EXTRACTION OF BUILDINGS IN VHR SAR IMAGES USING FULLY CONVOLUTION NEURAL NETWORKS

Muhammad Shahzad [1,2], *Member, IEEE*, Michael Maurer [4], Friedrich Fraundorfer [3,4], Yuanyuan Wang [2,3], *Member, IEEE*, Xiao Xiang Zhu [2,3], *Senior Member, IEEE*

(1) National University of Sciences and Technology (NUST), School of Electrical Engineering and Computer Science (SEECS), H-12 Islamabad, Pakistan

(2) Technische Universität München (TUM), Signal Processing in Earth Observation (SiPEO), Arcisstrasse 21, 80333 Munich, Germany (Email: muhammad.shahzad@tum.de, Tel: +49-81-5328-3096)

(3) German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Oberpfaffenhofen, 82234 Wessling, Germany (Emails: xiao.zhu@dlr.de, Tel: +49-81-5328-3531; wang@bv.tum.de, Tel: +49-81-5328-2386)

(4) Technical University of Graz (TU Graz), Institute for Computer Graphics and Vision (ICG), Inffeldgasse 16, 8010 Graz, Austria (Emails: fraundorfer@icg.tugraz.at, Tel: +43-316-873-5020; maurer@icg.tugraz.at, Tel: +43-316-873-5044)

ABSTRACT

Modern spaceborne synthetic aperture radar (SAR) sensors, such as TerraSAR-X/TanDEM-X and COSMO-SkyMed, can deliver very high resolution (VHR) data beyond the inherent spatial scales (on the order of 1m) of buildings, constituting invaluable data source for large-scale urban mapping. Processing this VHR data with advanced interferometric techniques, such as SAR tomography (TomoSAR), enables the generation of 3-D (or even 4-D) TomoSAR point clouds from space. In this paper, we present a novel and generic workflow that exploits these TomoSAR point clouds in a way that is capable to automatically produce benchmark annotated (buildings/non-buildings) SAR datasets. These annotated datasets (building masks) have been utilized to construct and train the state-of-the-art deep Fully Convolution Neural Networks with an additional Conditional Random Field represented as a Recurrent Neural Network to detect building regions in a single VHR SAR image. The results of building detection are illustrated and validated over TerraSAR-X VHR spotlight SAR image covering approximately 39 km^2 − almost the whole city of Berlin − with mean pixel accuracies of around 93.84%.

Index Terms— Synthetic Aperture Radar (SAR), Fully Convolution Neural Networks, SAR Tomography, Building Detection, OpenStreetMap, TerraSAR-X/TanDEM-X

1. INTRODUCTION

Automatic detection of man-made objects in particular buildings from single very high resolution (VHR) SAR image is of great practical interest especially when it comes to applications having stringent temporal restrictions e.g., emergency responses. However, due to inherent complexity of SAR images caused by the so-called speckle effect together with radiometric distortions mainly originating due to side looking geometry, scene interpretation from SAR images is highly challenging especially in the context of object recognition and 3-D reconstruction.

A variety of algorithms has been published in the literature that aims at detection and reconstruction of buildings from SAR images. Typically, most of the developed approaches rely on auxiliary information e.g., multi-sensor data provided by optical and LiDAR sensors, Geographic Information System (GIS) data e.g., 2-D building footprints, multi-dimensional data e.g., polarimetric SAR (PolSAR), or multi-view/multi-aspect data (i.e., the information extracted by imaging the area under investigation more than once with different viewing configurations) e.g., interferometric SAR (InSAR). These approaches improved the feature extraction process by providing complimentary information. To our knowledge, the literature using only single SAR image in context of building detection is rather sparse [1] [2] [3]. These approaches aimed to extract buildings in an unsupervised (or data-driven) manner and therefore were able to extract either simple or isolated buildings only.

