EFFICIENT SEMANTIC SEGMENTATION OF MAN-MADE SCENES USING FULLY-CONNECTED CONDITIONAL RANDOM FIELD

Weihao Li[a], Michael Ying Yang[a,b*]

[a] TU Dresden, Computer Vision Lab Dresden, Dresden, Germany - weihao.li@mailbox.tu-dresden.de
[b] University of Twente, ITC Faculty, EOS department, Enschede, The Netherlands - yang@tnt.uni-hannover.de

Commission III, WG III/4

KEY WORDS: Man-made Scene, Semantic Segmentation, Fully Connected CRFs, Mean Field Inference

ABSTRACT:

In this paper we explore semantic segmentation of man-made scenes using fully connected conditional random field (CRF). Images of man-made scenes display strong contextual dependencies in the spatial structures. Fully connected CRFs can model long-range connections within the image of man-made scenes and make use of contextual information of scene structures. The pairwise edge potentials of fully connected CRF models are defined by a linear combination of Gaussian kernels. Using filter-based mean field algorithm, the inference is very efficient. Our experimental results demonstrate that fully connected CRF performs better than previous state-of-the-art approaches on both eTRIMS dataset and LabelMeFacade dataset.

1. INTRODUCTION

Semantic segmentation of man-made scenes is one of the fundamental problems in photogrammetry and computer vision. Man-made scenes, e.g. street scene, as shown in Figure 1, may be the most familiar scenes in our life. Applications of man-made scene interpretation include 3D city modeling, vision-based outdoor navigation, intelligent parking etc. Man-made scenes exhibit strong contextual and structural information in the form of spatial interactions among components, which may include buildings, cars, doors, pavements, roads, windows or vegetation. The eTRIMS (Korč and Förstner, 2009) and LabelMeFacade (Fröhlich et al., 2010, Brust et al., 2015) image databases are two popular dataset for man-made scene semantic segmentation, which have irregular facades and do not follow strong architectural principles. In this paper, we will explore semantic segmentation of this kind of man-made scenes using fully connected Conditional random fields (CRFs).

Conditional random field (CRF) (Lafferty et al., 2001, Shotton et al., 2006) is a popular method for modeling the spatial structure of images in semantic segmentation problem. The key idea of the semantic segmentation is to combine the low-level pixel object classifiers information and spatial contextual information within a CRF model, then running a maximum a posteriori (MAP) or maximum posterior marginal (MPM) inference method to obtain the segmentation results. However, low-connected standard (e.g. 4-connected or 8-connected) CRF works on a local level and cannot model the long range dependencies of the images, so the object boundaries of these results are excessive smoothing.

CRFs with higher-order potentials, i.e. P^n Potts model (Kohli et al., 2009), and hierarchical CRF (Ladicky et al., 2009, Yang and Förstner, 2011) have been proposed to improve semantic segmentation accuracy by enforcing label consistency in image segments (or superpixels). Both P^n Potts model and hierarchical CRF are based on unsupervised image segmentation, which is used to compute the segments or superpixels, e.g. normalized cuts (Shi and Malik, 2000), mean shift (Comaniciu and Meer, 2002) and SLIC (Achanta et al., 2012). However, accurate unsupervised image segmentation is still an unsolvable problem. Segment-based P^n Potts model and hierarchical CRF model are limited by the accuracy of these unsupervised image segmentation. Mistakes in the initial unsupervised image segmentation cannot be recovered in the inference step, if regions cross multiple object classes.

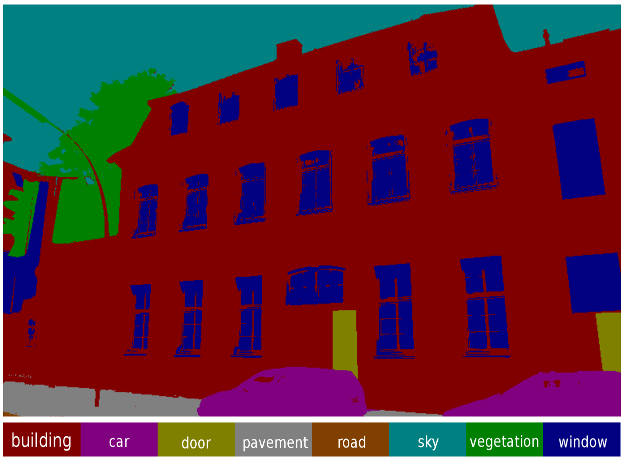

Figure 1. Example images of eTRIMS dataset (1st row), the corresponding semantic segmentation results using fully-connected CRF (2nd row), and ground truth labeling (3rd row). The last row is the legend.

