REGISTRATION OF HIGH RESOLUTION SAR AND OPTICAL SATELLITE IMAGERY USING FULLY CONVOLUTIONAL NETWORKS

Stefan Hoffmann[1], *Clemens-Alexander Brust*[1], *Maha Shadaydeh*[1], *Joachim Denzler*[1,2]

[1]Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany
[2]Michael Stifel Center Jena, Jena, Germany

ABSTRACT

Multi-modal image registration is a crucial step when fusing images which show different physical/chemical properties of an object. Depending on the compared modalities and the used registration metric, this process exhibits varying reliability. We propose a deep metric based on a fully convolutional neural network (FCN). It is trained from scratch on SAR-optical image pairs to predict whether certain image areas are aligned or not. Tests on the affine registration of SAR and optical images showing suburban areas verify an enormous improvement of the registration accuracy in comparison to registration metrics that are based on mutual information (MI).

***Index Terms*—** Remote Sensing, SAR-optical Image Registration, Fully Convolutional Network

1. INTRODUCTION

Fusing images taken by different sensors leads to an increase in information density and allows spatially-resolved comparison of different physical/chemical properties of a viewed object. For accurate fusion, it is crucial to achieve spatial alignment between those images. Multi-modal image registration is therefore a challenging task in a wide variety of topics, including medical science, computer vision and remote sensing.

While SAR images can be taken independent of sunlight or cloud coverage and with higher spatial resolution than optical remote sensing images, SAR images are harder to interpret. Fusion of SAR and optical images enables geolocalization refinement of optical remote sensing images [1], as well as giving new possibilities of learning SAR image interpretation [2].

As SAR and optical sensors measure with different wavelengths and are using different viewing angles, the resulting images look very dissimilar w.r.t. shadowing effects and spatial intensity distribution. Consequently registration metrics based on information theory, e.g. Kullback-Leibler divergence or MI, may not be reliable when registering SAR and optical images. We propose a deep metric which is able to detect misaligned areas of a SAR-optical image pair in a spatially-resolved manner. To this end, we convert the task of evaluating the alignment of two images into a binary segmentation task, estimating whether or not the corresponding areas of SAR and optical images are aligned. The deep metric is tested on affine transformations including translations, rotations and scaling.

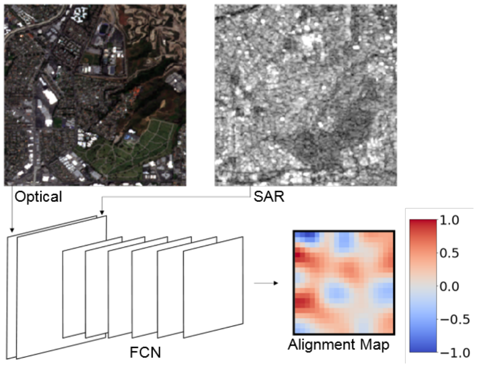

Fig. 1: Scheme of the proposed deep metric. The concatenated input of the FCN consists of a fixed optical image and a SAR image which is transformed during registration. The scalar deep metric value is calculated by the averaging of the spatially-resolved output of the FCN that predicts the alignment of its input channels.

Due to the fact that the network is fully based on convolutions [3], it is possible to use it regardless of the size of the compared images, with the constraint that the compared images must have the same size. As shown in Figure 1, the FCN provides a two-dimensional output, in which cells contain a spatially-resolved prediction of the alignment of their receptive field. A scalar deep metric value for an image pair is calculated through average pooling.

2. RELATED WORK

Although there a number of publications studying deep learning methods for segmentation, multi-modal image fusion, and multi-temporal change detection of remote sensing images [4, 5, 6, 7], deep learning is unpopular in SAR image processing, mainly because of the limited amount of available data [8].

There are two public data sets containing aligned SAR and optical images: the SARpital data set proposed in [9], which was created through three-dimensional reconstruction of the measured areas [10], and the SEN1-2 data set [11], which we use for training and registration.

