Semantic Segmentation of Objects from Airborne Imagery

Thuy Thi Nguyen[1,3], Sang Viet Dinh[2], Nguyen Tien Quang[2], and Huynh Thi Thanh Binh[2]

[1]Faculty of Information Technology, VietNam National University Agriculture
[3]R&D Department, Anvita. JSC, Vietnam
[2]School of Information and Communication Technology, HUST, Vietnam

***Abstract*—Extraction of objects from images acquired by airborne sensors is the one of the most important topics in Aerial Photograph Interpretation (API). The task is challenging due to the very heterogeneous appearance of man-made and natural objects on the ground. Meanwhile images acquired by airborne sensors are very high-resolution, which requires high computational costs. This paper presents an efficient approach for automated extraction of objects at pixel level. We propose to combine a powerful classifier and an efficient contextual model for semantic segmentation of objects in images. Multiple image features are used to train the classifier, other features are used to learn the contextual model. We employ Random forest (RF) as classifier which allows one to learn very fast on big data. The outputs given by RF are then combined with a fully connected conditional random field (CRF) model for improving classification performance. Experiments have been conducted on a challenging aerial image dataset from a recent ISPRS Semantic Labeling Contest. We obtained state-of-the-art performance with a reasonable computational demand.**

Keywords*—Image segmentation, Semantic labeling, Object detection, Machine learning, Random forest, Deep learning, Aerial image, Remote sensing.

I. Introduction

Remote Sensing has been trending towards producing a great detail of its data, providing towards centimeter-level street-side image data. It is also taking advantage of an increase in methodological sophistication, which is greatly supported by rapid progress of the computational methods and computing environment. Recent development in remote imaging technology (satellite or aircraft-based sensor) has made not only improvement of traditional applications (e.g. mapping), but also opening up a number of advanced topics. Notable ones are the automatic creation of 3D models of cities for applications such as urban planning, disaster monitoring, Internet applications (Microsoft Virtual Earth, Google Earth), GIS and military applications, even Human mobility [5], [37]. Large area airborne images with details of objects at encourage models and applications for the location awareness of the Internet, object-level scene understanding. For which, automated Aerial photograph interpretation (API) is an essential need. These advanced topics in many cases only become feasible with the recent development of aerial imaging technology and new methods in computer vision.

The problem of automatic extraction of objects from aerial and space images has been an active research for decades [4], [24]. In recent years, the field has been extremely advanced by the computer vision research community for automated processing [18], [21], [25].

The problem of semantics object segmentation in aerial images is difficult for many reasons. Objects are of many kinds, including man-made (buildings, road nets) and natural ones (trees, grass, water surface) with very heterogeneous appearance. For example, let's consider buildings object. Buildings are complex object with many architectural details. Rooftops are usually composed of different materials with different reflectance properties, which may have low contrast to the ground. Buildings are located in urban scene with various objects which are in close proximity or disturbing, such as parking lots, vehicles, ground street, lamps, trees. These difficulties make the problem of building detection challenging. Beside that, images acquired by airborne sensors are often very high-resolution data, which requires high computational resources.

A number of methods have been proposed for solving the problem of objects detection in remote sensing images in literature [8], [17], [21], [23], [25]. These methods used different data modalities for object class modeling and different measurements for evaluations. No attempt has been made to exploit and combine multiple modalities of image data at pixel level effectively in an efficient learning model, to obtain satisfying performance for the detection and segmentation of objects. In this paper, we present a new approach based on a combination of a strong classifier (i.e., Random forest, RF) and a fully connected conditional random field (CRF) model on image features for the segmentation of objects in aerial images. The main contributions are: (1) a survey of state-of-the-art approaches for automated objects detection from aerial images and applications; (2) a pipeline for automated and segmentation of objects; and (3) an efficient learning framework for the proposed system.

