BUILDING DETECTION AND SEGMENTATION USING A CNN WITH AUTOMATICALLY GENERATED TRAINING DATA

Xiangyu Zhuo[1], *Friedrich Fraundorfer*[2,1] , *Franz Kurz*[1] *and Peter Reinartz*[1]

[1] Remote Sensing Technology Institute, German Aerospace Center, Germany
[2] Institute for Computer Graphics and Vision, Graz University of Technology, Austria

ABSTRACT

Significantly outperforming traditional machine learning methods, deep convolutional neural networks have gained increasing popularity in the application of image classification and segmentation. Nevertheless, deep learning-based methods usually require a large amount of training data, which is quite labor-intensive and time-demanding. To deal with the problem in generating training data, we propose in this paper a novel approach to generate image annotations by transferring labels from aerial images to UAV images and refine the annotations using a densely connected CRF model with an embedded naive Bayes classifier. The generated annotations not only present correct semantic labels, but also preserve accurate class boundaries. To validate the utility of these automatic annotations, we deploy them as training data for pixel-wise image segmentation and compare the results with the segmentation using manual annotations. Experiment results demonstrate that the automatic annotations can achieve comparable segmentation accuracy as the manual annotations.

Index Terms— Image segmentation, Automatic image annotation, Label propagation, Deep learning

1. INTRODUCTION

As one of the hottest topic in machine learning research, deep learning has been widely applied in image classification and segmentation and demonstrated significant improvement compared to traditional machine learning methods. Nevertheless, a large number of ground truth annotations are required to train the deep convolutional neural networks. Though there are several open image databases like ImageNet[1] and LabelMeFacade[2], they are only applicable for specific scenes. When it comes to the applications in photogrammetry and remote sensing, it is usually inevitable to create annotations manually, which costs plenty of time and labor.

In order to tackle the lack of training data, a couple of attempts have been made in automatic generation of image annotations. [3] proposed to generate synthetic images with pixel-level annotations, but this method relies on rich generative models to generate new and distinctive annotations. Video-based algorithms [4] propagate the labels of annotated frames to neighboring frames, however, such methods are vulnerable to occlusions of different classes. Some methods exploit LiDAR point cloud or 3D reconstruction of the scene as an intermediary. [5] proposed to transfer labels from 3D to 2D, i.e., manually annotate a point cloud in 3D domain and then project the labels back into 2D domain. The results are promising, however, the task of labeling the point cloud is still indispensable and quite labor intensive. Considering the fact that aerial images usually have much larger coverage than UAV images, we seek to propagate the labels from one aerial image to multiple co-registered UAV image and the UAV point cloud. Theoretically, we simply need to annotate one aerial image manually and then transfer the labels to numerous co-registered UAV images of the same area. To this goal, we propose in this paper a new pipeline for automatic image annotation generation. This approach consists of three steps: 1. label one or two aerial images manually; 2. transfer the pixel labels to multiple UAV images via the UAV point cloud; 3.refine the generated annotations with a densely connected CRF model and a naive Bayes classifier. To validate the accuracy of automatically generated annotations, we also train a deep convolutional neural network with the automatic annotations for image segmentation and compare the performance with manual annotations.

2. METHODOLOGY

2.1. Annotation

Main differences between aerial imagery and UAV imagery lie in their resolution, scale and viewing directions. Due to occlusion, one aerial image may not contain the same objects as UAV images, e.g., the building facades. In order to exploit the complete information from the oblique view aerial images, we label two aerial images from the surveying area, one is left-view and the other is right-view, denoted by I_l and I_r respectively. Figure 1 depicts an oblique UAV imagery and the corresponding region in the left-view and right-view aerial images. It can be seen that the combination of the two views contribute to complete representation of the scene.