Recently, the Convolution Neural Networks (CNNs) – type of multi-layered neural networks – have significantly outperformed other methods and became state-of-the-art in image classification. Use of CNNs over SAR images is up till now limited but consistently increasing [4]–[7]. As can be imagined, the precondition for application of CNNs or any other supervised learning frameworks is the availability

978-1-5386-7150-4/18/$31.00 ©2018 IEEE

of annotated datasets. They are necessary not only to analyze and validate the performance of classification algorithms but are also required in training phase where parts of annotated data are utilized to optimize prediction models. Lack of such annotated datasets is one of the major issues in application of CNNs over SAR images. Manual (or somewhat interactive) annotation, as is done in the aforementioned approaches, is one potential solution. However, it often requires expert's knowledge and easily become impractical when large scenes need to be processed. Thus, in view of above, automatic annotation of SAR images, if possible, is essential.

The objective of this paper is twofold: First is to demonstrate the potential of automatic preparation of SAR training datasets for larger regions; Secondly, using the automatically prepared dataset to train deep CNN architecture to detect buildings in a single very high resolution SAR image.

2. AUTOMATIC ANNOTATION OF SAR IMAGES

Annotating an image is fundamental for application of any supervised learning technique for segmentation/ classification purposes. For this reason, we propose a novel method that utilize the SAR tomography (TomoSAR) point clouds to automatically annotate (buildings/non-buildings) SAR images of the area of interest. Before proceeding further, we briefly introduce these point clouds and later demonstrate their usage in such automatic annotation.

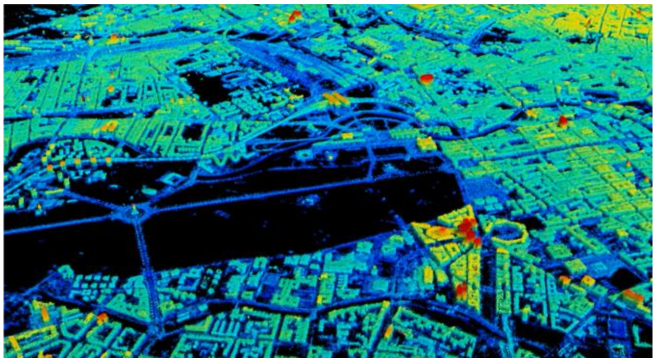

Figure 1: TomoSAR point clouds generated from TerraSAR-X data stacks of ascending and descending orbits (Site: city of Berlin). The color represents height. Black areas are temporally decorrelated objects, e.g. vegetation or water.

2.1. TomoSAR Point Cloud

SAR tomography (TomoSAR) is an advanced interferometric technique that aims at real and unambiguous 3-D SAR imaging. By exploiting stack(s) of SAR images taken from slightly different positions, it builds up a synthetic aperture in the elevation direction that enables retrieval of precise 3-D position of dominant scatterers within one azimuth range SAR image pixel [8]. The retrieved scatterer information when geo-coded into world coordinates produces TomoSAR point clouds [9], capable of containing not only the 3-D positions of the scatterer location but also the estimates of seasonal/temporal deformation. Figure 1 shows the generated TomoSAR point cloud of the city of Berlin, Germany, using DLR's tomographic processing system – Tomo-GENESIS [10].

2.2. Automatic Annotation

In this paper, we utilized these TomoSAR point clouds in generating labelled SAR images. The basic idea is to classify each 3-D point as belonging to buildings and non-buildings and later geo-code them back into their corresponding SAR (i.e., in azimuth and range) coordinates. The classification of each point is obtained by exploiting information pertaining to already available 2-D building footprints. To elaborate, the 2-D building footprints from OpenStreetMap (OSM) are downloaded from Geofabrik's website[1] which are subsequently utilized to automatically annotate the SAR image. I.e., all points within the building polygons are classified as buildings. Extracted building points are then projected back to SAR coordinates. A morphological dilation operator is then employed to finalize the building mask.

3. PROPOSED NETWORK ARCHITECTURE

The network architecture of the fully convolutional network (FCN) is based on the FCN structure of Long et al. [11]. To additionally integrate binary potentials we add a CRF represented as RNN [12]. This gives us an end to end trainable network as depicted in Figure 2.