II. Related work

There is a large amount of works in literature about semantic segmentation in remote sensing image data. With the success of the aerial imaging technology, high resolution images can be obtained cost-effectively. The acquired images often contain several types of data, i.e. color, infrared and panchromatic images [11], [24], [35]. Airborne images are usually taken from above, although with some constraints on the viewpoint, the appearance of the objects on the ground can vary significantly. Some works use single chanel of image data only [14], while other works employed data from multiple image chanels, including color and high field data [8], [20], [24], [32], [38].

978-1-5386-2490-6/17/$31.00 ©2017 IEEE

Training a traditional classifier for visual object detection is mainly based on locally extracted features, in which visual information, such as color and texture are usually combined together to represent the object class. A concatenating of different feature types into a single vector for learning may cause a drawback, i.e. one feature type may inhibit the performance of another; the learned classifier could be overfitting due to redundancy and correlation in the input data. Besides that, standard learning algorithms, such as logistic regression and support vector machines (SVM) assume that the data is independent and identically distributed. This is not true in many cases, as image pixels possess dependencies, e.g. if a pixel is labeled as building, it is likely that a neighboring pixel is also labeled as building; non-building pixels tend to be next to other non-building pixels. These dependencies should be exploited to improve the classification performance. There have been a number of works attempted to exploit contextual model to improve classification result, e.g. taking into account evidences surround an object to decide keeping or rejecting a hypothesis [11], [19], [29]. It has been shown that using surround evidences (contextual information) the performance of a state-of-the-art learning method can be improved. Conditional random field has been used to model contextual information for detection of urban areas [36] or objects from aerial images [20], [33], [34]. In our previous work [28] we have developed a similar framework for the segmentation of objects in satellite image. We obtained a good performance, however the work focused on only one object class, i.e. buildings.

The very recent research in machine learning, deep learning algorithms, have been applied for semantic segmentation of remote sensing imagery [24], [25]. This model can give state-of-the-art results on the segmentation accuracy. However, those models are complicated, requires large training data, computational demanding, and the model architecture is difficult to understand to most of off-the-topic researchers.

Recently, ISPRS has launched a benchmark data set for semantic object labeling contest [7]. The data set provides ground truth for evaluation of various proposed methods. There have been a number of research groups participated in the contest. The results of very recent works reported on the website show attempts of researchers in developing efficient methods for automated object detection and segmentation from aerial imagery. Despite that, the problem of how to effectively detection and segmentation of objects at pixel-level from high resolution aerial images remains a challenge. In this work, we extend our previous work by surveying more recent works, applying the proposed method on all five classes and making further analysis of the utilized techniques. Experimental results show that the performance of our approach is comparable to stat-of-the-art methods in terms of segmentation accuracy while with less computational costs.

III. PROPOSED METHOD

A. Background

1) Random Forest: A randomized forest consists of an ensemble of T trees in a hierarchical structure. The nodes of each tree include decisions for splitting down the tree until a leaf node is reached. Each leaf node s in a given maximal depth is associated with a learned class distribution $P(x|s)$, where x can take any value from a predefined set of labels $L = \{1, 2, \ldots, l\}$. The final class distribution over the whole forest can be achieved by averaging class distributions over the leaf nodes $S = \{s_1, s_2, \ldots, s_T\}$ $P(x|S) = \frac{1}{T}\sum_{t=1}^{T} P(x|s_t)$. As demonstrated in [13], [26], RF classifiers give robust and accurate classification results in both binary and multi-class classification problems. Due to the use of only a subset for training and the efficient decisions on evaluation, RFs are extremely fast for training and testing on large amounts of data, which is appropriate in the case of large aerial images.

2) Conditional Random Field Model: A Conditional Random Field (CRF), used in the context of pixel-wise label prediction, models pixel labels as random variables that form a Markov Random Field (MRF) when conditioned upon a global observation. The global observation is usually taken to be the entire image.

