SEMANTIC BUILDING FAÇADE SEGMENTATION FROM AIRBORNE OBLIQUE IMAGES

Y. Lin, F. Nex, M.Y. Yang

ITC Faculty of Geo-Information Science and Earth Observation, University of Twente, The Netherlands
(y.lin-1, f.nex, michael.yang)@utwente.nl

Commission II, **WG** II/4

KEYWORDS: Semantic Segmentation, Point Cloud, 3D features, Fully Connected Conditional Random Field

ABSTRACT:

With the introduction of airborne oblique camera systems and the improvement of photogrammetric techniques, high-resolution 2D and 3D data can be acquired in urban areas. This high-resolution data allows us to perform detailed investigations on building roofs and façades which can contribute to LoD3 city modeling. Normally, façade segmentation is achieved from terrestrial views. In this paper, we address the problem from aerial views by using high resolution oblique aerial images as the data source in urban areas. In addition to traditional image features, such as RGB and SIFT, normal vector and planarity are also extracted from dense matching point clouds. Then, these 3D geometrical features are projected back to 2D space to assist façade interpretation. Random forest is trained and applied to label façade pixels. Fully connected conditional random field (CRF), capturing long-range spatial interactions, is used as a post-processing to refine our classification results. Its pairwise potential is defined by a linear combination of Gaussian kernels and the CRF model is efficiently solved by mean field approximation. Experiments show that 3D features can significantly improve classification results. Also, fully connected CRF performs well in correcting noisy pixels.

1. INTRODUCTION

With population explosion in urban areas, detailed 3D city modeling is demanded for scientific urban planning, disaster management and tourism. Building façade interpretation is a subproblem, contributing to the Level of Detail 3 of CityGML. It aims to detect building façades and distinguish its components, like walls, windows, doors and balconies. However, time-inefficient human interpretation is the main hurdle in generating detailed 3D city models. Therefore, automated façade segmentation is required at urban scales.

Machine learning techniques are possible solutions to automated interpretation. Random forest (Frohlich et al., 2010), boosting scheme (Shotton et al., 2006) and deep convolutional neural networks (DCNNs) (Chen et al., 2015) are commonly used in scene interpretation, while all of them have few limitations. For traditional classifiers, like random forest and boosting scheme, results from pixel-wise classification are quite noisy because a single label is independently assigned to each pixel without considering labels of surrounding pixels. In contrast, DCNNs are capable to learn features from data and capture neighboring information by convolutional filters while the repetitive use of downsampling layers and max-pooling layers lead to large receptive fields. This often gives rise to coarse outputs and can consequently generate blob-like shapes and non-sharp boundaries (Chen et al., 2015). Conditional random field (CRF) models show their capabilities to take advantage of contextual information and deal with boundaries. 4-connected and 8-connected CRFs can capture short-range interactions between pixel labels and produce smoothed boundaries of façade elements. To improve classification accuracy and achieve better visualization on object boundaries, fully connected CRF is applied to refine outputs from classifiers (Chen et al., 2015). In fully connected CRF, both local and global spatial dependencies can be modeled. Globally connected structures model long-range interactions, to disambiguate object boundaries and figure out delicate structures.

Currently, images used for semantic façade segmentation are commonly from street views (CMP (Tyleček and Šára, 2013), ECP (Teboul, 2010), eTRIMS (Korč and Förstner, 2009)). However, when data are required to cover large areas for 3D city models, collecting data from terrestrial platforms is time-consuming. In contrast, acquiring data by airborne equipment is more efficient for wide coverage application. In high resolution airborne oblique images, building façades are visible. With the help of photogrammetric techniques, multi-views of ground objects make it possible to generate large coverage point clouds in urban areas. Comparing with those traditional datasets that only include a single view for each façade, this point cloud provides additional 3D geometrical cues to solve the problem. Although some previous studies employ multi-view façade images for façade segmentation (Gadde et al., 2017; Martinović et al., 2015), they only focus on terrestrial images and no semantic segmentation has been done from airborne images. This work aims to explore the potential of airborne images to address the problem.

The main objective of this work is to develop and apply a semantic classification method to differentiate different components of building façades from airborne oblique images. 2D and 3D features are integrated and CRF models are used to improve classification accuracy and achieve better visualization.

This paper is organized as follows: in Section 2, related works in façade interpretation are discussed. In Section 3, methods for feature extraction and segmentation are explained. In Section 4, model parameters and experimental results are shown and discussed. Conclusions and possible future work are described in Section 5.

2. RELATED WORK

In general, façade segmentation approaches can be divided into two categories. The first one is top-down method using shape

https://doi.org/10.5194/isprs-annals-IV-2-209-2018

grammar to parse façades. The second one is bottom-up method, applying the multi-class classifier to pixels or superpixels and then employing CRF or other optimization methods to refine classification results.

For top-down methods, a façade is represented by a tree and tree nodes keep splitting based on predefined rules and images characteristics. These rules or shape grammars are always manually defined counting on strong prior knowledge of façade structures. Teboul et al., (2010) define six rules to constrain the global layout of building façades. The splitting of façades considers pixel-wise classification results obtained from the random forest. However, their rules only fit Haussmannian style buildings in Paris and can fail when they are applied to other architectural styles. Instead of relying on prior knowledge, Martinović and Van Gool (2013) learn splitting grammars from labeled images while their method still focuses on grid-shape objects with good alignment and cannot deal with orientated façade objects from oblique airborne images.

