IDENTIFYING CORRESPONDING PATCHES IN SAR AND OPTICAL IMAGERY WITH A CONVOLUTIONAL NEURAL NETWORK

Lichao Mou[1], *Michael Schmitt*[1], *Yuanyuan Wang*[1], *Xiao Xiang Zhu*[1,2]

[1]Signal Processing in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

In this paper, we investigate making use of a convolutional neural network (CNN) to solve the task of identifying corresponding patches in very high resolution (VHR) optical and SAR imagery of complicated urban scenery. By doing so, the binary decision function is learnt directly from automatically generated training data and does not resort to any handcrafted features. First evaluations show great potential for further studies towards a generalized multi-sensor matching procedure.

Index Terms— synthetic aperture radar (SAR), optical imagery, data fusion, deep learning, convolutional neural networks (CNN), image matching

1. INTRODUCTION

The identification of corresponding image patches is used extensively in computer vision and remote sensing-related image analysis, especially in the framework of stereo applications. While quite some established feature-based approaches, specifically designed for the matching of optical images, exist (e.g., the well-known and widely used SIFT approach), to this date the matching of images acquired by different sensors still remains an open challenge. Identifying correspondences of SAR and optical patches is a non-trivial task, as there are a couple of challenges that are caused by two completely different sensing modalities: SAR imagery collects information about the physical properties of the scene and follows a range-based imaging geometry, while optical imagery reflects the chemical characteristics of the scene and follows a perspective imaging geometry. Hence, particularly structures elevated above the ground level, such as buildings in urban areas, show strongly different appearances in both SAR and optical images (cf. Fig. 1).

CNNs trained by backpropagation have recently shown promising performance in large-scale image classification [1]. This gave the beginning to a surge of studies exploiting CNNs for various visual analysis tasks. Meanwhile, CNNs are well known to be very good at learning input-output relations given enough labeled training data.

Fig. 1. Two examples for the different appearance of urban objects in non-rectified VHR SAR and optical data. Left column: TerraSAR-X amplitude image (range direction: top-down), middle and right column: airborne optical imagery with different viewing angles.

As has been shown before, CNNs also provide a powerful means for the matching of homologue image patches [2]. While almost all of the hitherto published work deals with the classical problem of optical-to-optical image matching, recently we have shown that CNNs also allow to learn the identification of corresponding patches in seemingly very different SAR and optical images of complex urban scenes [3]. In order to solve this task, we have proposed a pseudo-siamese architecture with two identical yet completely seperate convolutional streams, whose information is only fused in a final, fully connected decision layer.

2. HOW TO GENERATE TRAINING DATA?

As is well-known, it is necessary to use a large number of training samples to learn the huge number of parameters of a deep CNN. In our case, the first major problem is to get hold of these training data, as even for human experts, the matching of homologue image patches in VHR SAR and optical images of complicated urban scenery is a non-trivial task. In order to deal with this challenge, we utilize an object-space-based matching procedure developed for mapping textures

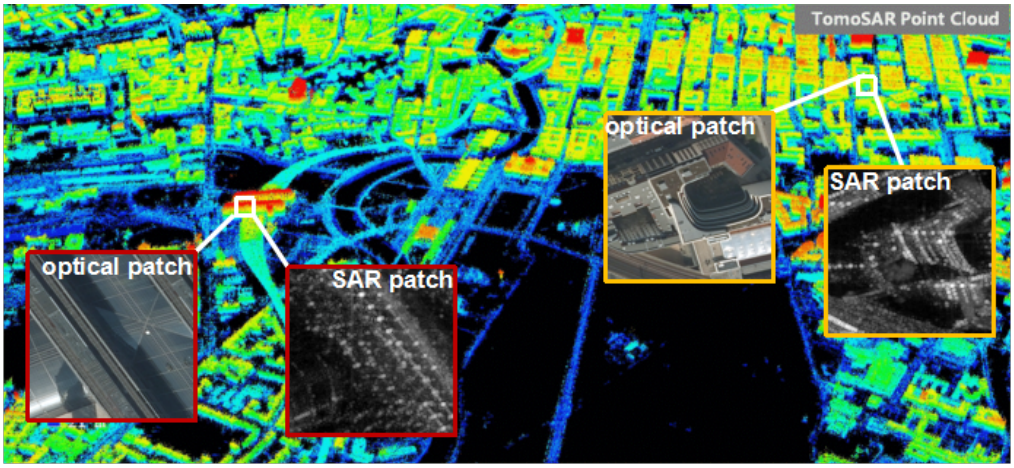

Fig. 2. The used 3D point clouds and the corresponding SAR-optical patches in the 3D space.

from optical images onto 3D point clouds derived from SAR tomography [4].