Recently, the fully connected CRF (Krähenbühl and Koltun, 2011) gains popularity in the semantic segmentation problems. Fully connected CRF establishes pairwise potentials on all pairs of pixels in the image and has the ability to model long-range connections and capture fine edge details within the image. In contrast with local-range CRFs (Rother et al., 2004, Shotton et al., 2006), which are solved by an expensive discrete optimization problem (Kappes et al., 2015), mean field approximation inference for the fully-connected CRF is much more efficient (Krähenbühl and Koltun, 2011). In this paper, we propose to use fully connected CRF to model semantic segmentation of man-made scene problem and demonstrate it leads to state-of-the-art results.

The whole pipeline of our system consists of two parts, as shown

*Corresponding author

doi:10.5194/isprsarchives-XLI-B3-633-2016

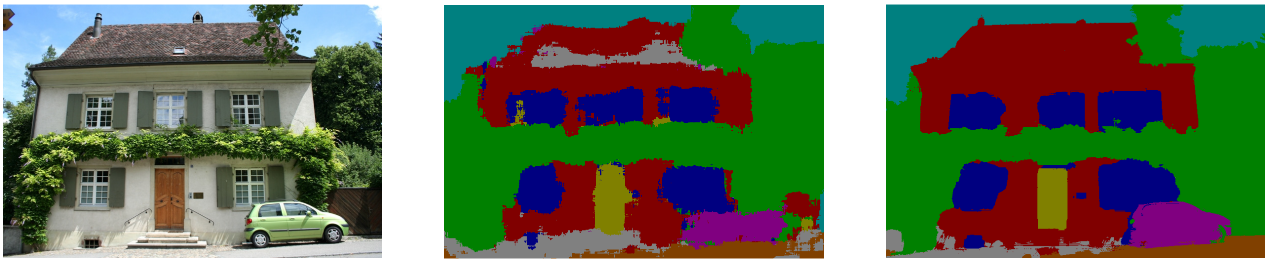

Figure 2. The pipeline of our method: first, we train Textonboost as the unary classifiers for each pixel; second, we run mean-field inference of fully connected CRF to get maximum posterior marginal results.

in Figure 2: first, we train the textonboost (Shotton et al., 2006) as the unary classifier for each pixel; second, we run mean-field inference (Krähenbühl and Koltun, 2011) for fully connected CRF to obtain maximum posterior marginal (MPM) results. Surprisingly, we find that the experimental results based fully connected CRF are better or more efficient than all previous approaches which are based on low connected CRFs on eTRIMS dataset (Korč and Förstner, 2009) and LabelMeFacade dataset (Fröhlich et al., 2010).

The following sections are organized as follows. The related works are discussed in Section 2 In Section 3 the method for the segmentation problem is described. In Section 4 experimental results are presented. Finally, this work is concluded and future work is discussed in Section 5

2. RELATED WORK

Man-made scene semantic segmentation approaches can be generally classified into two categories. One class methods are based on multi-class classifiers, e.g. randomized decision forest and boosting, for pixel or superpxiels, then use Conditional random fields or unsupervised segmentation methods to refine the classification results. this class method often is called as bottom-up method, such as (Yang and Förstner, 2011, Martinovic et al., 2012, Jampani et al., 2015).

Frohlich et al. (Fröhlich et al., 2010) presented a man-made scene image labeling method, which is using a random decision forest classifier and local features. Their method use an unsupervised segmentation, e.g. mean shift, to refine the classification results. In (Yang et al., 2010, Yang and Förstner, 2011), a hierarchical CRF model is proposed to solve man-made scene images semantic segmentation problem. In this hierarchical CRF model, multi-scale mean shift algorithm (Comaniciu and Meer, 2002) is used to segment the images into multi-scale superpixels. Unary potentials are the probability results of a randomized decision forest classifier, and then the spatial and hierarchical structures of the superpixels are connected as pairwise potentials. However, the superpixel-based hierarchical CRF model is limited by the accuracy of unsupervised image segmentation. Mistakes in the initial unsupervised segmentation cannot be recovered in the inference step, if superpixels cross multiple object classes.

Martinovic et al. (Martinovic et al., 2012) present a three-layered approach for semantic segmentation of building facades and man-made scenes. In the first layer, they train a recursive neural network (Socher et al., 2011) to get label probabilities of superpixels, which are got by oversegmenting the input image using mean shift algorithm (Comaniciu and Meer, 2002), as the unary potentials. In the second layer, using a grid CRF model to merge initial labeling and specialized object detectors (Dollár et al., 2009). In the third layer, weak architectural principles are used as a post-processing step. However, the accuracy of the three-layered method is also restricted by the precision of unsupervised image segmentation, similar as (Yang and Förstner, 2011). The system of (Jampani et al., 2015) uses a sequence of boosted decision trees, which are stacked using Auto-context features (Tu and Bai, 2010) and trained using the stacked generalization. They construct a CRF which is a pairwise 8-connected Potts model and unary classifiers are obtained directly from image, detection, and auto-context features. Their inference method is alpha expansion (Boykov et al., 2001), which costs about 24 seconds for an image on average. In contrast, the filter-based mean field approximation inference of fully-connected CRF only costs about 1 second per image of eTRIMS dataset. Therefore, fully connected CRF is much more efficient.