Bottom-up methods aim to label façades at pixel or superpixel levels by using machine learning classifiers. Yang and Förstner (2011a) use random forest as the classifier for façade segmentation. Results are noisy due to the lack of contextual information. Rahmani et al. (2017) propose an approach using a structured random forest to produce nearly noise-free façade segmentation. Schmitz and Mayer (2016) use fully convolution network to achieve façade interpretation. As building façade components always possess symmetry in shape, Liu et al. (2017) present an approach to incorporate this symmetry in loss function when training neural networks.

The conditional random field is commonly used to refine pixel-wise classification results by modeling contextual interactions. Yang and Förstner (2011b) propose a hierarchical CRFs to exploit contextual dependencies from local to global for façade interpretation using mean shift superpixels at different levels. In addition to a unary term from random forest and a pairwise term that represents class compatibility between nearby labels, a hierarchical term is added to demonstrate segment relationships among different scales. Li and Yang (2016) implement fully connected CRF to semantic façade segmentation. They choose Textonboost to get unary potentials and pick linear combinations of Gaussian kernels (Krähenbühl and Koltun, 2011) as fully connected pairwise potentials. Their model is not only good at enforcing the label consistency but also capable of detecting small façade components and delineating crisp boundaries. Martinović et al. (2012) propose a three-layered approach to solve façade interpretation. In the first layer, a Recursive Neural Network is trained to get label probability at superpixel level. In the middle layer, object detectors are used to obtain probabilities of window and door over the image. This information is combined with the output of RNN in a CRF model. In the top layer, weak architectural constraints are introduced to achieve more structured façade configurations.

All above studies are attempts to interpret façades by using image features, while urban scene interpretation also benefits from involving 3D data. Vetrivel et al. (2017) use dense matching point clouds to facilitate building damage detection in aerial oblique images. 3D characteristics are extracted in 3D space, like linearity, planarity and scattering. The combination of two types of features achieves 3% higher average classification accuracy than the approach only using 2D features (Vetrivel et al., 2017). Fooladgar and Kasaei (2015) also combine 2D and 3D information at image pixel level to achieve semantic segmentation of indoor RGB-D images. A CRF model is proposed where unary potentials are from random forest and Potts model is set to calculate pairwise potentials. Instead of assigning labels to image pixels, Martinović et al. (2015) design a 3D pipeline to take advantages of 2D images and 3D point clouds from Structure from Motion for 3D labeling. Height, depth, normal vector and spin image descriptors at different scales are 3D features used in a random forest classifier. A 3D CRF, considering 4 nearest neighboring points, is used as a post-processing to smooth results. They find the CRF model that utilizes both 2D and 3D features in 3D space and incorporates with superpixel and object detectors achieves the best accuracy.

Currently, most of the studies only use single view images for façade segmentation and few of them incorporate 3D features obtained from multi-view images. Some studies use aerial images for scene understanding but studies in semantic segmentation at façade level are quite rare. This work explores potentials of airborne images to address the problem with a fully connected CRF.

3. METHOD

Figure 1 is the workflow for this paper. Firstly, façades are annotated in 2D images and cropped from 3D point clouds and the dataset is split into three parts. Then 2D and 3D features are extracted and combined. These features are fed into a fully connected CRF model. 45 façades are used to train random forest, 15 façades are used to tune CRF parameters and 45 façades are used to test models. More details are explained in the following sections.

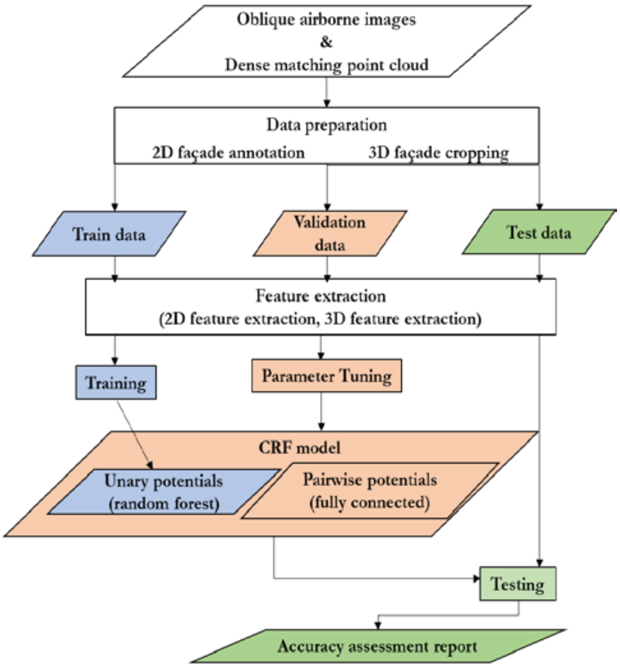

Figure 1 Workflow for this work.

3.1 Feature extraction

Here, 2D features for every façade are extracted from images and 3D features are extracted from crop 3D point clouds. Then, 3D features for points are projected back to 2D space and combined with image features.

3.1.1 2D feature extraction

Three different typologies of 2D features have been adopted:

Color features In this paper, color information is stored in RGB color space. Intensities in red, green and blue channels are used as three features to represent spectral information.

SIFT SIFT descriptor is made up of 128 features. These features are extracted from grayscale images in a grid region at a fixed scale and fixed orientations (Liu et al., 2011). More detailed explanation refers to Liu et al. (2011) and Lowe (2004).