Let $\mathbf{X}$ be a random field over the set of random variables $\mathbf{X} = \{X_1, X_2, ..., X_N\}$, where N is the number of pixels in the image, and X_i is the random variable corresponding to pixel i, which denotes the predicted label of pixel i.X_i can take arbitrary value from a predefined set $L = \{1, 2, \ldots, l\}$. Let $\mathbf{Y}$ represent the features corresponding to image pixels. The pair of random fields $(\mathbf{Y}, \mathbf{X})$ can be treated as a CRF model, which can be described by a Gibbs distribution:

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

where $E(\mathbf{x})$ is the energy of the label assignment $\mathbf{x} \in L^N$ and $Z(\mathbf{Y})$ is the normalization function [12]. The energy $E(\mathbf{x})$ in the fully connected pairwise CRF model can be expressed as follows:

$$E(\mathbf{x}) = \underbrace{\sum_{i=1}^{N} \psi_u(x_i)}_{unary} + \underbrace{\sum_{i<j} \psi_p(x_i, x_j)}_{pairwise}. \tag{2}$$

where the unary energy components $\psi_u(x_i)$ measure the cost of the pixel i assigned the label x_i, and pairwise energy components $\psi_p(x_i, x_j)$ indicate the cost of assigning labels x_i, x_j to pixels i, j respectively. In our model, unary energies are achieved by applying a RF classifier, which independently predicts class labels for pixels without regarding the interrelations between neighboring pixels. The pairwise energies indicate the smoothness of label assignment. As in [10], pairwise energies are defined by weighted Gaussians:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w^{(m)} k_G^{(m)}(\mathbf{y}_i, \mathbf{y}_j). \tag{3}$$

where $k_G^{(m)}$, $m = 1, ..., M$, is a Gaussian kernel applied on feature vectors, $\mathbf{y}_i$ is the feature vector of pixel i. The function $\mu(.,.)$ introduces a penalty when neighboring similar pixels are assigned by different class labels.

B. Learning the model

The framework of our proposed model is presented in Fig.1. There are several learning steps in the model to obtain the final result.

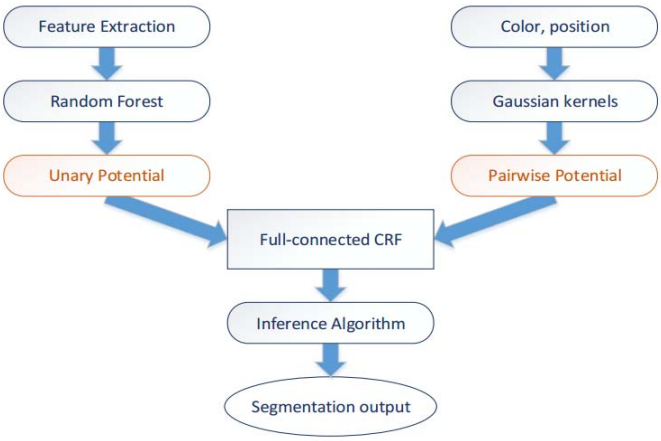


Fig. 1. The architecture of our proposed framework for semantic labeling of images.

After extracting feature vectors, we train a random forest classifier and compute the unary energies for CRF models. Random forest model used in this work is the CART-RF of Breiman [2]. The training algorithm for random forest applies the general technique of bootstrap aggregating (bagging) to tree learners. Given a training set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$ where $\mathbf{y}_j$ is the feature vector of pixel j, with true labels $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ where $x_j \in L = \{1, \ldots, l\}$, we bagging repeatedly select random samples with replacement from the training set and fit trees to these samples:

for $t = 1, \ldots, T$ **do**
Sample with replacement n training samples $(\mathbf{Y}_t, \mathbf{X}_t)$ from $(\mathbf{Y}, \mathbf{X})$.
Train a decision tree f_t on $(\mathbf{Y}_t, \mathbf{X}_t)$.
endfor

In training a random forest model, the selection of number of trees in the forest is important for the performance of the RF classifier. Fig. 2, we show how the number of trees in the random forest affects the validation accuracy during the training phase. Based on this observation, we choose the random forest with 200 trees for our next experiments.