Another class of facade labeling method is shape grammar (Teboul et al., 2011, Martinovic and Gool, 2013), which is called as top-down approach, The shape grammar methods represent the facede using a parse tree and compute a grammar by a set of production rules. However these methods are not pixelwise labeling and not suitable for irregular man-made scene images, such The eTRIMS (Korč and Förstner, 2009) and LabelMeFacade (Fröhlich et al., 2010, Brust et al., 2015) datasets. Recently, convolutional patch network, which is one type of convolutional neural networks, is presented by (Brust et al., 2015). Since both the eTRIMS (Korč and Förstner, 2009) and LabelMeFacade (Fröhlich et al., 2010, Brust et al., 2015) image databases are relative small, this limit the classification and labeling ability of the convolutional patch networks.

3. METHOD

Conditional random field (CRF) (Lafferty et al., 2001) is a popular method for modeling the spatial structure of images in semantic segmentation problem. CRF model can combine the low-level pixel object classifiers information and spatial contextual information. Fully connected CRF (Krähenbühl and Koltun, 2011) establishes pairwise potentials on all pairs of pixels in the image and has the ability to model long-range connections, as shown in Figure 3, and capture fine edge details within the image. Each pairwise term of fully connected CRF is defined as a linear combination of Gaussian kernels. In contrast with local-range CRFs (Rother et al., 2004, Shotton et al., 2006), which are solved by an expensive discrete optimization problem (Kappes et al., 2015), mean field approximation inference for the fully-connected CRF is much more efficient (Krähenbühl and Koltun, 2011). The whole pipeline of our system consists of two parts, as shown in Figure 2: first, we train the textonboost (Shotton et al., 2006) as the unary classifier for each pixel independently; second, we run filter-based mean-field approximation inference (Krähenbühl and

Koltun, 2011) for fully connected CRF to obtain maximum posterior marginal (MPM) results.

3.1 Fully Connected CRF

We define a random field $\mathbf{X}$ over a set of variables $\{X_1, ..., X_N\}$ which is conditioned on pixels $\{I_1, ..., I_N\}$ of a man-made scene image $\mathbf{I}$. Each random variable X_j takes a label value from the label sets $\mathcal{L} = \{l_1, ..., l_L\}$, i.e. X_j is the label of pixel I_j. The conditional random field is defined as a Gibbs distribution

$$P(\mathbf{X} = \mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{x}|\mathbf{I})) \tag{1}$$

where $E(\mathbf{x})$ is the corresponding energy of the labeling $\mathbf{x} \in \mathcal{L}^N$ conditioned on $\mathbf{I}$. $Z(\mathbf{I})$ is the partition function. For convenience, we drop the conditioning on $\mathbf{I}$ in the notation. In the fully connected pairwise CRF model, the corresponding energy function is given by

$$E(\mathbf{x}) = \sum_i \phi_i(x_i) + \sum_{i<j\in N} \phi_{ij}(x_i, x_j) \tag{2}$$

where the unary potential $\phi_i(x_i)$ is the cost computed for pixel i taking the label x_i by a classifier given image features, and the pairwise energy potential $\phi_{ij}(x_i, x_j)$ encourage coherence in pixels x_i and x_j when they have similar features, such as, the RGB values and positions.

3.2 Unary Potentials

The image features used in our paper include 17-dimensional filter bank responses (Shotton et al., 2006), RGB color, Histogram of Oriented Gradient(HOG) (Dalal and Triggs, 2005), SIFT (Lowe, 2004) and pixel location information. Given these image features, we compute the unary potential $\phi_i(x_i)$ for each pixel i by a multiclass classifier that produces a probability distribution over the labeling x_i independently. The form of unary potential $\phi_i(x_i)$ is the negative log likelihood, i.e. corresponding probability distribution of the labeling assigned to pixel i.

$$\phi_i(x_i) = -\log P(x_i|\mathbf{I}) \tag{3}$$

The unary potentials incorporate shape, texture, location, and color descriptors, which are derived from TextonBoost (Shotton et al., 2006, Ladicky et al., 2009). We use the extended TextonBoost framework, which boosts classifiers defined on above mentioned features together. The implementation used here is the Automatic Labelling Environment (ALE) (Ladicky et al., 2009). The result of unary classifiers is usually noisy, as shown in the middle image of Figure 2.

3.3 Pairwise Potentials

The pairwise potentials in fully connected CRF model have the form

$$\phi_{ij}(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \tag{4}$$

where $k^{(m)}$ is a Gaussian kernel, $w^{(m)}$ is weight of the kernel, μ is label compatibility function, and $\mathbf{f}_i, \mathbf{f}_j$ are feature vectors for pixel i and j, which are color values and pixel positions as (Krähenbühl and Koltun, 2011).