In detail, the first part of our network calculates a feature for each input pixel. Therefore, we exploit a fully convolutional network (FCN) with in-network upsampling and skip and fuse architecture to fuse coarse, semantic and local, appearance information [13]. As we are using a FCN we exploit the ability to not only classify a single pixel as proposed in [5] [14] [4] but we perform image segmentation for input images of arbitrary size at once. Thus, we eliminate overhead calculations resulting from the sliding window approach.

The second part of the network adds binary potentials. This means it adds constraints to give neighboring pixels with similar intensity the same label. This is typically done using a Markov Random Field or to be more precise the special case of a fully connected Conditional Random Field (CRF) as presented by Krähenbühl et al. [15]. As an end to end trainable network is preferable, we added the dense CRF represented as a Recurrent Neural Network (RNN) further called CRF-RNN as proposed by [12]. This network was

[1] GEOFABRIK downloads, http://download.geofabrik.de/europe/germany/berlin.html

then modified to get a pixel wise two class classification representing building and non-building.

4. IMPLEMENTATION OF TRAINING ALGORITHM

We performed staged training as mentioned in [38] because it is less prone to divergence. First the single-stream FCN-32s is trained then the training continued with the two-stream FCN-16s and the three-stream FCN-8s. Next, the CRF-RNN is added and trained by keeping the FCN-8s part constant. Finally, a fine-tuning of the complete net has been performed. Each stage was trained for 400,000 iterations with constant learning rate ($1e^{-10}$, $1e^{-12}$, $1e^{-14}$, $1e^{-12}$ and $1e^{-12}$ for each stage respectively) a momentum of 0.99, weight-decay of 0.0005 and a pixel wise soft-max loss (that has been averaged over 100 images each epoch).

As the network contains convolutional layers as well as pooling layers the resulting segmented image is reduced in dimension. This is compensated by in-network upsampling layers whose parameters are initialized as bilinear filtering and further refined while training.

5. EXPERIMENTAL RESULTS & VALIDATION

5.1. Dataset

To validate our approach, we employed a TerraSAR-X high-resolution spotlight image and a 3-D TomoSAR point cloud of Berlin. The SAR image was acquired from ascending orbit with incidence angle of 36° which almost provides a full coverage of the whole city. For automatic annotation, the 3-D TomoSAR point clouds have been generated from stacks of 102 TerraSAR-X high spotlight images from ascending and descending orbits covering almost the whole city of Berlin using the Tomo-GENESIS software developed at the German Aerospace Center (DLR) [10].

5.2. Results of Automatic Annotation

Figure 3 shows the SAR intensity image covering almost the whole city of Berlin (around 39 km^2) while Figure 4 demonstrates the resulting mask of building regions obtained automatically using the OSM building footprints. Experimental results of training and testing/validation of deep learning based staged network architecture exploiting both these automatic annotations are presented in the subsequent sections.

5.3. Preparation of Training Data

We prepared the dataset for training by taking 11 out of 16 of the pre-classified input images covering almost the whole city of Berlin (using OSM + TomoSAR point cloud) and created patches of 256 x 256 pixels with an overlap of 32 pixels. Further, these patches are augmented by flipping and rotation. Finally, we got 26,312 image patches for training and used the remaining 5 out of 16 of the annotated input images for testing. It is also important to mention that to reduce speckle effect, we first performed non-local filtering of the SAR images prior to training using the algorithm as proposed in [16].

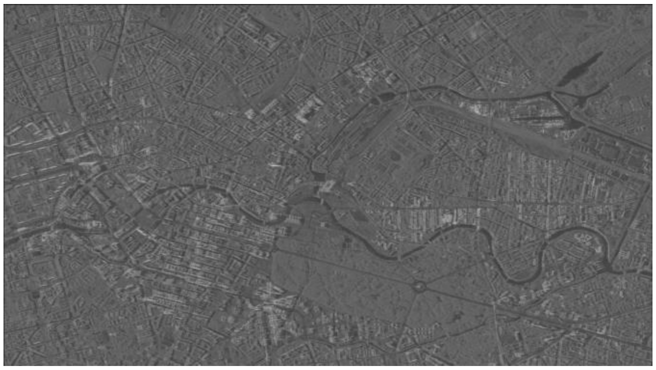

Figure 3: SAR intensity image covering almost the whole city of Berlin.