LM filter 48 texture features are derived from Leung-Malik filter bank (Varma and Zisserman, 2005). The filter bank[1] is a combination of 18 first order and 18 second order derivatives of Gaussian kernels, 8 Laplacian of Gaussian kernels and 4 Gaussian kernels.

3.1.2 3D feature extraction

Normal vector and planarity are features extracted from local neighbors. Normal vector is helpful to separate points on different planes, like points on roof and wall surfaces. Planarity is an efficient indicator to assess whether the surface is flat or curved and distinguish objects with different kinds of surfaces (Vosselman et al., 2017). Planarity is derived from three eigenvalues ($\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$) of the covariance matrix which is calculated based on local neighbors. In equation (1), three eigenvalues are normalized by $\lambda_1 + \lambda_2 + \lambda_3$ and therefore $e_1+e_2+e_3=1$.

$$Planarity = \frac{e_2 - e_3}{e_1} \tag{1}$$

As both normal vector and planarity are computed based on local neighbors, how to define neighboring points is critical in this work. Here 'k-nearest neighbors' is the method to search for nearby points (Weinmann et al., 2015). Figure 2 shows how unsupervised classification results are influenced by scales of the search range. Planarity is extracted from 20, 100 and 500 nearest points respectively. Although these 3D features extracted from a single scale are not sufficient to distinguish objects, the change of the planarity in different scales is a signature for different classes, especially for those objects on plane surfaces (Brodu and Lague, 2012). Balcony horizontal surfaces on wall and chimney vertical surfaces on roofs can be well detected in figure 2 d. Thus, instead of extracting normal vector and planarity at a fixed scale (4 features), we extract features from 20, 100 and 500 nearest points respectively (12 features in total) and then fed them into a classifier.

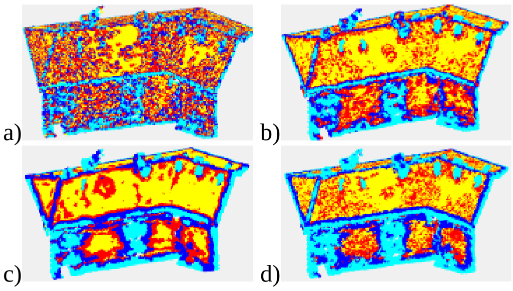

Figure 2. Unsupervised segmentation based on 3D features extracted from different scales (ground truth refers to figure 3). Planarity is calculated based on different sizes of neighboring points. a): 20, b):100, c):500, d): {20, 100, 500}.

3.1.3 Feature combination

To combine 2D and 3D features, 3D features are projected back to oblique images using PMatrix generated by Pix4D mapper for every oblique image. Projected 2D coordinates can be related to different patch sizes in 2D images. Pixels within a patch share same 3D features. If multiple points fall in a same patch, averaged 3D features are assigned to that patch (Johnson & Hebert, 1999). With increasing patch size, although the percentage of void pixels keeps decreasing, projected 3D features are coarser and loss more detailed information as an effect of averaging. As we did not use the full resolution point cloud in Pix4D, to avoid holes in projected 3D features, 4*4 pixel is defined as an optimal patch size when projecting 3D points to images.

3.2 Random forest

A random forest classifier is an ensemble of independent decision trees and classification results are votes from those trees. Every decision tree is a function of ***a*** to get ***b*** where ***a*** is a sample consisting of n features and recursively classified by branching down the tree until the sample reaches a leaf node. ***b*** is a probability distribution for each class assigned by the leaf node based on feature values in sample ***a***. For each node, a split function is used to decide whether the sample should go left or right. Splitting terminates until a leaf node is reached. For each node, the split function is learned from training dataset.

3.3 Conditional random field

Conditional random fields are commonly used to refine noisy segmentation results. They combine results from simple classifiers with contextual information. In 4-connected and 8-connected CRFs, neighboring systems are established based on 4 and 8 nearest points which can only capture short-range spatial interactions. In fully connected CRF, pairwise potentials are built on all possible pairs of pixels over the whole image and this full connection makes it possible to model long-range interactions within an image. Inference of fully connected CRF by traditional algorithms is computationally expensive. Krähenbühl and Koltun (2011) apply a linear combination of Gaussian kernels as the pairwise term and use mean field approximation to efficiently solve fully connected CRF.

In this paper, a random field $\boldsymbol{X}$ is constructed by a set of random variables $\{x_1, \dots, x_N\}$, where N is the number of pixels over the image. For each random variable in $\boldsymbol{X}$, its domain is a set of labels $\mathcal{L} = \{l_1, \dots, l_k\}$, where k is the number of classes. Random field $\boldsymbol{X}$ is conditioned on image $\boldsymbol{I}$ which consists of image pixels $\{I_1, \dots, I_N\}$. This conditional random field $(\boldsymbol{I}, \boldsymbol{X})$ is a Gibbs distribution and written as:

$$P(\boldsymbol{X}|\boldsymbol{I}) = \frac{1}{Z(I)} exp(-\sum\nolimits_{c \in \boldsymbol{C}_{\mathcal{G}}} \phi_c(\boldsymbol{x_c}|\boldsymbol{I})) \tag{2}$$

$$E(\boldsymbol{x}) = \sum\nolimits_{c \in \boldsymbol{C}_{\mathcal{G}}} \psi_c(\boldsymbol{x_c}) \tag{3}$$

$$Z(I) = \sum\nolimits_{\boldsymbol{x}} \exp(-E(\boldsymbol{x})) \tag{4}$$

Here $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph established on $\boldsymbol{X}$. Potential functions $\phi_c(\boldsymbol{x}_c|\boldsymbol{I})$ are defined over variables ($\boldsymbol{x}_c = \{x_i\ ,\ i \in c\}$) within a clique c. $\boldsymbol{C}_{\mathcal{G}}$ is a set of all cliques in a graph. $E(\boldsymbol{x})$ is a Gibbs energy function to label $\boldsymbol{x} \in \mathcal{L}^N$.