In this paper, we use Potts model, $\mu(x_i, x_j) = \delta(x_i \neq x_j)$. For man-made scene semantic segmentation we use contrast-sensitive two-kernel potentials,

$$\begin{aligned} k^{(1)}(\mathbf{f}_i, \mathbf{f}_j) &= w^{(1)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}), \\ k^{(2)}(\mathbf{f}_i, \mathbf{f}_j) &= w^{(2)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}). \end{aligned} \tag{5}$$

where I_i and I_j are the color vectors and p_i and p_j are positions. The first part $k^{(1)}(\mathbf{f}_i, \mathbf{f}_j)$ is the smoothness kernel which help to remove small isolated regions, and the second part $k^{(2)}(\mathbf{f}_i, \mathbf{f}_j)$ is the appearance kernel which encourages nearby pixels to have the same labels when they have similar color.

3.4 Inference

Following (Krähenbühl and Koltun, 2011), we use a mean field method for approximate Maximum Posterior Marginal (MPM) inference. The mean field approximation computes an alternative distribution $Q(\mathbf{X})$ over the random variables $\mathbf{X}$, instead of computing the posterior distribution $P(\mathbf{X})$ directly. Distributions $Q(\mathbf{X})$ is a product of independent marginals, i.e. $Q(\mathbf{X}) = \prod_i Q_i(\mathbf{X}_i)$. The mean field approximation minimizes the KL-divergence $D(Q||P)$ between distribution Q and the exact distribution P. The mean field inference performs the following message passing iterative update until convergence:

$$\begin{aligned} Q_i(x_i = l) &= \frac{1}{Z_i} \exp\{-\phi_u(x_i) \\ &\quad - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(x_j = l') \phi_{ij}(x_i, x_j)\} \\ &= \frac{1}{Z_i} \exp\{-\phi_u(x_i) \\ &\quad - \sum_{l' \in \mathcal{L}} \sum_{m=1}^{K} \mu(l, l') w^{(m)} \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l')\} \end{aligned} \tag{6}$$

where Z_i is the marginal partition function of pixel i used to normalize the marginal. Updating the message passing iteration in sequence across pixels, KL-divergence will be convergence (Koller and Friedman, 2009). Directly computing this message passing iterative is intractable, because for each pixel, evaluating the sum of all of other pixels is required. This is the computational bottleneck of the message passing iterative. To make this update tractable and efficient, a high dimensional Gaussian filter can be used (Adams et al., 2010, Krähenbühl and Koltun, 2011). The transformation is:

$$\begin{aligned} \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l) &= \sum_j k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l) - Q_j(l) \\ &= [G_m \otimes Q(l)](\mathbf{f}_i) - Q_j(l) \end{aligned} \tag{7}$$

where G_m is the corresponding Gaussian kernel of $k^{(m)}$ and $\otimes$ is the convolution filtering. Using the permutohedral lattice (Adams et al., 2010), which is a highly efficient convolution data structure, the approximate message passing can be updated in time $O(Nd)$ (Krähenbühl and Koltun, 2011). Of cause, other filtering methods also can be used for the approximate message passing, e.g. domain transform filtering (Gastal and Oliveira, 2011, Vineet et al., 2014). The first smoothness kernel is a Gaussian blur. And the second appearance kernel actually is the joint or cross bilateral filtering (Tomasi and Manduchi, 1998, Petschnigg et al., 2004, Eisemann and Durand, 2004), in which $Q(l)$ is the input image and I is the reference (or guidance) image. After running the update step in a fixed number, in this paper we update 10 times

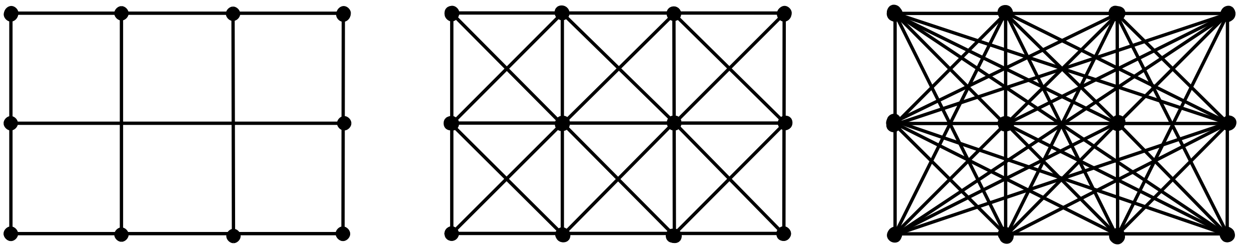

Figure 3. Fully connected CRF establishes pairwise potentials on all pairs of pixels in the image and has the ability to model long-range connections. From left to right: four-connected CRF, eight-connected CRF and fully-connected CRF.

iteration, then we get the MPM result from the final iteration,

$$x_i \in \arg\max_{l} Q_i(x_i = l). \tag{8}$$

3.5 Learning

For training unary potentials, we set the parameters of the low-level feature descriptors, such as, HOG, Texton and SIFT, using the setting of Automatic Labelling Environment (ALE). For the parameters of the CRFs, we use 5 folders cross-validation to learning the weights of the unary responses and Gaussian kernels.