The core idea of this approach is to match the SAR and the optical images in 3D space in order to deal with the inevitable differences caused by different geometrical distortions. This often would require an accurate digital surface model (DSM) of the area to link homologue image parts via a known object space. In contrast, the approach in [4] creates two separate 3D point clouds one from SAR tomography and one from optical stereo matching, which are then registered in 3D space to form a "SARptical" point cloud, which serves as the necessary representation of the object space. An illustration of the data generation setup can be found in Fig. 2.

To estimate the 3D positions of the individual pixels in the images, an interferometric stack of SAR images is required, as well as at least a pair of optical stereo images. The matching of the two point clouds in 3-D guarantees the matching of the SAR and the optical images. Finally, we can project the SAR image into the geometry of the optical image via the "SARptical" point cloud, and vice versa.

In this work, we made use of a stack of 109 TerraSAR-X high resolution spotlight images of Berlin acquired between 2009 and 2013 with about 1 meter resolution, and of 9 UltraCAM optical images of the same area with 20cm ground spacing. After the 3D point cloud reconstruction, 32,446 pixels were selected from the SAR images and projected into the optical images, yielding 89,502 optical patches. Image patches of 112×112 pixels are centered at a given SAR pixel, and a similarly large patch around the projected position in the optical image is cropped to generate a pair of corresponding SAR-optical patches. Proper corrections, including rotation and adjustment of the pixel spacing, has been applied on the corresponding patches, so that they align with each other at a first approximation. The reason for the different number of patches is that the 9 optical images are acquired at multiple viewing angles, so that one SAR image patch may have a maximum of 9 corresponding optical image patches, depending on the visibility of the SAR pixel from the respective optical point of view. Fig. 1 shows two examples of the extracted corresponding patches, where the left most column is the selected SAR image patch, and the other two columns are the corresponding optical patches, respectively. The SAR and optical patches are shown in their original geometry. As we can see, it is still visually difficult to tell if the patches correspond to each other, due to the complex 3D geometry of the buildings. In addition, the optical patches are slightly different because of the different viewing angle of the camera.

3. HOW TO DESIGN THE NETWORK?

3.1. Network Architecture

Conventional CNNs have shown promising performance in various visual tasks. Yet, such network architectures are single-input and single-output (SISO) systems. Since SAR and optical images can be considered to lie on different manifolds, in theory it is not suitable to handle the comparison task in the focus of this paper. In order to cope with this deficiency, we make use of a network with two separate, yet identical convolutional streams, which process the SAR patch and the optical patch in parallel, and only fuse the resulting information at a later decision stage. Using this architecture, the network is constrained to first learn meaningful representations of the input SAR patch and the optical patch separately, and to combine them on a higher level.

3.2. Detailed Configuration

The exact architecture of the network we train (cf. Fig. 4) is mainly inspired by the philosophy of VGG Nets [1].

In general, it follows two rules: 1) Having the same feature map size and the same number of filters in each convolutional layer of the same block; and 2) increasing the size of