[1] This filter bank is available at: http://www.robots.ox.ac.uk/~vgg/research/texclass/filters.html

$\phi_c(\boldsymbol{x}_c|\boldsymbol{I})$ is simplified as $\psi_c(\boldsymbol{x}_c)$. $Z(I)$ is a partition function which is a normalization constant. The maximum a posteriori (MAP) labeling $\boldsymbol{x}^*$ is defined as below:

$$\boldsymbol{x}^* = argmax_{x \in \mathcal{L}^N} P(\boldsymbol{X}|\boldsymbol{I}) \tag{5}$$

Optimal labeling can be found by minimizing the energy function $E(\boldsymbol{x})$.

In this paper, fully connected CRF is a pairwise model. Therefore, the energy function can be written as below:

$$E(\boldsymbol{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{6}$$

where unary potential $\psi_u(x_i)$ is derived from probability distribution over labels from classifiers. Pairwise potentials $\psi_p(x_i, x_j)$ enforce consistency in pixels that share similar features in image space.

3.3.1 Unary Potentials

Feature extraction is explained in section 3.1. Taking those 2D and 3D features, random forest mentioned in section 3.2 gives multi-class label prediction. For every pixel i, probability distribution over label set $\mathcal{L}$, $P(x_i|\boldsymbol{I})$, is independently generated by classifiers. Unary potentials for pixel i are defined as the negative log of probability, shown as below:

$$\psi_u(x_i) = -logP(x_i|\boldsymbol{I}) \tag{7}$$

3.3.2 Pairwise Potentials

The pairwise term in fully connected CRF is formed by a linear combination of Gaussian kernels (Krähenbühl and Koltun, 2011), defined as below:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_m^1 w^{(m)} k^{(m)} (\boldsymbol{f}_i, \boldsymbol{f}_j) \tag{8}$$

$$\mu(x_i, x_j) = \{\begin{matrix} 0 & if\ x_i = x_{j,} \\ 1 & otherwise, \end{matrix} \tag{9}$$

$$w^{(1)} k^{(1)}(\boldsymbol{f}_i, \boldsymbol{f}_j) = w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \tag{10}$$

$$w^{(2)} k^{(2)}(\boldsymbol{f}_i, \boldsymbol{f}_j) = w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \tag{11}$$

$k^{(m)}$ are Gaussian kernels and $w^{(m)}$ are weights of kernels. $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ are feature vectors taking color values and positions for neighboring pixels i and j. Here, $\mu(x_i, x_j)$, a Potts model, is applied as a label compatibility function. It assigns penalties when two neighboring pixels have different labels. Two kernel potentials, $k^1(\boldsymbol{f}_i, \boldsymbol{f}_j)$ and $k^2(\boldsymbol{f}_i, \boldsymbol{f}_j)$, make the Potts model contrast-sensitive. p_i, p_j are position vectors and I_i, I_j are color vectors using RGB values. $k^1(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is an appearance kernel encouraging two pixels, which are close in position and have similar colors, to have a same label. In other words, a high penalty will be introduced when two pixels with different labels have similar features vectors and this penalty should be minimized to achieve coherency between pixels. θ_α and θ_β are used to control the extents of nearness and color similarity when determining penalties. $k^2(\boldsymbol{f}_i, \boldsymbol{f}_j)$ is a smoothness kernel to clean small and isolated parts.

3.3.3 Inference

According to Krähenbühl and Koltun (2011), an approximate CRF distribution is applied for Maximum Posterior Marginal labeling. This alternative distribution Q(X) is obtained based on the mean field approximation to an exact distribution P(X). Q(X) which is a product of independent marginals, can be computed by minimizing the KL-divergence D(Q||P), the difference between P(X) distribution and Q(X) distribution. Following the update equation (equation 12), the inference is calculated by iterative message passing, compatibility transform and local update within the approximate field until convergence.

$$\begin{aligned} &Q_i(x_i = l) \\ &= \frac{1}{Z_i} exp \left\{ -\psi_u(x_i) \right. \\ &\left. - \sum_{l' \epsilon \mathcal{L}} \mu(l, l') \sum_{m=1}^{K} w^{(m)} \sum_{j \neq i} k^{(m)}(\boldsymbol{f}_i, \boldsymbol{f}_j) Q_j(l') \right\} \end{aligned} \tag{12}$$

In fully connected CRF, direct message passing computation is intractable because, for every pixel, the sum of all other pixels is supposed to be evaluated (Krähenbühl and Koltun, 2011). The complexity of this problem is quadratic to the number of pixels in images. Thus, high dimensional Gaussian filtering is applied to reduce the complexity from quadratic to linear. The transformed message passing algorithm is shown as below:

$$\begin{aligned} &\sum_{j \in \mathcal{V}} k^{(m)} (\boldsymbol{f}_i, \boldsymbol{f}_j) Q_j(l) - Q_i(l) \\ &\qquad = [G_{\Lambda^{(m)}} \otimes Q(l)](\boldsymbol{f}_i) - Q_i(l) \end{aligned} \tag{13}$$