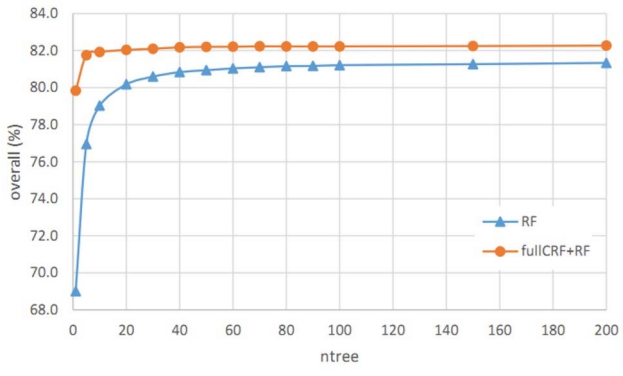


Fig. 2. Effects of the number of trees in a RF on the validation accuracy.

After training, one can predict the label of a new sample $\mathbf{y}'$ by averaging the predictions from all the individual decision trees in the forest:

$$\hat{f} = \frac{1}{T}\sum_{b=1}^{T} f_t(\mathbf{y}'). \tag{4}$$

The use of random forests has several advantages such as the computational efficiency, the probabilistic output, the seamless handling of a large variety of visual features and the inherent feature sharing of a multi-class classifier. However, random forests treat the image pixels independently without regarding the interrelations between them. Therefore, in the later process, we can further improve the segmentation results by employing an efficient inference model (CRF) that can exploit the relationships between neighboring pixels.

Minimizing the CRF energy $E(\mathbf{x})$ in 2 yields the most probable label assignment $\mathbf{x}$ for the given image. Since the exact minimization of $E(\mathbf{x})$ is intractable, an approximate approach of optimization should be used instead. Particularly, the Mean-Field method [10] can be used to compute a distribution $Q(\mathbf{X})$ that best approximates the probability distribution $P(\mathbf{X})$ of the model. $Q(\mathbf{X}) = \prod_i Q_i(X_i)$ is a product of independent marginals over each of the variables. Each of the marginals is constrained to be a proper probability distribution: $\sum_{x_i} Q_i(X_i = x_i) = 1$ and $Q_i(X_i) \geq 0$. The Mean-Field method minimizes the KL-divergence as follows:

$$\begin{aligned} D(Q \parallel P) &= \sum_i Q_i(x_i) \log \frac{Q_i(x_i)}{P_i(x_i)} \\ &= \sum_i Q_i(x_i) \log Q_i(x_i) + Q_i(x_i) \sum_i \psi_i(x_i) \\ &+ Q_i(x_i) \sum_{i<j} \psi_p(x_i, x_j) + \log Z(\mathbf{Y}). \end{aligned} \tag{5}$$

Traditional mean field inference [9] performs the following message passing update on each marginal Q_i in turn until all marginal probabilities are converged:

$$Q_i(x_i) = \frac{1}{Z_i} exp\left(-\psi_u(x_i) - \sum_{j \neq i} \sum_{x_j} \psi_p(x_i, x_j) Q_j(x_j)\right). \tag{6}$$

where Z_i is the marginal normalization function. Each iterator is guaranteed to decrease the KL-divergence, thus this inference algorithm is guaranteed to converge to a local optimum [9], [30]). In message passing the computational bottleneck is the evaluation of the sum $\sum_{j \neq i} \sum_{x_j} \psi_p(x_i, x_j) Q_j(x_j)$. The computational complexity of a single update of a marginal $Q_i(X_i)$ is $O(N)$ and the complexity of updating all the marginals is $O(N^2)$. Fortunately, Krahenbuhl [10] observed that a high dimensional Gaussian filter can be used to update all the mean field marginals concurrently in time O(N), that makes inference tractable.

IV. Experiment and Discussion

A. Data set

In this paper, we use the data set released by ISPRS WG III/4 for the Urban classification test project [1]. The test requires detailed 2D semantic segmentation that assigns labels

[1] http://www2.isprs.org/commissions/comm3/wg4/tests.html

to multiple object categories. Six categories/classes have been defined, include: Impervious surfaces (RGB: 255, 255, 255), Building (RGB: 0, 0, 255), Low vegetation (RGB: 0, 255, 255), Tree (RGB: 0, 255, 0), Car (RGB: 255, 255, 0), and Clutter/background (RGB: 255, 0, 0).