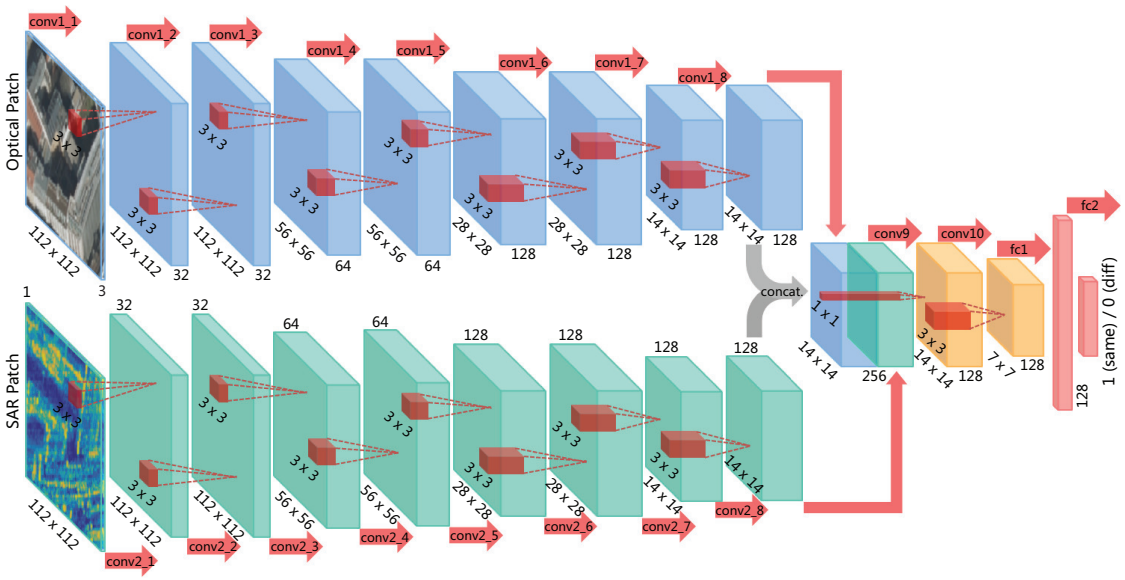

Fig. 3. The architecture of the proposed two-stream CNN for identification of similar patches in SAR and optical imagery.

the features in the deeper layers, roughly doubling after each max-pooling layer. The traits of our network can be summarized as follows:

- The SAR and optical image patches are passed through a stack of convolutional layers, where we make use of 3×3 convolutional filters, rather than using larger ones, such as 5×5 or 7×7. That is because the 3×3 receptive field is the smallest kernel to capture patterns in different directions, such as center, up/down, and left/right. In addition, another prop is the use of small convolutional filters will increase the nonlinearities inside the network and thus make the network more discriminative [1].

- We utilize a convolutional layer with a 1×1 kernels in the fusion stage of the network, which can be regarded as nonlinear transformation of the input channels [5]. The 1×1 convolutional layer is used to reduce the dimensionality by a factor of two, and is capable of modeling weighted combinations of two feature maps produced separately by the SAR and the optical convolution streams at the same spatial location. When implemented as trainable filters in the network, 1×1 convolutional filters are able to learn a proper fusion rule of the two feature maps, which minimizes the final loss function.

- The convolution stride in our network is fixed to 1 pixel; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution, i.e., the padding is 0 for the 1×1 convolutional layer, and is 1 pixel for the 3×3 convolutional layers.

- Spatial pooling is achieved by carrying out seven max-pooling layers, which follow some of the convolutional layers. Max-pooling is performed over 2×2 pixel windows with stride 2.

3.3. Loss Function

Let $\mathbf{X} = \{(\mathbf{x}_1^{sar}, \mathbf{x}_1^{opt}), (\mathbf{x}_2^{sar}, \mathbf{x}_2^{opt}), \cdots, (\mathbf{x}_n^{sar}, \mathbf{x}_n^{opt})\}$ be a set of SAR-optical pairs, where $\mathbf{x}_i^{sar}, \mathbf{x}_i^{opt} \in \mathbb{R}^{D\times D}, \forall i = 1, \cdots, n$, whereas y_i is the 0/1 label for the pair $(\mathbf{x}_i^{sar}, \mathbf{x}_i^{opt})$ (with 0 and 1 denoting a non-matching and a matching pair, respectively). We then seek to minimize the following error

$$E = \frac{1}{n}\sum_{i=1}^{n}((1-y_i)\hat{y}_i^2 + y_i(\max(0, \lambda - \hat{y}_i))^2)\,, \quad (1)$$

where λ is the margin, $\hat{y}_i$ is the output of network, given SAR-opt pair $(\mathbf{x}_i^{sar}, \mathbf{x}_i^{opt})$ and current network parameter settings θ, as follows,

$$\hat{y}_i = f(\mathbf{x}_i^{sar}, \mathbf{x}_i^{opt}; \theta)\,. \quad (2)$$

4. EXPERIMENT

4.1. Training Details

For training the network, we use the Adamax algorithm [6], because it shows faster convergence than standard stochastic gradient descent with momentum. The parameters of Adamax are fixed to $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate $lr = 0.002$ as recommended. In the training, fairly large mini-batches of 128 SAR-optical patch pairs are used. All weight matrices in the network and all bias vectors are initialized from a uniform distribution in the range [-0.1,0.1]. To train

