semantic segmentation in difficult street scenes. We have presented a novel unified framework capturing both, the appearance and geometry features, defined at superpixels and have demonstrated superior results than the previous technique. The semantic segmentation of the street view scenes requires special attention because of their practical importance, difficulty, and impossibility of standard techniques to score equally well as on standard object datasets.

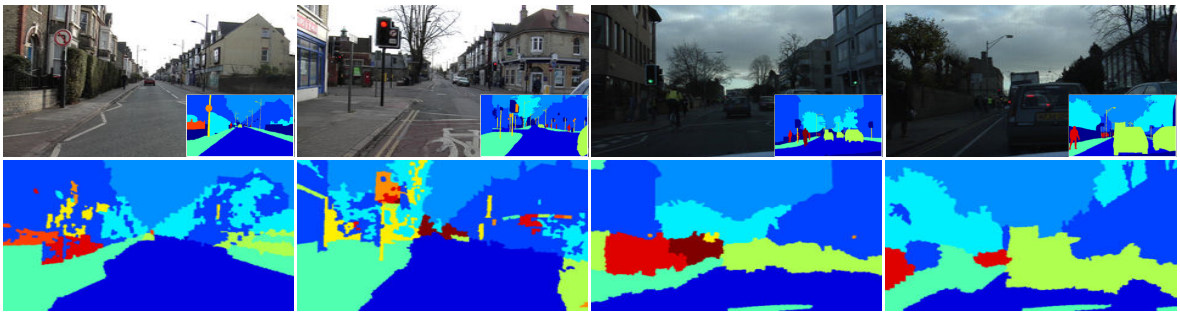
References

[1] A. C. Berg, F. Grabler, and J. Malik. Parsing images of architectural scenes. In *ICCV*, pages I: 44–57, 2007.

[2] G. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[3] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, pages I: 44–57, 2008.

[4] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *Int. Journal of Robotics Research*, 27(6):647–665, 2008.



(a) DayTest #0450 (b) DayTest #2460 (c) DuskTest #8550 (d) DuskTest #9180

Figure 5. Sample segmentation results on the CamVid dataset. On top the input images with ground-truth in-sets are shown.

	building	tree	sky	car	sign-symbol	road	pedestrian	fence	column-pole	sidewalk	bicyclist	Average	Global
Appearance + Geometry													
Textonboost + SfM [3]	46.2	**61.9**	**89.7**	68.6	**42.9**	**89.5**	53.6	**46.6**	0.7	60.5	22.5	53.0	69.1
Our, co-occ & wshedLoG splxs	**71.1**	56.1	89.5	**76.5**	12.5	88.4	**59.1**	4.8	**11.4**	**84.7**	**28.8**	53.0	**77.1**
Our, no co-occ & wshedLoG splxs	81.1	53.7	85.7	74.3	1.9	93.7	22.7	2.0	9.3	65.7	7.9	45.3	76.7
Our, co-occ & regular splxs	75.0	56.9	90.9	68.1	2.2	87.9	38.5	3.4	7.1	78.4	26.4	48.6	77.0
Our, co-occ & Felz. splxs [6]	61.9	60.0	94.2	72.4	12.9	89.6	56.5	2.8	26.1	83.1	15.0	52.2	76.4
Appearance only													
Textonboost [20]	38.7	60.7	**90.1**	**71.1**	**51.4**	**88.6**	**54.6**	**40.1**	1.1	55.5	23.6	**52.3**	66.5
Our, co-occ & wshedLoG splxs	**66.1**	**62.6**	88.2	70.8	9.4	84.0	49.3	3.1	**18.1**	**79.2**	**32.3**	51.2	**74.5**

Table 1. Semantic segmentation results in pixel-wise percentage accuracy on the CamVid dataset.

[5] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro. Principal curvature-based region detector for object recognition. In *CVPR*, pages 1–8, 2007.

[6] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[7] B. Fulkerson, A. Vedaldi, and S. Soatto. Localizing objects with smart dictionaries. In *ECCV*, pages I:179–192, 2008.

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

[9] P. Kohli, L. Ladický, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, pages I: 44–57, 2008.

[10] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9):1465–1479, 2006.

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

[12] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages I: 384–393, 2002.

[13] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009.

[14] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages II:2161–2168, 2006.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[16] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[17] B. C. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *NIPS*, pages I: 44–57, 2007.

[18] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, pages 2033–2040, 2006.

[19] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, 2008.

[20] J. Shotton, J. 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.

[21] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, pages II:1470–1477, 2003.

[22] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, pages II: 1331–1338, 2005.

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

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

[25] H. Wildenauer, B. Micusik, and M. Vincze. Efficient texture representation using multi-scale regions. In *ACCV*, pages 65–74, 2007.