The experiments were conducted on Vaihingen 2D semantic labeling dataset [7] acquired over Vaihingen city in Germany. The dataset contains 33 large image patches, each of which consists of a true orthophoto (TOP) extracted from a larger TOP mosaic and a Digital Surface Model (DSM). Additionally, we also use a nDSM provided by [16]. The average size of such a patch is about 15MB, while the resolution of a patch is varied from 2336×1281 upto 3816×2550. The ground sampling distance of both, the TOP and the DSM, is 9 cm. Ground truth is provided only for 16 patches, meanwhile the remaining 17 patches are withheld by the challenge organizers for testing. We then divide the 16 annotated patches into training and validation sets. The training set consists of 11 patches selected randomly, the validation set consists of 5 remaining patches.

B. Feature extraction for object class representation

Feature extraction is for the representation of objects, one of the most important steps in learning for object recognition system. The more informative features are extracted, the better the description of objects and therefore the learning process. In this work, features should be well represented for object at pixel-level. Aerial image data often provides rich image features. Based on our experiences of working with remote sensing image data, in this work, the following features are extracted for the description of object classes.

Texton: Texton is widely used to represent a unit of image texture. This representation has been shown to be effective because the pixels represented with textons will contain more useful information than in the form of normal color [31].

Color: We employ the CIELab color space for image representation. Lab color is designed to approximate human vision. It aspires to perceptual uniformity, and its L component closely matches human perception of lightness.

Saturation: We use of CIR image as some previous works have shown that the saturation is helpful to further support the separation of vegetation and impervious surfaces.

Entropy: Entropy is gathered over a 9×9 neighborhood from the DSM to exploit spatial context information of a pixel (neighboring).

NDVI (the normalized digital vegetation index): computed from the first (IR) and the second channels (R) of the CIR true-orthophoto (TOP) image data.

$$NDVI = \frac{IR - R}{IR + R} \tag{7}$$

The use of the NDVI is based on the fact that green vegetation has low reflectance in the red spectrum (R) due to chlorophyll and much higher reflectance in infrared spectrum (IR) due to its cell structure. Hence, this is a good feature to distinguish green vegetation from other classes.

NDSM: the difference between the DSM and the derived DTM, which distinguishes pixels into ground and off-ground ones.

$$NDSM = DSM - DTM \tag{8}$$

This feature allows to separate object classes with high elevation from the low ones.

TABLE I. ACCURACY OF FULL_REFERENCE TYPE OF RESULTS

→ Reference ↓ Predicted	Imp_surf	Building	Low_veg	Tree	Car	clutter
Imp_surf	**0.907**	0.043	0.032	0.017	0.001	0.0
Building	0.080	**0.899**	0.008	0.012	0.001	0.0
Low_veg	0.112	0.038	**0.688**	0.162	0.0	0.0
Tree	0.032	0.012	0.096	**0.860**	0.0	0.0
Car	0.743	0.092	0.012	0.006	**0.147**	0.0
clutter	0.811	0.162	0.017	0.005	0.005	**0.0**
Precision	0.785	0.902	0.818	0.819	0.792	-
Recall	0.907	0.899	0.688	0.860	0.147	0.0
F1	0.842	0.900	0.747	0.839	0.247	-
Overall	**82.9**					

TABLE II. ACCURACY OF NO_BOUNDARIES TYPE OF RESULTS

→ Reference ↓ Predicted	Imp_surf	Building	Low_veg	Tree	Car	clutter
Imp_surf	**93.1**	3.6	2.3	1.0	0.0	0.0
Building	7.0	**91.5**	0.5	0.8	0.1	0.0
Low_veg	10.1	2.9	**72.2**	14.8	0.0	0.0
Tree	2.1	0.8	8.0	**89.1**	0.0	0.0
Car	69.9	11.0	1.0	0.3	**17.8**	0.0
clutter	0.811	0.169	0.014	0.001	0.005	**0.0**
Precision	81.4	92.4	85.5	84.7	79.5	-
Recall	93.1	91.5	72.2	89.1	17.8	0.0
F1	86.9	92.0	78.3	86.9	29.0	-
Overall	**85.9**					

C. System setting

In the experiments, for each test set we run the system 20 times. All the programs are run on a machine with CPU Intel Core i7-4770K (8 CPUs), RAM 16GB DDRIII 1600Mhz, Window 8.1, and implemented in R and C++ program languages.