The FCN-based concept of deep image registration metrics, which are usable for non-rigid registration, has been proposed in [12]. There, a three-dimensional FCN is trained on spin-lattice- and spin-spin-relaxation MRI scans, predicting whether or not corresponding scan volumes are aligned. We adapt this concept and adjust it to remote sensing data, using a patch-wise training method as proposed in [13].

The FCN presented in [12] is a progression of [14], where CNNs are used to predict whether two images show the same object. As shown in [15], the CNNs that were introduced in [14] can be adapted to predict whether SAR and optical images show corresponding objects. To the best of our knowledge, our approach is the first using an FCN as a deep metric for SAR-optical image registration.

Instead of training a deep metric, [1] proposes directly predicting the displacement between SAR and optical images. Unlike our approach, it is only capable of image registration w.r.t. shifts.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1. Data selection and pre-processing

We select 7,100 image pairs for training, as well as 2,600 for validation and 200 for image registration tests. We use a subset of the SEN1-2 'Spring' data set which is structured according to the regions shown in the SAR and optical images. The three subsets show different regions of the Earth's surface to guarantee data independence.

Optical images are transformed into grayscale images. A Lee filter with kernel size 5 is used on SAR images to reduce speckle noise [16]. Images of both types are normalized w.r.t. mean and variance.

3.2. Architecture and Training

The general structure of the deep metric is shown in Table 1. While linear activation is used on convolutional layer 6, the layers 1 to 5 contain Leaky-ReLU activation functions with $\alpha = 0.1$. SAR and optical images are concatenated to create an input with dimensions $x \times y \times 2$, whereby the minimum size for each dimension x and y is 37.

The FCN is trained patch-wise on paired SAR and optical patches of size $37 \times 37 \times 2$. We randomly select image pairs from a set of perfectly aligned multi-modal images. For each pair we crop a 37×37 area out of each image and concatenate them to a multi-modal patch. We constrain that 50% of the multi-modal patches have to show exactly the same area in each modality channel. Those are labeled with 1. The other 50% show areas which are displaced by a random number between 1 to 10 pixels along x and y axis and were labeled with -1.

Table 1: FCN architecture used for training with concatenated input images of size $x = y = 37$.

Name	Kernel / Stride	Output size / Channels
Conv. 1	5,5 / 2	37,37 / 512
Conv. 2	5,5 / 2	17,17 / 512
Conv. 3	3,3 / 2	7,7 / 512
Conv. 4	3,3 / 1	3,3 / 512
Conv. 5	1,1 / 1	1,1 / 512
Conv. 6	1,1 / 1	1,1 / 1

The training is run for 4 million iterations with batchsize 128, using SGD with a momentum of 0.9, a learning rate of 10^{-2} and a weight decay of 10^{-4}. We use a hinge loss, where l and $\hat{l}$ are the real and estimated patch labels and n is the batch size, defined as follows:

$$HL(l, \hat{l}) = \frac{\sum_i^n \max(0, 1 - l_i \cdot \hat{l}_i)}{n} \tag{1}$$

To further characterize the training, the percentage of correctly assigned samples per batch is calculated. Correct assignment is defined by whether or not a negative label for misaligned or a positive label for aligned multi-modal patches is predicted correctly. After training, the network achieves a hinge loss of 0.45, assigning 78% of the patches correctly.

In contrast to results presented in [14], the usage of pseudo-siamese layers does not improve the network performance, as splitting layers leads to an oscillation of the loss function during training.

3.3. Image Registration

We randomly select 200 aligned multi-modal image pairs to investigate the registration performance of the deep metric. The set of test pairs is created by transforming the SAR images. Transformation parameters are taken randomly out of an interval of -6 to 6 pixels for translation, degrees for rotation and percent for scaling. Misalignment by translation is only applied along the x-axis.

We implement the actual registration task as a grid search, searching for the parameters needed to transform the SAR images back to their original state, using the optical images as a fixed reference. The grid has a distance between the grid points of 1 pixel for translation, 1 degree for rotation and 2 percent for scaling. 3,375 parameter constellations are tested, corresponding to 15 different values for each transformation parameter.