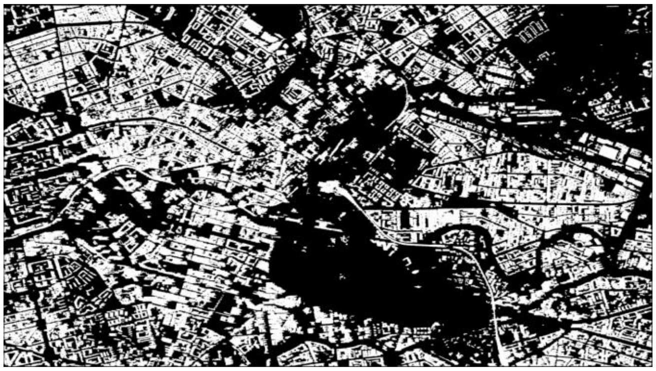

Figure 4: Automatically generated mask of building regions using OSM + TomoSAR point clouds for the SAR intensity image shown in Figure 3.

5.4. Results Analysis

The experimental results have been obtained after applying staged training where the results obtained after single-stream, then upgraded to two-stream and three-stream are depicted as FCN-32s (32x upsampled prediction), FCN-16s (16x upsampled prediction), and FCN-8s (8x upsampled prediction) respectively. In each respective stage, the network is learned from end-to-end in a cascaded manner i.e., all initialization parameters of the previous stage are fed as input to the subsequent one. If we denote the automatically generated annotated dataset using OSM + TomoSAR point cloud as OSM-Ref, then the Tables I and II depict the results acquired over the whole area of Berlin in different stages of the network architecture. As mentioned earlier, the OSM dataset is prone to errors introduced as a consequent of crowd sourcing, therefore for fair evaluation of network architecture, we manually prepared a more accurate annotated dataset (denoted as OSM-GT) for test sub images and utilized it for evaluation as depicted in Table II. In general, the upgraded three-stream FCN-8s with CRF-

RNN tends to show superior performance in distinguishing buildings from non-buildings. Figures 5 shows this result of FCN-8s with CRF-RNN trained using OSM-Ref overlaid onto the SAR image of Figure 3 covering almost the whole region of Berlin.

Table I: Accuracy analysis of obtained results using different stages of the trained network with following details: training & testing/validation using OSM-REF data. Evaluation metrics [11] [13]: PA - Pixel Accuracy, MA - Mean Accuracy, MIU - Mean IU, FWIU - Frequency Weighted IU, FAR - False Alarm Rate, QR - Quality Rate.

Network Architecture	PA	MA	MIU	FWIU	FAR	QR
FCN-32s	83.28	82.40	69.47	71.98	20.39	71.58
FCN-16s	83.46	82.43	69.70	72.22	20.12	71.85
FCN-8s	83.49	82.45	69.75	72.27	20.06	71.90
FCN-8s (CRF-RNN)	83.54	82.60	69.86	72.35	20.00	71.97

Table II: Accuracy analysis of obtained results using different stages of the trained network with following details: training & testing/validation using OSM-REF data.