Message passing algorithm expressed as $\sum_{j \in \mathcal{V}} k^{(m)} (\boldsymbol{f}_i, \boldsymbol{f}_j) Q_j(l) - Q_i(l)$, is converted into a function where all pixels are summed up by convolutions $G_{\Lambda^{(m)}}$ but sum of Q_i is excluded (Krähenbühl and Koltun, 2011). Here, $G_{\Lambda^{(m)}}$ is a low passing filter. Based on sampling theorem, the function can be reconstructed and the convolution can be achieved by downsampling $Q(l)$, applying $G_{\Lambda^{(m)}}$ to sampled $Q(l)$ and then upsampling results in feature space (Krähenbühl and Koltun, 2011). Gaussian kernels $G_{\Lambda^{(m)}}$ can be approximately converted to truncated Gaussians, where values are turned to zero if they are not within two standard deviations. As sample spacing is in proportion to standard deviations, there are fixed number of samples in truncated kernels. Therefore, by summing over limited number of nearby pixels, convolution can be approximately calculated. This suggests that the inference can be completed in O(N) time. Permutohedral lattice, an efficient convolution data structure, is applied to simplify the calculation to be O(Nd) time. By Cholesky decomposition, high dimension kernels are separated into 1 dimensional kernels that allows the inference tractable (Krähenbühl and Koltun, 2011).

3.3.4 Learning

Features mentioned in section 3.1 are extracted from 45 façades and used to train random forest. 15 façades are used as validation dataset to tune parameters in fully connected CRF.

https://doi.org/10.5194/isprs-annals-IV-2-209-2018 | © Authors 2018. CC BY 4.0 License.

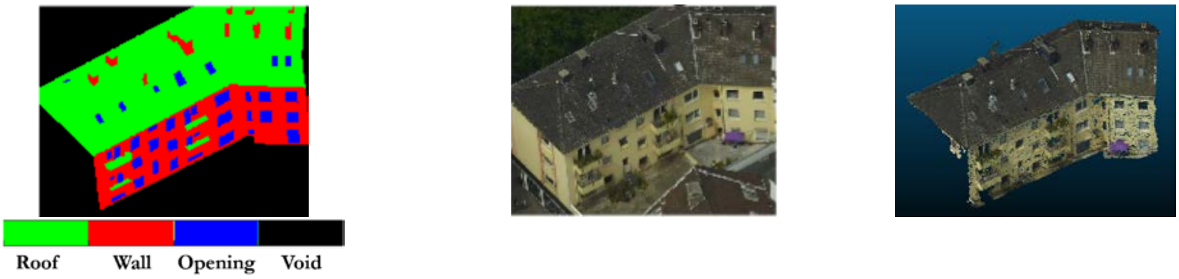

Figure 3 Example of our dataset. Left: annotated façade, Middle: façade image, and Right: cropped façade point cloud.

4. EXPERIMENTS

Airborne images used in this paper were acquired by an IGI Pentacam system over the city of Dortmund (Germany) on July 7th, 2016. Average ground sampling distance for oblique images is 4.5 cm. High-resolution images have been oriented and used to generate the point cloud in Pix4D.

In this paper, 3 classes of building components are defined regarding functionality, namely, roof, wall and opening. Roof is defined as a structure horizontally covering a building and wall is an element vertically covering a building. In this scenario, balconies on façades are divided into roof segments and wall segments (figure 3 Left)) and chimneys are also separated into two parts (figure 3 Left)). Opening includes windows and doors because both structures allow air, sound and light to pass. Also, in urban areas, especially for commercial buildings, doors are made of glass, the same material as windows. Online annotation tool LabelMe is used to delineate component boundaries on building façades. Façades of interest are manually cropped from dense matching point clouds.

4.1 Model parameters

4.1.1 Random forest

For random forest, a larger number of decision trees can achieve better results but it takes longer time to train those trees. To achieve the balance between time and accuracy, 50 trees are used in our work. Minimum leaf size is the minimum number of observations in each leaf. If it is small, branches are likely to go deep. Although out of bag prediction error is small in this case, the forest can be overfitting and have poor performance on testing images. Thus, minimum leaf size is set to be 50. This creates shallow trees but avoids overfitting. Number of predictors to sample defines how many features are selected at random to feed to each node. If it is too large, the strength of an individual tree increases and the correlation between different trees increases. As reliable performance of a random forest counts on the independence between individual decision trees, the high correlation is not allowed (Breiman, 2001). As a result, the square root of the total number of features is calculated to be the value of the number of predictors to sample (in this case: 14).

4.1.2 Fully connected CRF

Parameter setting tuned by 15 validation façades are shown as below:

$$w^{(1)} = 1,\ \theta_\alpha = 4,\ \theta_\beta = 11,\ w^{(2)} = 2,\ \ \theta_\gamma = 1$$

In our case, the optimal spatial standard deviation θ_α is 4 pixels and the optimal value for color standard deviation θ_β is 11. The influence of θ_α and θ_β on overall pixel accuracy are assessed qualitatively (figure 4) and quantitatively (figure 5). For this assessment, $w^{(1)}$ is kept as 1 and $w^{(2)}$ is set to be 0. As θ_α increases, the accuracy increases at first and then steadily decreases. Long-range connections cause some failures (figure 5). In contrast to what is mentioned in Krähenbühl and Koltun's experiment (2011), most of spatial standard deviations larger than 35 pixels, relatively short-range connections are more suitable for façade interpretation from aerial oblique images.