For comparison, identical tests are performed on our deep metric, the MI and the normalized mutual information (NMI) with 64 bins [17]. NMI is implemented as follows, where

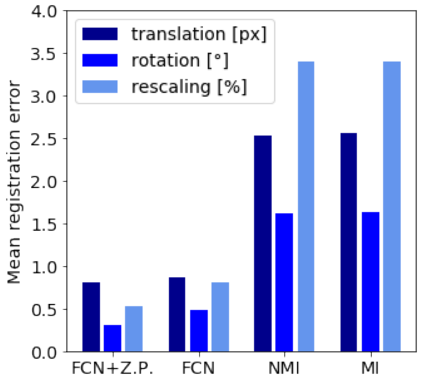

Fig. 2: Mean registration error $\Delta\hat{\theta}$ for translation, rotation, and rescaling using the deep metric with and without Zero Padding, compared to NMI and MI.

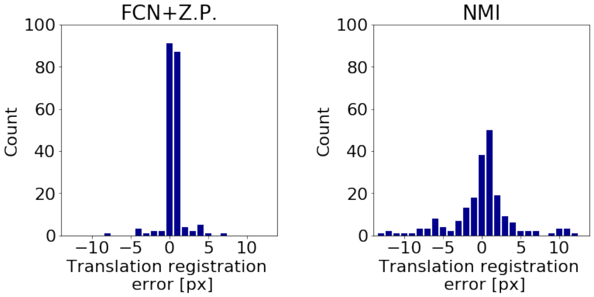

Fig. 3: Distribution of the translation registration error of the deep metric and the NMI.

$H(X)$ and $H(Y)$ is the entropy function of two images X and Y, while $H(X,Y)$ is their joint entropy:

$$\mathrm{NMI}(X,Y) = \frac{H(X)+H(Y)}{H(X,Y)} \tag{2}$$

Metrics are calculated on a patch containing central image areas with size $x = y = 157$ after each step of the grid search. In addition to using the described deep metric, we test an approach where zero padding of (18,18) ia applied to the patch before calculating the metric. This enlarges the number of receptive fields used by the FCN, without adding new information. We further use average pooling over the whole two-dimensional output to create a scalar deep metric value, after skipping values larger than 1 or smaller than -1, replacing them with 1 or -1.

We define the constellation of back-transformation parameters which lead to the highest metric value as the predicted registration parameter set $\hat{\theta}$. As the true registration parameters θ^* are known during testing, performance is evaluated by the mean registration error $\Delta\hat{\theta}$, which is calculated as follows, where N is the number of registered pairs:

Table 2: Comparison of the registration accuracy of different metrics. The registration accuracy is defined by the percentage of image pairs for which a registration error below certain threshold is achieved.

Methods	Registration Accuracy			
	≤ 1 px	≤ 2 px	$\leq 1°$	$\leq 2\%$
MI	53.0%	69.0%	68.5%	69.0%
NMI	54.0%	69.5%	68.5%	69.5%
FCN	89.5 %	91.0%	91.0 %	92.0%
FCN+Z.P.	90.0 %	93.0%	96.5%	96.0%

$$\Delta\hat{\theta} = \frac{\sum_i^N |\theta^* - \hat{\theta}|}{N} \tag{3}$$

As shown in Figure 2, the deep metric outperforms MI and NMI. Individual zero padding before calculating the deep metric further increases its registration accuracy w.r.t. predicting rotation and rescaling parameters. With zero padding the deep metric leads to a reduction of $\Delta\hat{\theta}$ by a factor of c. 3 for translation parameters and a reduction by factor c. 6 for rotation and rescaling in comparison to the NMI.

As demonstrated in Figure 3, the deep metric predicts mostly small misalignment, while the results of MI and NMI are widely spread. The results shown in Table 2 confirm that our method is achieving registration errors of at the most 1 pixel, 1 degree or 2 percent in about 90% of cases.