D. Result

To demonstrate the performance of our framework, we compare experimental results of our approach to some of the recent proposed systems and to the state-of-the-art deep learning approaches.

In the following, we compare our results with recent methods reported in [7] on the ISPRS Semantic Labeling Benchmark dataset.

Following the description by the ISPRS contest committee, results should be provided in two types, with full_reference and with no_boundaries. The full_reference result is produced over all the pixels of the test patches. Meanwhile, in the no_boundary reference result, the boundaries of objects are eroded by a circular disc of 3 pixel radius. Those eroded areas are then ignored in order to reduce the impact of uncertain border definitions on the evaluation.

The evaluation is shown in the form of an accumulated confusion matrix. We then compute recall, precision and F1_score per class as follows:

$$Precision = \frac{\#true\ positive}{\#true\ positive + \#false\ positive}. \tag{9}$$

TABLE III. SEGMENTATION RESULTS OF DIFFERENT METHODS

Method	Imp_surf	Building	Low_veg	Tree	Car	Overall
Stair Vision Library (SVL_3) [3]	86.6	91.0	77.0	85.0	55.6	84.8
Rule-based [27]	84.3	88.7	74.5	82.0	9.9	81.8
FCN (UZ_1)	89.2	92.5	81.6	86.9	57.3	87.3
FCN (UOA) [15]	89.8	92.1	80.4	88.2	82.0	87.6
FCN trained with no-downsampling (DST_1) [25]	90.19	94.49	77.69	87.24	76.77	87.70
FCN trained with no-DS + RF + CRF (DST_2) [25]	90.41	94.73	78.25	87.25	75.57	87.90
Our RF + Full CRF	86.9	92.0	78.3	86.9	29.0	85.9

$$Recall = \frac{\#true\ positive}{\#true\ positive + \#false\ negative}. \quad (10)$$

$$F1\_score = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (11)$$

The overall accuracy is derived from the normalization of the trace from the confusion matrix. These measures are computed for both full_reference and no_boundaries.

As described by the ISPRS contest committee, the boundaries between classes are eroded by a circular disc of 3 pixel radius. Those eroded areas are then ignored during evaluation. The motivation is to reduce the impact of uncertain border definitions on the evaluation. Our results are shown in Table I and Table II. Description of the results can be seen in [6].

The comparison between different methods is shown in Table III. As one can see from the Table III, our proposed framework achieves slightly better results than recent approaches such as Rule-based [27] and Stair Vision Library (SVL_3) [3], and yields competitive accuracies with other state-of-the-art deep learning methods.

It can be seen that, deep learning methods require much more computational time for training and testing than our framework. For example, ONE_5 [22] ensembles three separate CNNs with three different input sizes: 16×16, 32×32, 64×64. To demonstrate the performance of CNN-based methods, we evaluate a simplified CNN model on image patches of size 32×32, and is implemented using Torch7 library with CUDA support [1]. The simplified CNN has the same architecture as described in [22]. We test this model on a powerful Dell Precision T7610 Workstation with Intel Xeon 8 Core E5-2650V2 2.60 GHz, 32GB DDR3 RAM and Tesla K20c. The average time required for training and test phrases for both our framework, the simplified CNN model and another effective FCN [25] is shown in Table IV.

