A CNN for the Identification of Corresponding Patches in SAR and Optical Imagery of Urban Scenes

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)
Arcisstr. 21, 80333 Munich, Germany

[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR)
Münchener Str. 20, 82234 Wessling, Germany

***Abstract*—In this paper we propose a convolutional neural network (CNN), which allows to identify corresponding patches of very high resolution (VHR) optical and SAR imagery of complex urban scenes. Instead of a siamese architecture as conventionally used in CNNs designed for image matching, we resort to a pseudo-siamese configuration with no interconnection between the two streams for SAR and optical imagery. The network is trained with automatically generated training data and does not resort to any hand-crafted features. First evaluations show that the network is able to predict corresponding patches with high accuracy, thus indicating great potential for further development to a generalized multi-sensor matching procedure.**

I. Introduction

The identification of corresponding image patches certainly is a frequently needed task in remote sensing-related image analysis, especially in the framwork 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 [1]), to this date the matching of images acquired by different sensors still remains an open challenge. This particularly holds for a joint exploitation of SAR and optical imagery. In this case, not only a slightly different radiometric appearance, e.g. caused by changing illumination conditions, makes the matching a non-trivial task. Instead, the challenge is caused by two completely different sensing modalities: While optical imagery reflects the chemical characteristics of the scene and follows a perspective imaging geometry, SAR imagery collects information about the physical properties of the scene and follows a range-based imaging geometry. Thus, particularly structures elevated above the ground level, such as buildings in urban areas, show strongly different appearances in both image types (cf. Fig. 1).

In order to deal with the problem of multi-sensor image matching, several sophisticated approaches have been proposed, mostly exploiting the structural content of the images, e.g. through implicit similarity determination [2] or using phase congruency as a generalization of gradient information [3]. However, none of them is reliably able to deal with resolutions in the (sub-)meter domain and with densely built-up urban scenes, which is probably caused by the fact that manually designed descriptors reach their limitations for such highly resolving data, which – in the SAR case – to this day can still only be interpreted by long-time experts.
In contrast, our work aims at learning a multi-sensor similarity function for SAR and optical image patches of state-of-the-art VHR data. In order to prepare for a learnt similarity descriptor, in this paper we deal with the question: Can we automatically learn to identify corresponding image patches in SAR and optical images by making use of a convolutional neural network? This is related to the work of [4], who investigated CNNs for image similarity in the framework of depth map generation for optical images. The major difference to our work is that we focus on the afore-mentioned, distinctly more complicated multi-sensor setup, and therefore make use of a network architecture with two separate, yet identical convolutional streams for processing SAR and optical patches in parallel, instead of a weight-shared siamese network in order to deal with the heterogeneous nature of the input imagery.

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.

II. Network Architecture

A. "SARptical" Convolutional Network

Since SAR and optical images can be considered to lie on different manifolds, it is not advisable to compare them directly by descriptors designed for matching optical patches. Neither suitable are conventional CNNs, which have originally been designed as single-input single-output (SISO) systems, as the matching of SAR and optical image patches with strongly different properties belongs to the class of multiple-input single-output (MISO) systems. Therefore, also SISO networks

978-1-5090-5808-2/17/$31.00

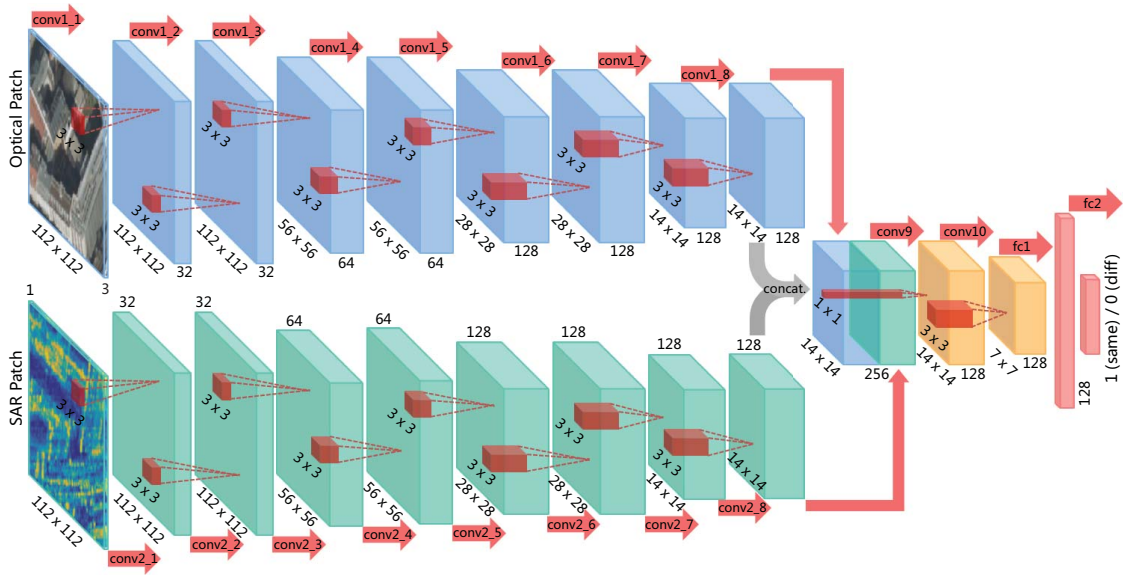

Fig. 2. The architecture of the proposed two-stream CNN for identification of similar patches in SAR and optical imagery.

such as the weight-shared siamese network proposed in [4] are not able to handle the comparison task in the focus of this paper. In order to cope with this deficiency, we propose a MISO 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 pseudo-siamese 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.

