COMPARATIVE EVALUATION OF SIGNAL-BASED AND DESCRIPTOR-BASED SIMILARITY MEASURES FOR SAR-OPTICAL IMAGE MATCHING

Chunping Qiu[1], *Michael Schmitt*[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

This paper compares different similarity measures for the matching of very-high-resolution SAR and optical images over urban areas. It is meant to provide guidance about the performance of both signal-based and descriptor-based similarity measures in the context of this non-trivial case of multi-sensor correspondence matching. Using an automatically generated training dataset, thresholds for the distinction between correct matches and wrong matches are determined. It is shown that descriptor-based similarity measures outperform signal-based similarity measures significantly.

Index Terms— synthetic aperture radar (SAR), optical image, remote sensing, data fusion, image matching, similarity measures

1. INTRODUCTION

Automatic matching of remote sensing images acquired by optical cameras to images acquired by synthetic aperture radar systems has drawn the attention of scientists for many years, e.g. in the context of image registration or multi-sensor stereogrammetry. This is caused by the fact that determining homologous image regions for very-high-resolution SAR and optical images in general is a highly non-trivial case of multi-sensor alignment [1, 2], in particular, when highly complex urban areas are the focus of interest.

Over the years, many different approaches for SAR-optical image matching were proposed. While some investigations are carried out regarding the applicability of conventional similarity measures (e.g. [3, 4]), most of them rely on more or less complicated pipelines which go beyond simple similarity determination for potential tie points. For example, Lehereau et al. [5] estimate the translations between a SAR and an optical image by exploiting the Fourier-Mellin invariant calculated from line and edge images, respectively. Hellwich et al. [6] integrate matching with geocoding in order to robustify the results of classical tie point matching based on SIFT and SURF features. Incorporating prior knowledge in the form of previously extracted roundabouts and junctions, Palubinskas & Reinartz [7] employ template-based matching for identification of sparsely distributed, yet robust tie points.

In contrast to these full-fledged pipelines, which comprise a number of different processing steps, we intend to focus purely on similarity measures that can be used to identify tie points corresponding to each other across both image domains. As a framework for this investigation, we rely on the matching procedure for SAR-optical stereogrammetry proposed in [8].

2. MEASURING THE SIMILARITY BETWEEN IMAGE PATCHES

In the context of this paper, we follow the generic approach of Inglada & Giros [9], who define the similarity measure between two images I and J as a strictly positive scalar function

$$S_c(I, J) = f(I, J; c), \tag{1}$$

where c is a to-be-defined similarity criterion. S_c has the maximum when the two images are identical according to the similarity criterion. In the framework of this paper, we extend this definition by allowing negative values so that similarity measure such as the correlation coefficient whose value range by definition is $[-1; +1]$, can be considered as a similarity measure as well.

Distinguished by the similarity criterion, there are two basic categories of similarity measures: signal-based similarity measures, and descriptor-based similarity measures. Some exemplary similarity measures of both categories are described in the following.

2.1. Signal-based Similarity Measures

Signal-based similarity measures are calculated based on the original or pre-processed signals, i.e. gray values of pixels in the image processing case. In this paper, we investigate two widely used measures:

- *Normalized Cross-Correlation (NCC)*
  The normalized cross-correlation coefficient

$$\rho(x, y) = \frac{1}{N-1} \sum_{x,y} \frac{\left(I(x, y) - \bar{I}\right)\left(J(x, y) - \bar{J}\right)}{\sigma_I \sigma_J} \tag{2}$$

correlates two image patches I and J, where N is the number of the pixels in the image patch, while implicitly normalizing them to reduce the effects of changing image brightness.

- *Mutual Information (MI)*
  Mutual information is defined as the function of the joint entropy $H(I, J)$ and the marginal entropies $H(I), H(J)$ of two images I, J. We employ its normalized version in this paper [10].