4.2 Accuracy assessment

Performances of random forest and the fully connected CRF model are evaluated by annotated testing façades and classification results are estimated by 3 measures. Overall pixel classification accuracy for entire images and averaged pixel-wise accuracy for each class are two standard measures. Intersection over union (IoU) score (Everingham et al., 2010) is calculated for each class and then averaged. These three measures are computed in terms of true positives (TP), false positives (FP) and false negatives (FN). Followings are equations:

Overall accuracy: ‘$TP/(TP + FN)$’ is calculated over the whole image.
Average accuracy: ‘$TP/(TP + FN)$ is calculated for every class and then averaged.

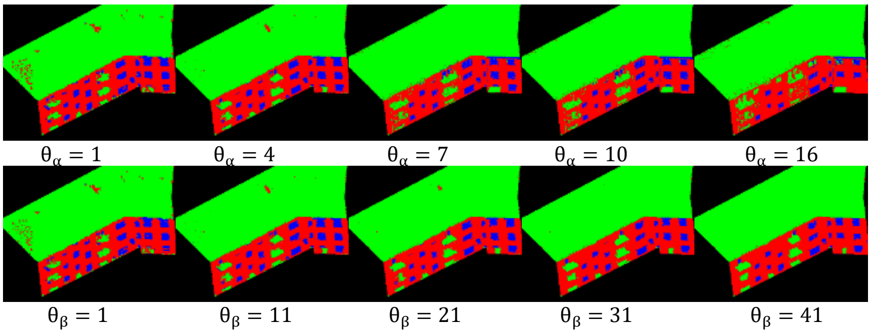

Figure 4 Qualitative assessment of the influence of connections in fully connected CRF (ground truth refers to figure 3).

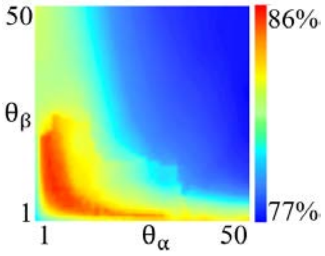

Figure 5 Quantitative assessment of the influence of connections in fully connected CRF.