4. EXPERIMENTS

We evaluate the fully-connected CRFs on all two irregular man-made scene images benchmark datasets: eTRIMS dataset (Korč and Förstner, 2009) and LabelMeFacade dataset (Fröhlich et al., 2010). For eTRIMS dataset, we perform a 5-fold cross-validation as in (Yang and Förstner, 2011) mentioned by dividing 40 images into a training set and 20 images into a test set randomly. For LabelMeFacade dataset, we use the pre-separated training and testing as the same as (Fröhlich et al., 2010, Brust et al., 2015) mentioned. We compare our results with against (Jampani et al., 2015) and (Brust et al., 2015).

4.1 Datasets

eTRIMS dataset (Korč and Förstner, 2009) includes 60 man-made scene images, which are labeled with 8 classes: *building, car, door, pavement, road, sky, vegetation and window.* And each image have an accurate pixel-wise annotation. For evaluation, we perform a 5-fold cross-validation as in (Yang and Förstner, 2011) by dividing 40 images into a training set and 20 images into a test set randomly. Then we run the experiment five times, and report the average accuracy.

LabelMeFacade Dataset is presented by (Fröhlich et al., 2010), which are also labeled with 8 classes: *building, car, door, pavement, road, sky, vegetation and window.* The images of LabelMeFacade dataset are taken from LabelMe dataset (Russell et al., 2008). There are 945 images in the dataset, which are split as two sets, 100 images for training set and 845 images testing set. Similar with eTRIMS dataset, facades in LabelMeFacade dataset are highly irregular.

4.2 Results

We compare our approach with the state-of-the-art man-made scene image segmentation method on eTRIMS dataset and LabelMeFacade dataset, i.e. facade segmentation using Auto-Context (Jampani et al., 2015) and Convolutional Patch Networks (Brust et al., 2015). We choose the average, overall and intersection over union (IoU) score as the evaluation measures. Overall is the pixel-wise labeling accuracy, which is computed over the whole image pixels for all classes. Average is the pixel-wise labeling accuracy computed for all classes and the averaged over these classes. IoU is defined as $TP/(TP+FP+FN)$. TP represents the true positive, FP means false positive and FN indicates the false negative.

We show the quantitative experimental results of eTRIMS dataset in Table 1. Our method outperforms the previous state-of-the-art approaches (Jampani et al., 2015) on eTRIMS dataset. We get the Average accuracy 0.8185 and IoU accuracy 0.6481. Our Average and IoU are highest, and we get five classes out of eight higher than the Auto-Context method (Jampani et al., 2015), which is the benchmark on the eTRIMS dataset before. Note that the Auto-Context method (Jampani et al., 2015) uses detection as pre-processing step. We do not using any detection information, and our approach is very efficient, which only need about one second in the inference step. Figure 4 shows some qualitative segmentation results of our method. Fully connected CRF obtains more accurate and detailed results than Textonboost and the low connected CRF. Our results demonstrate that dense pixel-level connectivity leads to significantly more accurate pixel-level classification performance (see window/vegetation class labeling results in most images).

We show the quantitative experimental results of LabelMeFacade dataset in Table 2. Our method outperforms the previous state-of-the-art approaches (Jampani et al., 2015) and (Brust et al., 2015) on LabelMeFacade dataset. Since (Brust et al., 2015) only provide Average and Overall result, we just compare (Brust et al., 2015) with these two measures. In contrast with (Jampani et al., 2015), they regard the 'various' as a class, we do not consider 'various' as a class. So LabelMeFacade dataset have eight classes as eTRIMS dataset. We get the Overall accuracy 0.7927 and IoU accuracy 0.4848. Our Average and IoU are highest, and we get five classes out of eight higher than the Convolutional Patch Networks method (Brust et al., 2015), which is the current benchmark on the eTRIMS dataset. Figure 5 shows some qualitative segmentation results of our method. Our method obtains more accurate and detailed results than the low connected CRF. Since building facades in LabelMeFacade dataset are highly irregular, there are sometimes wrongly labeled classes, e.g. parts of windows are labeled as building in some example images.