TABLE IV. AVERAGE TIME FOR TRAINING AND TESTING PHRASES

Method	Average training time(s)	Test time per image(s)
FCN trained with no-downsampling [25]	223200	60
Our reimplemented simplified CNN [22] model with CUDA support	26640	4320
Our framework with parallel processing with 8 threads	**843**	**50**

From Table IV, we observe that the CNN method, even with its simplified version, requires much more computational time than our framework. Specifically, our framework is about 25 times faster in training compared to the simplified CNN model trained on Dell Precision T7610 Workstation with CUDA support, and about 50 times in test phrase. Compared to the FCN model [25], our framework can be around 265 times faster in training. In the test phase, our frame work is slightly faster than the FCN [25].

Finally, Fig. 3 demonstrates the improvement of segmentation result given by the fully connected CRF model to the probabilistic results of RF on the test image patch number 11. It is noticed that the fully connected CRF can work on eliminating the misclassified pixels (considered as noise) from the unary classifier's output RF.

V. CONCLUSION

We have presented an efficient framework for automated extraction of objects from images acquired by airborne cameras. The model combines classification results from a powerful classifier (Random forest) with contextual potential given by a fully connected conditional random field. The RF classifier is performed on informative features extracted from image data, meanwhile CRF works on color and location of pixels in images. We obtain a promising classification results while keeping a low computational cost. For future work, we will explore further on object detection models, such as for the detection of objects with regular shapes, i.e. cars. This will help to improve classification accuracy for the unary classifier, and give more information for the contextual model (cars are likely appear on road). On the other hand, the architecture deep learning models should be further explored to find suitable ones for efficiently working with large aerial image data.

REFERENCES

[1] Torch 7 library. http://http://torch.ch/. Accessed: 2017-09-10.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] M. Gerke. Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen). 2014.

[4] A. Gruen. *Automatic Extraction of Man-Made Objects from Aerial and Space Images (II)*. Birkhauser Boston, 1998.

[5] W. Huang and S. Li. Understanding human activity patterns based on space-time-semantics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016.

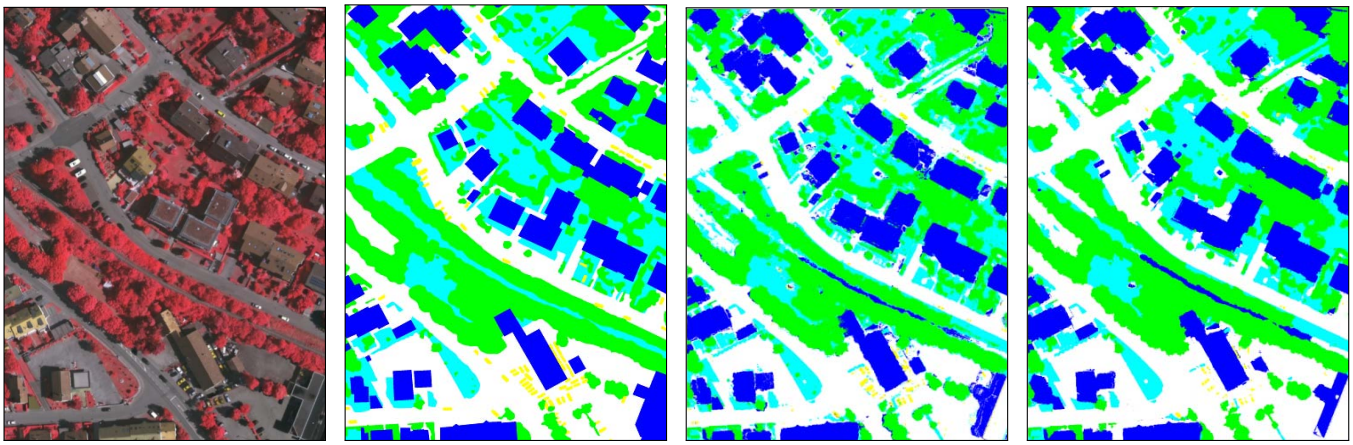
[6] ISPRS Working group III/4. Detailled semantic labeling (2d) result for hust team. http://www.itc.nl/external/ISPRS_WGIII4/ISPRSIII_4_Test_results/2D_labeling_vaih/2D_labeling_Vaih_details_HUST/index.html. Accessed: 2017-09-10.