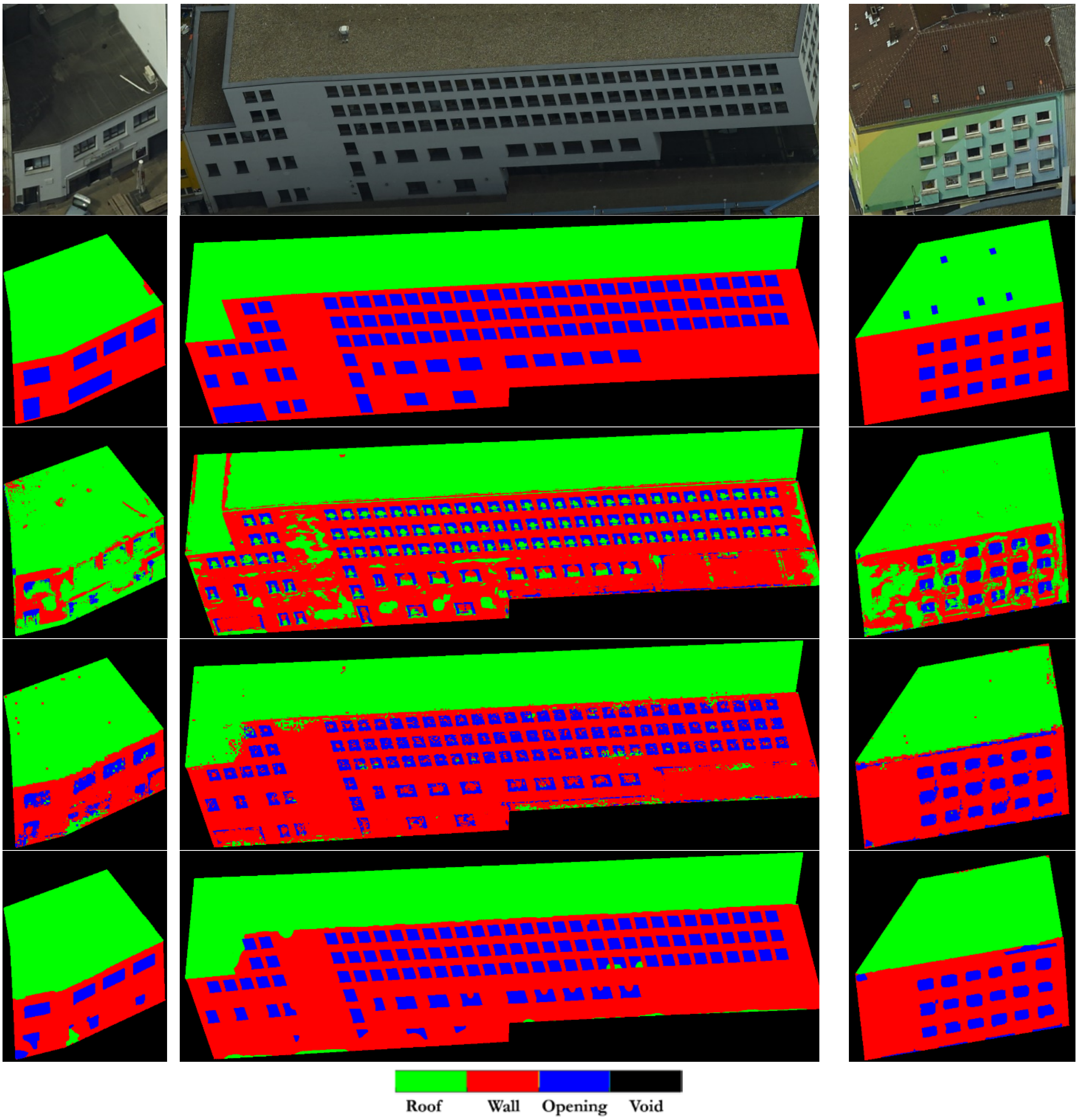

Figure 6 Examples from our dataset. First row: cropped façade images from oblique aerial images. Second row: ground truth. Third row: results from random forest using 2D features. Fourth row: results from random forest using 2D and 3D features. Fifth row: results from fully connected CRF and the unary term is obtained from random forest using 2D and 3D features.

IoU score: ‘$TP/(TP + FN + FP)$ is calculated for every class and then averaged.

4.3 Results and discussion

We used 45 façades to test models. Statistical results got from three models are demonstrated in table 1. Performance of the 2D feature classifier is worst in terms of all accuracy indicators. Although many studies show that color, SIFT and Texton features are capable to describe and distinguish different classes in terrestrial view dataset, like eTRIMS (Li and Yang, 2016), results from our work indicate that these 2D features are insufficient to distinguish roof, wall and opening in aerial view dataset with various architectural styles. It can be seen from table 2 that, based on 2D features, wall and opening pixels are likely to be misclassified as roof pixels.

Involvement of 3D features significantly corrects confusions between roof and pixels on vertical surfaces like wall and opening (table 3) and leads to an increase in IoU by 26.36% (table 1). This is because normal vector can efficiently separate roof pixels from wall and opening pixels. Gadde et al. (2017) also combine 2D and 3D features for image labeling in a terrestrial view dataset and there is an increase from 60.5% to 62.7% in IoU by adding 3D geometrical features to 2D features. Compared with results from Gadde et al. (2017), aiming to delineate detailed façade objects (window, wall, balcony, door, roof and shop), our experiment suggests that 3D features play an essential role in façade interpretation (roof, wall, opening) from aerial oblique images. However, there are still confusions between wall and opening pixels to be solved (table 3).

In our experiment, fully connected CRF refines results from random forest using both 2D and 3D features. It improves the IoU by 4.67% (table 1) and reduces noise to achieve better visualization (figure 6).

Our work suggests that 3D features are critical in façade interpretation from aerial oblique images. Inaccurate point clouds produced by poor image dense matching cannot solve misclassification. Figure 7 gives an example where classifier using 2D features misclassifies wall pixels as roof pixels (figure 7 d). In figure 7 e, 3D features can correct most of the wrongly labeled wall pixels, while there are still few roof pixels on the wall. By checking the corresponding point cloud in figure 7 b, few wall points have similar normal vectors to roof points. This unsolved misclassification in random forest cannot be corrected by adding fully connected pairwise potentials.

Class	2DRF	2D3DRF	FCRF
Roof	91.66%	93.30%	96.11%
Wall	39.75%	81.56%	85.56%
Opening	35.78%	57.89%	60.20%
Average	55.73%	77.59%	80.62%
Overall	60.59%	82.42%	85.63%
IoU	39.64%	66.00%	70.67%

Table 1 Quantitative results got from 3 models (45 façades for testing). 2DRF represents the random forest trained by 2D features. 2D3DRF represents the random forest trained by both 2D and 3D features. FCRF is a fully connected CRF using outputs from 2D3DRF as the unary term.

Predict True	Roof	Wall	Opening
Roof	**91.66%**	6.60%	1.74%
Wall	57.13%	**39.75%**	3.11%
Opening	50.58%	13.64%	**35.78%**

Table 2 Pixelwise accuracy of random forest using only 2D features.

Predict True	Roof	Wall	Opening
Roof	**93.30%**	6.05%	0.65%
Wall	11.22%	**81.56%**	7.22%
Opening	8.32%	33.79%	**57.89%**

Table 3 Pixelwise accuracy of random forest using both 2D and 3D features.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we explored the semantic façade segmentation from airborne oblique images. It is an alternative investigation on state-of-the-art works which perform façade segmentation from terrestrial views. Different from traditional semantic image segmentation, 3D geometrical features were extracted from dense matching point clouds and then projected back to 2D space to facilitate image façade segmentation based on images features. Instead of extracting 3D features at a fixed scale, for each point, normal vector and planarity were extracted at different scales, taking as signatures for different classes. These 3D features contributed to a significant increase in IoU from 39.64% to 66.00%. Although the accuracy of semantic segmentation strongly relied on the quality of point clouds and errors in point clouds led to misclassification, in most of the cases, current photogrammetric algorithm allowed to generate reliable points and some confusions were solved by using CRF model. Fully connected CRF, a state-of-art model, was used to consider contextual information. It improved the IoU by 4.67%. The main limitation of this study was that we only implemented a three-class classification (roof, wall, opening). In the future, more classes could be identified, making contributions to more detailed 3D city modeling. Also, inspired by the huge application of deep learning in semantic image segmentation (Chen et al., 2015), CNN frameworks will be explored to perform semantic façade segmentation in our future research.