5. CONCLUSIONS

In this paper, we explored man-made scene semantic segmentation using fully connected CRF model, which is very efficient and only need about one second in the inference step. This simple method outperforms the previous state-of-the-art approaches on eTRIMS dataset and LabelMeFacade dataset, which obtains more accurate and detailed results. Inspired by the recent huge

Class	Textonboost	CRF	AC (Jampani et al., 2015)	Ours
Building	0.7438	0.8038	**0.925**	0.8414
Car	0.8158	0.8588	0.766	**0.8706**
Door	0.7740	0.8024	0.653	**0.7980**
Pavement	0.5902	0.6222	0.488	**0.6046**
Road	0.8002	0.8136	**0.821**	0.8116
Sky	0.9716	0.9842	0.989	**0.9932**
Vegetation	0.8970	0.9022	**0.929**	0.9174
Window	0.7766	0.7594	0.682	**0.7110**
Average	0.7962	0.8183	0.7814	**0.8185**
Overall	0.8012	0.8331	**0.8729**	0.8472
IoU	0.5944	0.6352	0.6354	**0.6481**

Table 1. The quantitative results on the eTRIMS dataset. Textonboost is trained using Automatic Labelling Environment. The CRF is a 4-connected CRF. AC is the Auto-context method. Our method is the fully connected CRF.

Class	Textonboost	CRF	AC (Jampani et al., 2015)	CPN (Brust et al., 2015)	CPNw (Brust et al., 2015)	Ours
Average	**0.6139**	0.6095	0.4904	0.4777	0.5898	0.5953
Overall	0.7547	0.7740	0.7523	0.7433	0.6341	**0.7927**
IoU	0.4685	0.4807	0.3957	-	-	**0.4848**

Table 2. The quantitative results on the LabelMeFacade dataset. Textonboost is trained using Automatic Labelling Environment. The CRF is a 4-connected CRF. AC is the Auto-context method. CPN is the Convolutional Patch Networks method. CPNw is the wighting version of Convolutional Patch Networks method. Our method is the fully connected CRF.

success of CNNs in computer vision and machine learning, e.g. image classification task (Krizhevsky et al., 2012) and semantic segmentation task (Long et al., 2015), we plan to explore this direction in our future research.

ACKNOWLEDGEMENTS

The work is partially funded by DFG (German Research Foundation) YA 351/2-1. The authors gratefully acknowledge the support.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. and Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34(11), pp. 2274–2282.

Adams, A., Baek, J. and Davis, M. A., 2010. Fast high-dimensional filtering using the permutohedral lattice. Comput. Graph. Forum 29(2), pp. 753–762.

Boykov, Y., Veksler, O. and Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), pp. 1222–1239.

Brust, C.-A., Sickert, S., Simon, M., Rodner, E. and Denzler, J., 2015. Convolutional patch networks with spatial prior for road detection and urban scene understanding. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp. 510–517.

Comaniciu, D. and Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24(5), pp. 603–619.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893.

Dollár, P., Tu, Z., Perona, P. and Belongie, S. J., 2009. Integral channel features. In: British Machine Vision Conference, pp. 1–11.

Eisemann, E. and Durand, F., 2004. Flash photography enhancement via intrinsic relighting. ACM Trans. Graph. 23(3), pp. 673–678.

Fröhlich, B., Rodner, E. and Denzler, J., 2010. A fast approach for pixelwise labeling of facade images. In: International Conference on Pattern Recognition, pp. 3029–3032.

Gastal, E. S. L. and Oliveira, M. M., 2011. Domain transform for edge-aware image and video processing. ACM Trans. Graph. 30(4), pp. 69:1–69:12.

Jampani, V., Gadde, R. and Gehler, P. V., 2015. Efficient facade segmentation using auto-context. In: IEEE Winter Conference on Applications of Computer Vision, pp. 1038–1045.

Kappes, J. H., Andres, B., Hamprecht, F. A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B. X., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B. and Rother, C., 2015. A comparative study of modern inference techniques for structured discrete energy minimization problems. International Journal of Computer Vision 115(2), pp. 155–184.

Kohli, P., Ladicky, L. and Torr, P. H. S., 2009. Robust higher order potentials for enforcing label consistency. International Journal of Computer Vision 82(3), pp. 302–324.

Koller, D. and Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Korč, F. and Förstner, W., 2009. eTRIMS Image Database for interpreting images of man-made scenes. Technical Report TR-IGG-P-2009-01.

Krähenbühl, P. and Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems, pp. 109–117.