The architecture of the proposed network is shown in Fig. 2. It is mainly inspired by the philosophy of the VGG Nets [5]. The SAR and optical image patches are passed through a stack of convolutional layers, where we make use of convolutional filters with a very small receptive field of 3×3, rather than using larger ones, such as 5×5 or 7×7. The reason is that 3×3 convolutional filters are the smallest kernels to capture patterns in different directions, such as center, up/down, and left/right, but still have an advantage: the use of small convolutional filters will increase the nonlinearities inside the network and thus make the network more discriminative. Simonyan and Zisserman [5] have reported that it can be difficult to initialize such a network equipped with small convolutional filters. However, we do not face this problem and simply train the network from scratch.

In addition, it is worth noting that we also utilize a convolutional layer with a 1×1 receptive field in the fusion stage of the network, which can be regarded as nonlinear transformation of the input channels [6]. 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 (cf. Fig. 3). Max-pooling is performed over 2×2 pixel windows with stride 2.

In a nutshell, the convolutional layers in our network – apart from the fusion layer – generally consist of 3×3 filters and follow two rules: 1) The layers with same feature map size have the same number of filters; and 2) the size of the feature maps increases in the deeper layers, roughly doubling after each max-pooling layer (except for the last convolutional stack in each stream), which is meant to preserve the time complexity per layer as far as possible. The last convolutional layer is then followed by two fully connected layers: the first one has 128 channels, while the second performs binary classification and thus contains only 1 channel. All layers in the network are equipped with a rectified linear unit (ReLU) [7] as activation function, except the last fully connected layer, which is activated by a sigmoid function. Fig. 3 shows the schematic diagram of the detailed configuration information of our network.

B. 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 patch 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 dissimilar and a similar pair, respectively). We then seek to minimize the error

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

which penalizes a distance smaller than the margin λ (with $\lambda = 1$ in this work) for corresponding pairs, while for non-corresponding pairs penalization occurs for distances larger than 0.

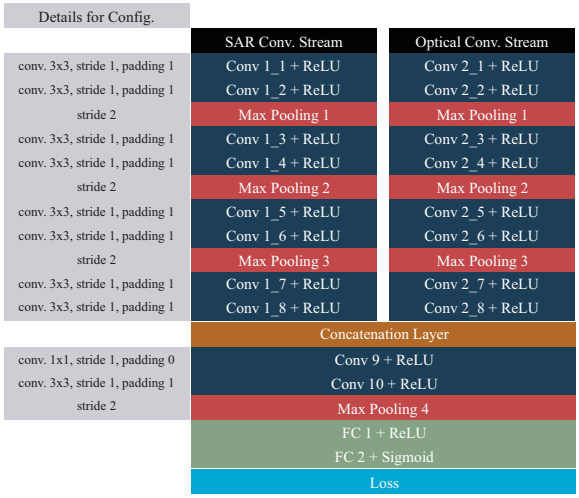

Fig. 3. Schematic diagram of our network configurations.

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

is the output of network, given the SAR-optical patch pair $(\mathbf{x}_i^{sar}, \mathbf{x}_i^{opt})$ and the current network parameter settings θ.

III. Training and Testing

A. Preparation of Similar Image Patches

As well known to the machine learning community, a large amount of training samples is necessary to learn the many parameters of a CNN. For the work presented in this paper, the major problem was to get hold of these training data, as the matching of homologue image patches in VHR SAR and optical images of complex urban scenes is a non-trivial task even for human experts. In order to deal with this challenge, we made use of an object-space-based matching procedure developed for mapping textures from optical images onto 3D point clouds derived from SAR tomography [8]. The core of this algorithm 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. Usually, this 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 [8] 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. The flowchart of the approach can be seen in Fig. 4.

In order to estimate the 3D positions of the individual pixels in the images, the algorithm requires an interferometric stack of SAR images, 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.