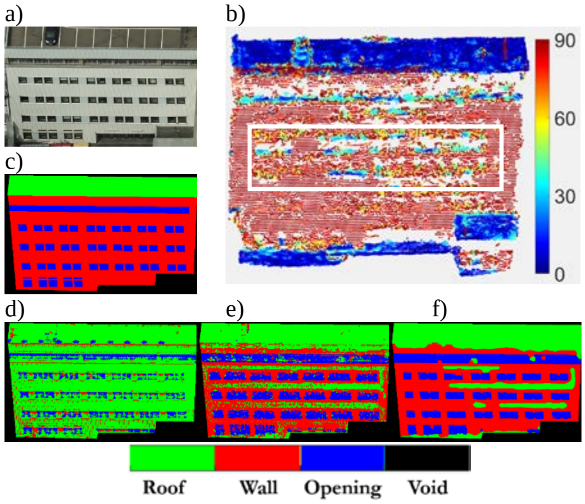

Figure 7 Misclassification caused by inaccurate dense matching point cloud. a) façade image, b) façade point cloud, color representing the angle (°) between normal vector and z-axis, c) ground truth, d) results from random forest using 2D features, e) results from random forest using 2D and 3D features, f) results from fully connected CRF.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to IGI mbH and Aerowest GmbH companies for providing the aerial image dataset over Dortmund city center.

REFERENCES

Breiman, L., 2001. Random Forests. *Machine Learning*, *45*(1),

Brodu, N., and Lague, D., 2012. 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology. *ISPRS Journal of Photogrammetry and Remote Sensing*, *68*, 121–134.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. , 2015. DeepLab: Semantic Image

Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv Preprint arXiv:1606.00915*.

Everingham, M., Luc, Gool, V., Williams, C. K. I., Winn, J., Zisserman, A., 2010. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision Manuscript*, *88*(2), 303–338.

Fooladgar, F., and Kasaei, S., 2015. Semantic Segmentation of RGB-D Images Using 3D and Local Neighbouring Features. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1–7). IEEE.

Frohlich, 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).

Gadde, R., Jampani, V., Marlet, R., and Gehler, P. V., 2017. Efficient 2D and 3D Facade Segmentation using Auto-Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Johnson, A. E., and Hebert, M., 1999. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(5), 433–449.

Korč, F., and Förstner, W., 2009. eTRIMS Image Database for Interpreting Images of Man-Made Scenes. *Technical Report TR-IGG-P-2009-01, University of Bonn, Dept. of Photogrammetry*.

Krähenbühl, P., and Koltun, V., 2011. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *Advances in Neural Information Processing Systems*, 109–117.

Li, W., and Yang, M. Y., 2016. Efficient Semantic Segmentation of Man-Made Scenes Using Fully-Connected Conditional Random Field. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *41*.

Liu, C., Yuen, J., and Torralba, A., 2011. SIFT Flow: Dense Correspondence across Scenes and Its Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 978–994.

Liu, H., Zhang, J., Zhu, J., and Hoi, S. C. H., 2017. DeepFacade: A Deep Learning Approach to Facade Parsing. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Martinović, A., Knopp, J., Riemenschneider, H., and Van Gool, L., 2015. 3D All The Way: Semantic Segmentation of Urban Scenes From Start to End in 3D. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4456–4465).

Martinović, A., Mathias, M., Weissenberg, J., and Van Gool, L., 2012. A Three-Layered Approach to Facade Parsing. *Computer Vision–ECCV 2012*, 416–429.

Martinović, A., and Van Gool, L., 2013. Bayesian Grammar Learning for Inverse Procedural Modeling. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 201–208). IEEE.

Rahmani, K., Huang, H., and Mayer, H., 2017. Facade Segmentation with a Structured Random Forest. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*.

Schmitz, M., and Mayer, H., 2016. A Convolutional Network for Semantic Facade Segmentation and Interpretation. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *41*, 709–715.

Shotton, J., Winn, J., 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). Springer.

Teboul, O., 2010. Ecole centrale paris facades database.

Teboul, O., Simon, L., Koutsourakis, P., and Paragios, N., 2010. Segmentation of Building Facades Using Procedural Shape Priors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3105–3112).

Tyleček, R., and Šára, R., 2013. Spatial Pattern Templates for Recognition of Objects with Regular Structure. In *German Conference on Pattern Recognition* (pp. 364–374).

Varma, M., and Zisserman, A., 2005. A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision*, *62*(1/2), 61–81.

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., and Vosselman, G., 2017. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Vosselman, G., Coenen, M., and Rottensteiner, F., 2017. Contextual segment-based classification of airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *128*, 354–371.

Weinmann, M., Jutzi, B., Hinz, S., and Mallet, C., 2015. Semantic Point Cloud Interpretation Based on Optimal Neighborhoods, Relevant Features and Efficient Classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, *105*, 286–304.

Yang, M. Y., and Förstner, W., 2011a. Regionwise classification of building facade images. *Photogrammetric Image Analysis*, Vol. 6952 LNCS, pp. 209–220.

Yang, M. Y., and Förstner, W., 2011b. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. IEEE *International Conference on Computer Vision Workshop*, pp. 196–203.