Network Architecture	PA	MA	MIU	FWIU	FAR	QR
FCN-32s	89.87	91.32	79.89	82.16	11.36	81.72
FCN-16s	91.35	92.78	82.52	84.54	9.54	84.17
FCN-8s	91.52	92.97	82.81	84.81	9.34	84.45
FCN-8s (CRF-RNN)	92.13	93.84	83.97	85.82	8.61	85.48

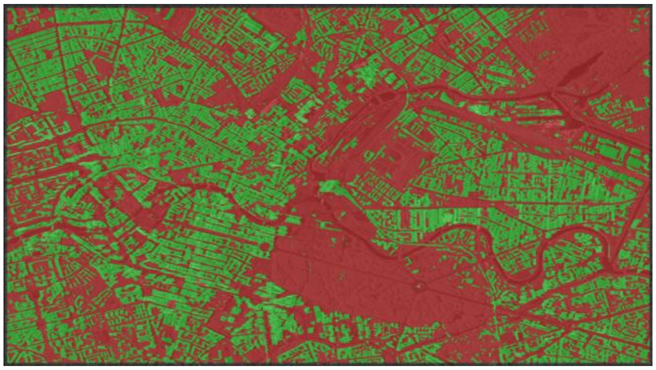

Figure 5: Input SAR image of Berlin city as depicted in Figure 6 with overlay of the semantic segmentation. Results computed using OSM-Ref annotated dataset with FCN-8s with CRF-RNN network.

6. CONCLUSION & OUTLOOK

In this paper, we have presented a deep learning based network architecture that is able to classify buildings from non-buildings in SAR images. Moreover, an automated annotation method able to generate reference building masks for training and testing the classifier has been presented. The method of automated annotation is quite generic and have the potential towards generation of large scale SAR reference datasets. The presented results are expected to further stimulate the research interest in exploiting SAR imagery using deep learning network architectures. In future, we also aim to extend the annotations for other objects available in the OSM dataset e.g., roads, coastlines etc.

REFERENCES

[1] M. Quartulli and M. Datcu, “Stochastic geometrical modeling for built-up area understanding from a single SAR intensity image with meter resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 9, pp. 1996–2003, 2004.

[2] A. Ferro, D. Brunner, and L. Bruzzone, “Automatic Detection and Reconstruction of Building Radar Footprints From Single VHR SAR Images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 935–952, 2013.

[3] L. Zhao, X. Zhou, and G. Kuang, “Building detection from urban SAR image using building characteristics and contextual information,” *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, pp. 1–16, 2013.

[4] J. Zhao, W. Guo, S. Cui, Z. Zhang, and W. Yu, “Convolutional Neural Network for SAR image classification at patch level,” in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 945–948.

[5] Y. Zhou, H. Wang, F. Xu, and Y. Q. Jin, “Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1935–1939, 2016.

[6] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, “Change Detection in Synthetic Aperture Radar Images Based on Deep Neural Networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[7] D. Malmgren-Hansen and M. Nobel-J∅rgensen, “Convolutional neural networks for SAR image segmentation,” in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, pp. 231–236.

[8] X. X. Zhu and R. Bamler, “Very High Resolution Spaceborne SAR Tomography in Urban Environment,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 12, pp. 4296–4308, 2010.

[9] X. X. Zhu and M. Shahzad, “Facade Reconstruction Using Multiview Spaceborne TomoSAR Point Clouds,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3541–3552, Jun. 2014.

[10] X. X. Zhu, Y. Wang, S. Gernhardt, and R. Bamler, “Tomo-GENESIS: DLR’s tomographic SAR processing system,” in *Proceedings of Joint Urban Remote Sensing Event (JURSE)*, Sau Paolo, Brazil, 2013, pp. 159–162.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[12] S. Zheng *et al.*, “Conditional Random Fields as Recurrent Neural Networks,” in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1529–1537.

[13] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[14] J. Li, R. Zhang, and Y. Li, “Multiscale convolutional neural network for the detection of built-up areas in high-resolution SAR images,” in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 910–913.

[15] P. Krähenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” *Adv Neural Inf Process Syst*, vol. 2, no. 3, p. 4, 2011.

[16] G. Baier, X. X. Zhu, M. Lachaise, H. Breit, and R. Bamler, “Nonlocal InSAR Filtering for DEM Generation and Addressing the Staircasing Effect,” in *Proceedings of EUSAR 2016: 11th European Conference on Synthetic Aperture Radar*, 2016, pp. 1–4.