[7] ISPRS Working group III/4. Isprs 2d semantic labeling contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. Accessed: 2017-09-10.

[8] C. Jaynes, E. Riseman, and A. Hanson. Recognition and reconstruction of buildings from multiple aerial images. *Computer Vision Image Understanding*, 90(1):68–98, 2003.

[9] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[10] V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*, 2011.



(a) Input image (b) Ground-truth (c) RF (d) RF+CRF

Fig. 3. An illustration of improving segmentation result using CRF over RF's output.

[11] F. Korc and W. Forstner. Interpretation terrestrial images of urban scenes using discriminative random fields. In *Proceedings of the Congress of the International Society for Photogrammetry and Remote Sensing*, pages B3a: 291–296, 2008.

[12] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[13] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1465–1479, 2006.

[14] C. Lin and R. Nevatia. Building detection and description from a single intensity image. *Int. Journal Computer Vision and Image Understanding*, 72(2):101–121, 1998.

[15] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[16] I. Markus Gerke. Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen).

[17] B. Matei, H. Sawhney, S. Samarasekera, J. Kim, and R. Kumar. Building segmentation for densely built urban regions using aerial lidar data. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[18] H. Mayer. Object extraction in photogrammetric computer vision. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(2):213–222, March 2008.

[19] S. Mueller and D. W. Zaum. Robust building detection in aerial images. In *ISPRS Workshop on Object Extraction for 3D City Models, Road Databases and Traffic Monitoring - Concepts, Algorithms, and Evaluation (CMRT05)*, 2005.

[20] T. T. Nguyen, S. Kluckner, H. Bischof, and F. Leberl. Aerial photo building classification by stacking appearance and elevation measurements. In *ISPRS TC VII Symposium*, 2010.

[21] F. Y. V. O. Firat, Gulcan Can. Representation learning for contextual object and region detection in remote sensing. In *In ICPR, 2014 22nd International Conference on*, 2014.

[22] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Hengel. Effective semantic pixel labelling with convolutional networks and conditional random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–43, 2015.

[23] M. Persson, M. Sandvall and T. Duckett. Automatic building detection from aerial images for mobile robot mapping. In *Symp. on Comp. Intel. in Robotics & Automation*, 2005.

[24] C. K. Ronald Kemker, Carl Salvaggio. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. arXiv:1703.06452.

[25] J. Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016.

[26] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[27] T. Speldekamp, C. Fries, C. Gevaert, and M. Gerke. Automatic semantic labelling of urban areas using a rule-based approach and realized with mevislab. 2015.

[28] N. T. Thuy, D. V. Sang, and H. T. T. Binh. An efficient framework for pixel-wise building. segmentation from aerial images. In *Prod. of SoICT2015*, 2015.

[29] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8, jun 2007.

[30] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[31] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.

[32] M. Xie, K. Fu, and Y. Wu. Building recognition and reconstruction from aerial imagery and lidar data. In *Proceedings of the International Conference on Radar*, pages 1–4, Oct. 2006.

[33] J. Yao and Z. M. Zhang. Semi-supervised learning based object detection in aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1011–1016, Washington, DC, USA, 2005. IEEE Computer Society.

[34] M. Z. Z. Zhang, M.Y. Yang. Multi-source multi-scale hierarchical conditional random field model for remote sensing image classification. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-3/W4*, 2015.

[35] L. Zebedin, A. Klaus, B. Gruber-Geymayer, and K. Karnera. Towards 3d map generation from digital aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(6):413–427, 2006.

[36] P. Zhong and R. Wang. Object detection based on combination of conditional random field and markov random field. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 160–163, 2006.

[37] Q. Zhu, M. Hu, Y. Zhang, and Z. Du. Research and practice in three-dimensional city modeling. *Geo-Spatial Information Science*, 12(1):18–24, March 2009.

[38] P. Zimmermann. A new framework for automatic building detection analyzing multiple cue data. *International Archives of Photogrammetry and Remote Sensing*, 33:1063–1070, 2000.