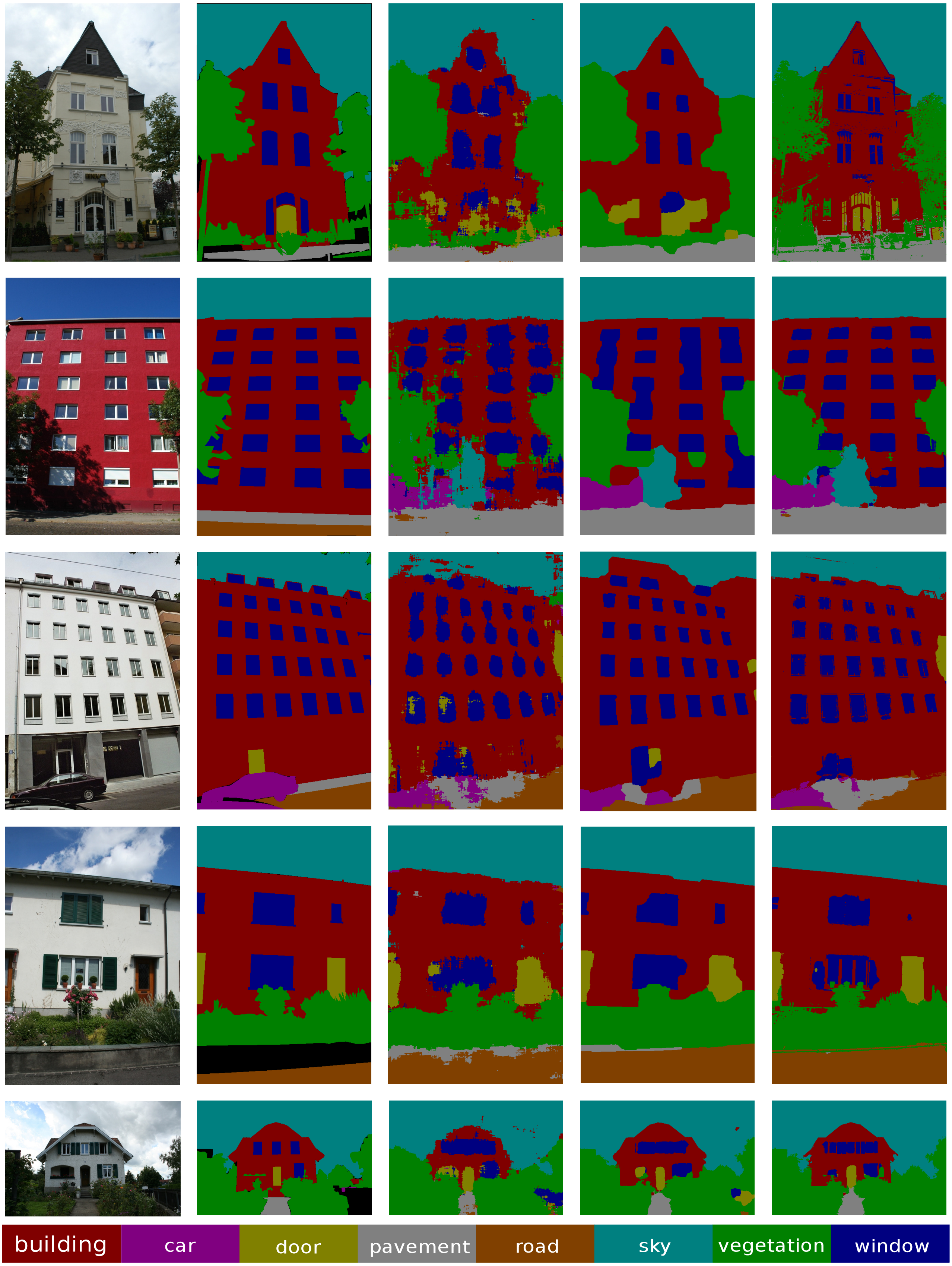

Figure 4. The qualitative results of the eTRIMS dataset. First column are examples of the testing images and 2nd-column are the corresponding ground truth. (3rd-column to 5th-column) man-made scene semantic segmentation results using the Textonboost classifier, the CRF model and the fully connected CRF model, respectively. The last row is the legend. The fully connected CRF model obtains more accurate and detailed results than the low connected CRF.