In our case, the interesting categories are *Building*, *Roof*, *Ground*, *Vegetation* and *Car*, while the indistinguishable objects are labeled as *Clutter* and will not be involved in segmentation. The labels are then transferred from aerial images to UAV images via the point cloud reconstructed from UAV images, i.e., we project all the 3D points into a labeled aerial image, thus all the non-occluded points get labeled, and then we project the 3D points into UAV images, transferring their labels to corresponding pixels. Figure 2 (a), (b) show a UAV image with labels transferred from left-view aerial image and right-view aerial image respectively. It needs to be noted that occlusions and moving cars result in label inconsistence at some areas, to tackle this problem, we propose to refine the raw annotations with the joint reasoning of pixel information in the image and auxiliary 3D information of the UAV point cloud.

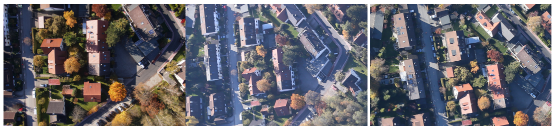
(a) (b) (c)

Fig. 1: Comparison of oblique aerial imagery and oblique UAV imagery. (a) is the UAV image, (b) and (c) show the corresponding region in the left-view and right-view aerial images

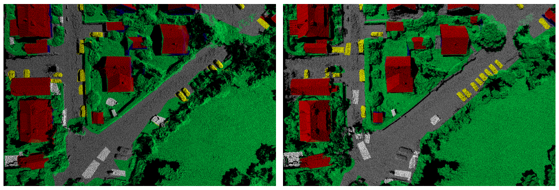
(a) (b)

Fig. 2: Raw UAV image annotation with labels transferred from aerial images. (a) is transferred from the left-view aerial image and (b) from the right-view

2.2. Model

Let $\mathcal{X}$ denote a random field over a set of variables $\{X_1, X_2, ..., X_N\}$ and the domain of each variable X_i is a set of semantic labels $\mathcal{L} = \{l_1, l_2, ..., l_k\}$. The Gibbs energy function of a label $x \in \mathcal{L}^N$ is:

$$E(x) = \sum_{i}^{N} \psi_u(x_i) + \sum_{i<j}^{N} \varphi_p(x_i, x_j) \quad (1)$$

The unary potentials $\psi_u(\cdot)$ feature the probability of a pixel taking label x_i. In our case, the pixel-wise probability $P(x_i)$ is initially derived from the transferred labels of UAV images. Due to occlusions, however, some pixels may have wrong labels, no labels or multiple labels, especially for classes *Building* and *Car*. To solve this problem, we exploit geometric information embedded in 3D points such as height and normal vectors. At ambiguous regions, these complimentary cues can help to distinguish different classes. More specifically, consider a set O defined over possible observation values, in our case the height value. Let $P(O \mid x_i)$ be the prior probability distribution of O defined based on image statistics. The likelihood of pixel i taking label s_i given the observation O can be obtained via a naive Bayes classifier:

$$\varphi_p\left(x_i \mid O\right) = \frac{P(O \mid x_i)P\left(xi\right)}{P(O)} \tag{2}$$

For the sake of notation convenience, we will use $\varphi_p\left(x_i\right)$ instead of $\varphi_p\left(x_i \mid O\right)$ in the following text. Pairwise potentials $\varphi_p(\cdot)$ encode the semantic label coherence of pixels. We adopt the contrast-sensitive Gaussian edge kernels as pairwise term, which have the form:

$$\varphi_p\left(x_i, x_j\right) = \omega_1\left(x_i, x_j\right) exp\left(-\frac{\mid p_i - p_j \mid^2}{2{\theta_\gamma}^2}\right) \\ + \omega_2\left(x_i, x_j\right) exp\left(-\frac{\mid p_i - p_j \mid^2}{2{\theta_\alpha}^2} - \frac{\mid I_i - I_j \mid^2}{2{\theta_\beta}^2}\right) \tag{3}$$

Where I_i is the color vector of pixel i and p_i denote its position. θ_α, θ_β and θ_γ control the degree of nearness, similarity and smoothness. The parameter values in our experiment are 20, 8, 3 respectively. In the end, the parameters in our CRF model are learned via minimization of the Gibbs energy defined in Eq. 1

3. EXPERIMENTAL RESULTS

3.1. Generated annotation

We performed learning and inference of the CRF model based on the implementation of [6] [1]. Figure 3 depicts the automatically generated annotation in comparison with the manually labeled annotation. It can be seen that the automatic annotation not only present correct semantic labels, but also preserve accurate class boundaries.