We also test another method of creating training patches. Instead of only misaligning 50% of the training patches labeled with -1 by random displacement, we also use random rotation and rescaling. This leads to a worse registration performance. Further, applying only displacement in 37.5% and displacement, rotation and rescaling in 12.5% of the cases leads, in comparison, to a decrease of the registration performance. Correctly predicting translation parameters is still the biggest weakness of our approach. Rotation and scaling parameters can still be predicted with relatively high accuracy by an FCN which is only trained on aligned or displaced multi-modal patches. In consequence, we recommend skipping misalignment through rotation and rescaling during training.

4. CONCLUSION

In this paper we proposed a deep metric which is capable of predicting the alignment of SAR and optical image areas in a spatially-resolved manner. By using average pooling we created a scalar metric value which describes the alignment of a SAR-optical image pair.

We tested our method on affine registration of remote sensing data of suburban areas. The registration was implemented as a grid search, using translation, rotation and rescaling. For comparison, equivalent tests were done using MI and NMI as registration metric. The results show that our method is outperforming MI and NMI by far, as it is more robust and more accurate.

As our network is fully based on convolutions, it is usable regardless of the size of a multi-modal image pair. Furthermore, although only affine registration was investigated, the deep metric should be capable of non-rigid image registration due to its spatially-resolved functionality.

Only a small subset of SEN1-2 was used for training and validation of the FCN, as we focused on suburban areas. Training our network on a larger subset also including unsettled terrain could further improve the applicability of our approach.

Acknowledgments

The study is supported by the Carl Zeiss Foundation within the framework of the project: A Virtual Workshop for Digitization in Science.

5. REFERENCES

[1] N. Merkle, W. Luo, S. Auer, R. Mller, and R. Urtasun. Exploiting deep matching and sar data for the geo-localization accuracy improvement of optical satellite images. *Remote Sensing*, 9(6):586, 2017.

[2] M. Schmitt, L. Hughes, M. Körner, and X. Zhu. Colorizing sentinel-1 sar images using a variational autoencoder conditioned on sentinel-2 imagery. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 422:1045–1051, 2018.

[3] 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*, pages 3431–3440, 2015.

[4] Li. Zhang, Le. Zhang, and B. Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

[5] D. Marcos, R. Hamid, and D. Tuia. Geospatial correspondences for multimodal registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5091–5100, 2016.

[6] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172, 2018.

[7] M. Vakalopoulou, C. Platias, M. Papadomanolaki, N. Paragios, and K. Karantzalos. Simultaneous registration, segmentation and change detection from multisensor, multitemporal satellite image pairs. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.

[8] D. Marmanis, W. Yao, F. Adam, M. Datcu, P. Reinartz, K. Schindler, J. Wegner, and U. Stilla. Artificial generation of big data for improving image classification: a generative adversarial network approach on sar data. *arXiv preprint arXiv:1711.02010*, 2017.

[9] Y. Wang and X. Zhu. The sarptical dataset for joint analysis of sar and optical image in dense urban area. *arXiv preprint arXiv:1801.07532*, 2018.

[10] H. Bagheri, M. Schmitt, P. dAngelo, and X. Zhu. A framework for sar-optical stereogrammetry over urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:389–408, 2018.

[11] M. Schmitt, L. Hughes, and X. Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:1807.01569*, 2018.

[12] M. Simonovsky, B. Gutirrez-Becker, D. Mateus, N. Navab, and N. Komodakis. A deep metric for multimodal registration. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, pages 10–18. Springer International Publishing, 2016.

[13] Clemens-Alexander Brust, Sven Sickert, Marcel Simon, Erik Rodner, and Joachim Denzler. Efficient convolutional patch networks for scene understanding. In *CVPR Workshop on Scene Understanding (CVPR-WS)*, 2015.

[14] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015.

[15] L. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. Zhu. Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn. *IEEE Geoscience and Remote Sensing Letters*, 15(5):784–788, 2018.

[16] J. Lee. Speckle analysis and smoothing of synthetic aperture radar images. *Computer graphics and image processing*, 17(1):24–32, 1981.

[17] R. Perko, H. Raggam, K. Gutjahr, and M. Schardt. Using worldwide available TerraSAR-X data to calibrate the geo-location accuracy of optical sensors. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 2551–2554, 2011.