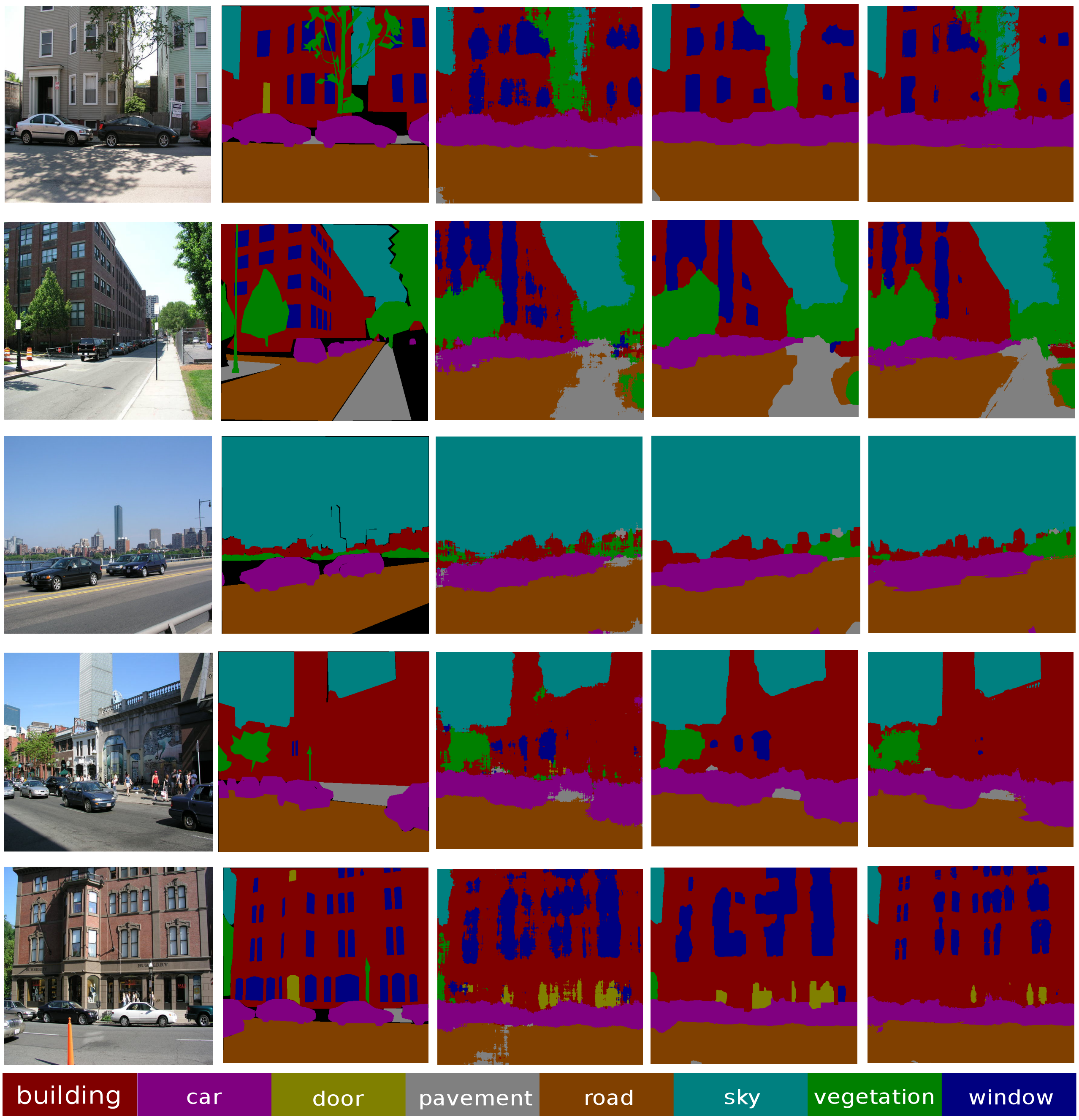

Figure 5. The qualitative results of the LabelMeFacade dataset. First column are examples of the testing images and 2nd-column are the corresponding ground truth. (3rd-column to 5th-column) man-made scene semantic segmentation results using the Textonboost classifier, the CRF model and the fully connected CRF model, respectively. The last row is the legend. The fully connected CRF model obtains more accurate and detailed results than the low connected CRF.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Ladicky, L., Russell, C., Kohli, P. and Torr, P. H. S., 2009. Associative hierarchical crfs for object class image segmentation. In: International Conference on Computer Vision, pp. 739–746.

Lafferty, J. D., McCallum, A. and Pereira, F. C. N., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, pp. 282–289.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), pp. 91–110.

Martinovic, A. and Gool, L. J. V., 2013. Bayesian grammar learning for inverse procedural modeling. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 201–208.

Martinovic, A., Mathias, M., Weissenberg, J. and Gool, L. J. V.,

2012. A three-layered approach to facade parsing. In: European Conference on Computer Vision, pp. 416–429.

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M. F., Hoppe, H. and Toyama, K., 2004. Digital photography with flash and no-flash image pairs. ACM Trans. Graph. 23(3), pp. 664–672.

Rother, C., Kolmogorov, V. and Blake, A., 2004. ”grabcut”: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23(3), pp. 309–314.

Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T., 2008. Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision 77(1-3), pp. 157–173.

Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), pp. 888–905.

Shotton, J., Winn, J. M., Rother, C. and Criminisi, A., 2006. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: European Conference on Computer Vision, pp. 1–15.

Socher, R., Lin, C. C., Ng, A. Y. and Manning, C. D., 2011. Parsing natural scenes and natural language with recursive neural networks. In: International Conference on Machine Learning, pp. 129–136.

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P. and Paragios, N., 2011. Shape grammar parsing via reinforcement learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2273–2280.

Tomasi, C. and Manduchi, R., 1998. Bilateral filtering for gray and color images. In: International Conference on Computer Vision, pp. 839–846.

Tu, Z. and Bai, X., 2010. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 32(10), pp. 1744–1757.

Vineet, V., Warrell, J. and Torr, P. H. S., 2014. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. International Journal of Computer Vision 110(3), pp. 290–307.

Yang, M. Y. and Förstner, W., 2011. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: ICCV Workshops, pp. 196–203.

Yang, M. Y., Förstner, W. and Drauschke, M., 2010. Hierarchical conditional random field for multi-class image classification. In: International Conference on Computer Vision Theory and Applications (VISSAPP), pp. 464–469.