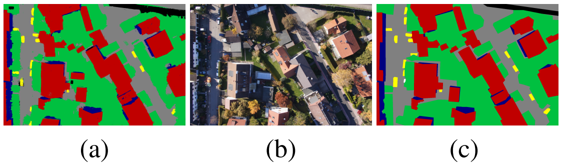

Fig. 3: Comparison of transferred annotations with manual annotations. (a) is automatic annotation and (c) is manual annotation, (b) is corresponding UAV image

3.2. Image segmentation with generated annotations

In order to validate the utility of the automatically generated annotations, we use them as training data for image semantic segmentation using a deep convolutional neural network and compare the performance with the segmentation using manual annotations. To be specific, we select 19 generated annotations featuring different regions of the survey site. In order to enrich the training data, we augment the annotations by scaling and rotating, resulting in around 8208 images with the size of 300×300. The corresponding manual annotations are also augmented in the same way with the same parameters.

The learning procedure is implemented under the deep learning framework Caffe. We fined tune the FCN-8s model [7] [2] with our dataset. Besides, we plug in the CRF-RNN [8] [3] layer in order to achieve sharp edges at class borders. Figure 4 shows the segmentation result of the trained network on the test data. Where, (a) shows the segmentation results using automatic annotations, (c) illustrates the results using manual annotations and (b) is the corresponding ground truth. The accuracy of each class is listed in Table 1. It can be seen that the segmentation using the automatic annotations achieves comparable accuracy with the segmentation using manual annotations.

[1] https://github.com/lucasb-eyer/pydensecrf

[2] https://github.com/shelhamer/fcn.berkeleyvision.org

[3] https://github.com/torrvision/crfasrnn

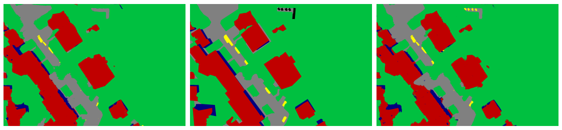

Fig. 4: Comparison of segmentation results. (a) is the segmentation result using automatic annotations, (c) is the segmentation using manual annotations, (b) is the corresponding ground truth.

Table 1: Segmentation accuracy (IoU) comparison between automatic annotations (A) and manual annotations (M).

	Building	Roof	Ground	Car	Veg
A	89.47	48.11	96.14	35.30	77.98
M	87.49	50.64	96.34	41.77	77.81

4. CONCLUSION

In this paper, we propose a novel approach to generate image annotations by transferring labels from aerial images to UAV images. Due to occlusion and inaccuracy of manual annotation, the transferred annotation may carry incomplete or wrong information, we refine the raw annotations using a densely connected CRF model with an embedded naive Bayes classifier. The generated annotations not only present correct semantic labels, but also preserve accurate class boundaries. To validate the effectiveness of the generated annotations, we deploy them as ground truth data in image segmentation. Experiment results demonstrate that the segmentation performance using the automatic annotations is comparable with manual annotations while saving the manual labor and time dramatically. The proposed method can be effectively applied in automatic generation of training data.

5. REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

[2] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman, "Labelme: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, May 2008.

[3] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] I. Budvytis, P. Sauer, T. Roddick, K. Breen, and R. Cipolla, "Large scale labelled video data augmentation for semantic segmentation in driving scenarios," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 230–237.

[5] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger, "Semantic instance annotation of street scenes by 3d to 2d label transfer," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *CoRR*, vol. abs/1210.5644, 2012.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[8] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr, "Conditional random fields as recurrent neural networks," in *International Conference on Computer Vision (ICCV)*, 2015.