B. Data

In this paper, we made use of a stack of 109 TerraSAR-X high resolution spotlight images of Berlin acquired between

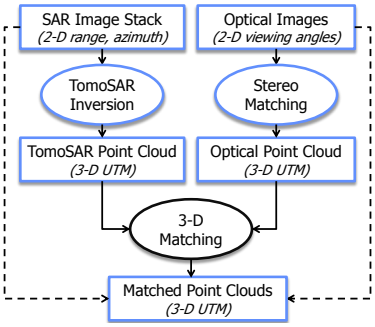

Fig. 4. Flowchart of the SAR optical image matching algorithm. The coordinate system of each dataset in the flowchart is indicated by the italic text in brackets. The dashed lines indicate that the SAR and optical images can be projected to each other through the matched 3-D point cloud.

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 different 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 correspond the patches to each other, due to the complex 3-D geometry of the buildings. In addition, the optical patches are slightly different because of the different viewing angle of the camera.

C. Training Details

For training the network, we use the Adamax algorithm [9], 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 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.

We first investigate the behaviour of our two-stream convolu-

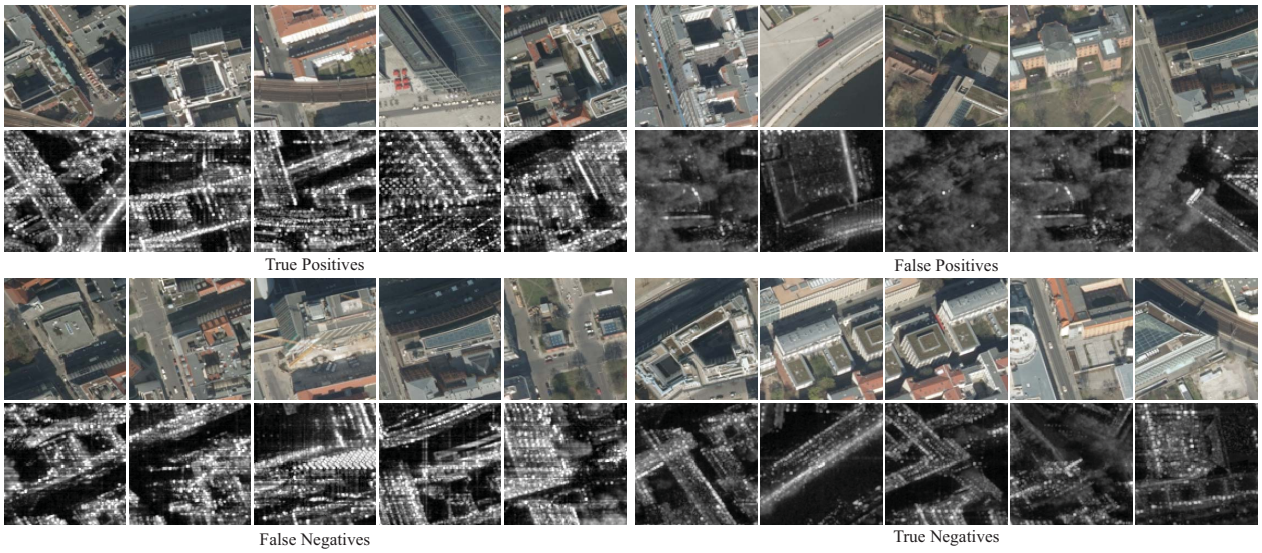

Fig. 5. Randomly selected examples

tional network during the training course, before we present the performance of the network on the actual identification task. The quality of the trained network can be reflected by learning curves. As shown in Fig. 6, our network starts greatly reducing errors on both the training and the validation set during the first few epochs, and finally achieves the error value of 5.98×10^{-7} on training samples and that of 3.57×10^{-7} on validation samples, which means the network can converge to a good solution. Moreover, since we do not observe overfitting in Fig. 6, the trained network can be thought as a good model for the follow-up test stage.

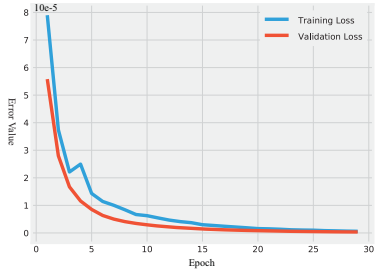

Fig. 6. Learning curves

D. Test 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 [4], 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. 5 some randomly selected examples computed by our network are shown.

IV. Conclusion & Outlook

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 very promising results that will help to pave the way for the future creation of SAR-optical tie point matching procedures exploiting a learnt generalized similarity measure. Future work will mainly comprise the generation of additional training data and test data from a completely different source, e.g. including imagery from different optical and SAR sensors.

References

[1] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.

[2] Y. Keller and A. Averbuch, "Multisensor image registration via implicit similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 794–801, 2006.

[3] Xiaochun Liu, Zhihui Lei, Qifeng Yu, Xiaohu Zhang, Yang Shang, and Wang Hou, "Multi-modal image matching based on local frequency information," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, pp. 1–11, 2013.

[4] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE CVPR*, 2015, pp. 4353–4361.

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

[6] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv:1312.4400*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[8] Yuanyuan Wang, Xiaoxiang Zhu, Bernhard Zeisl, and Marc 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. 1–13, 2016.

[9] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.