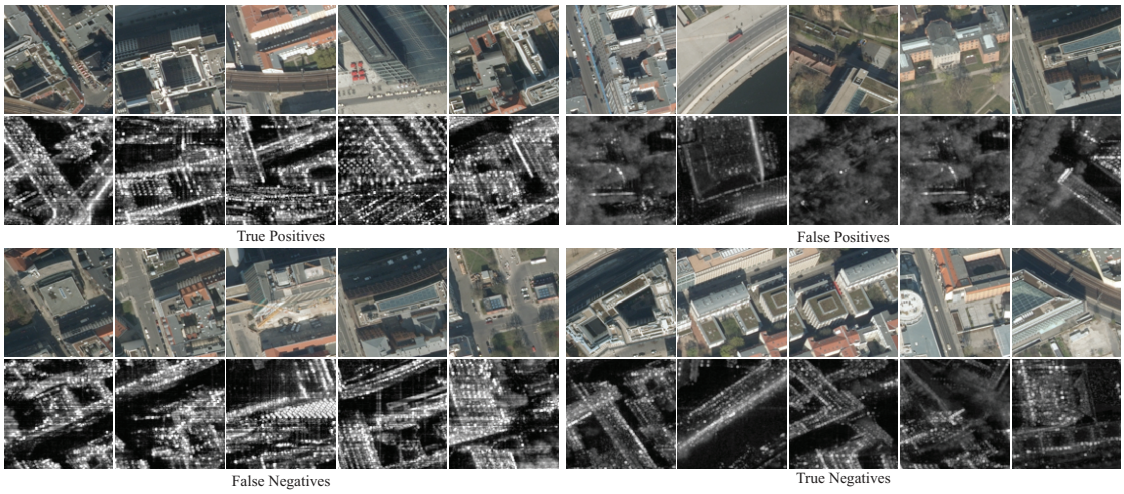

Fig. 4. Randomly selected examples.

the network, we randomly select 10,000 optical patches from the available patch data, and find their corresponding SAR patches to form the positive pairs of the training set. For the same 10,000 optical patches, negative pairs are generated by randomly assigning dissimilar SAR patches to them. Now, we have 20,000 pairs as the training set. Finally, to monitor the training course of network, we generate a validation set by randomly selecting 10% of the patch pairs from the training set.

4.2. Results

For testing purposes, we randomly select another 10,000 optical patches from the patch pool without any overlap between these new 10,000 optical patches and the optical patches in the training set. Repeating the same process as for the training data, i.e. assigning both 10,000 similar and 10,000 dissimilar SAR patches, we eventually create 20,000 test patch pairs.

To quantitatively evaluate the performance of our network, we make use of the widely used evaluation metric FPR95 [2], which stands for the false positive rate at 95% recall, i.e. the lower the FPR95 value, the better. Our network can give an FPR95 of 0.05%. In addition, our network is able to provide an overall accuracy of 97.48% with a false alarm rate of 0.05%. When maintaining 0% false positive rate, the highest overall accuracy of 93.43% can be achieved by the network. In Fig. some randomly selected examples computed by our network are shown.

5. CONCLUSION

In this paper, a CNN-based framework for learning to identify corresponding patches in SAR and optical images in a fully automatic manner has been presented. A first evaluation has shown promising results. Future work will mainly consist in enlarging the data pool so that completely independent training, validation and testing datasets can be formed. In addition, we will extend the network such that it is not only to cast a binary decision of correspondence, but to provide a continuous similarity measure, which will allow to better quantifiy how similar the different image patches actually are.

6. ACKNOWLEDGMENT

This work is supported by the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, www.sipeo.bgu.tum.de) and the German Research Foundation (DFG), grant SCHM 3322/1-1.

7. REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[2] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. of CVPR*, 2015.

[3] L. Mou, M. Schmitt, Y. Wang, and X. Zhu, "A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes," in *Proc. of JURSE*, 2017.

[4] Y. Wang, X. Zhu, B. Zeisl, and M. Pollefeys, "Fusing meter-resolution 4-d insar point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 14–26, 2017.

[5] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv:1312.4400*, 2014.

[6] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.