2.2. Descriptor-based Similarity Measures

Image descriptors are a well-established means to describe images on a global as well as a local scale. In the context of image matching, usually local descriptors are extracted around previously detected key points. Subsequently, the resulting feature vectors are compared using a suitable distance metric. In the scope of this paper, we resort to the negative L_2-norm as similarity metric. We chose the following descriptors in this investigation:

- *Histogram of Oriented Gradients (HOG)*
  The HOG descriptor was first proposed in 1986 [11] in the context of object detection. Its principle is to count occurrences of gradient orientation on a dense grid of uniformly spaced image cells, using overlapping local contrast normalization for improved accuracy.
- *Scale-Invariant Feature Transform (SIFT)*
  SIFT [12] is the most prominent example of a local feature descriptor that has found wide application in the fields of computer vision and optical image analysis for more than a decade. The SIFT feature vector usually contains 128 elements depicting the normalized values of previously computed orientation histograms – an analogy to HOG. In its original implementation, SIFT combines both feature point detection and descriptor extraction, so that the feature vector corresponds to a specific scale and orientation assigned to the detected key point. In this paper, we calculate the descriptor for a fixed scale and orientation of 10 and zero respectively.
- *Histogram of Orientated Phase Congruency (HOPC)*
  HOPC [13] is a relatively new local image descriptor that is also based on the analysis of oriented histograms, although the descriptor vector here is calculated from phase congruency [14] instead of gradient information. That makes it supposedly well-suited to the case of multi-sensor image analysis.

3. SIMILARITY THRESHOLD DETERMINATION

In order to decide between a correct and an incorrect match, usually a threshold is applied to the calculated similarity value. We determine individual thresholds for the similarity measures described above by analyzing a training dataset originally designed for learning a convolutional neural network that is able to identify corresponding image patches in SAR and optical very-high-resolution images of urban scenes. Details about this "SARptical" dataset can be found in [15] and [16]. In short, we employ 8840 correctly matched patch-pairs, where each pair consists of both a SAR image patch and an optical image patch pre-processed so that they are approximately aligned regarding orientation and pixel spacing. In addition, we created 8840 wrongly matched patch-pairs by random assignment. The resulting histograms of the similarity values corresponding to correct and incorrect matches are displayed in Fig. 1.

The calculation of a proper threshold can be cast in the framework of detection theory: A decision between H_0 (two image patches match) and H_1 (two image patches don't match) has to be made. One way to deal with this problem is to analyze the likelihood ratio

$$\Lambda(x) = \frac{f_x(x|H_0)}{f_x(x|H_1)}, \tag{3}$$

which reduces the problem of threshold determination to the question of how to balance the probabilities of: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Since TP and TN probabilities are always required to be as high as possible, a trade-off between FP and FN probabilities has to be found, which usually depends on the goal one has in mind. In the case of multi-sensor image matching, it is usually much worse to detect a match that is not correct than to miss a correct match, because wrong matches will always negatively affect the final result. Therefore, we seek to minimize the probability for FPs, while we don't care as much about the FN probability.

Based on these considerations, the threshold is determined based on the Neyman-Pearson criterion, which is based on the rationale to maximize the detection probability given a maximum allowable false alarm rate $P_{F,max}$. That means two patches are considered as correct matches, i.e. decision is made in favor of H_0 when

$$\Lambda(x) > \lambda \tag{4}$$

where λ is chosen so that $P_F = P_{F,max}$. With the FP set to 5%, the thresholds calculated using the NP criterion are depicted as blue bars in Fig. 1 and listed in Tab. 1 where the TP, TN, FP and FN rates calculated from the patch-pairs of the training dataset are also shown. Note that the FP rates vary and are not fixed at $P_{F,max} = 5\%$, as the values in Tab. 1 were calculated from the original similarity values, while the threshold was determined based on the fitted Gaussian distributions.

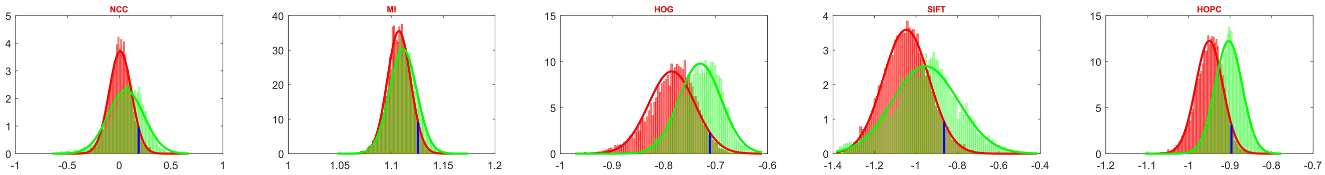


Fig. 1. Histograms of the similarity values calculated for the correct matches (green) and the incorrect matches (red) of the "SARptical" training data. The red curves and the green curves are the fitted Gaussian distributions corresponding to correct and incorrect matches, respectively. The thresholds calculated based on the Neyman-Pearson Criterion are depicted by the blue lines.

Table 1. NP-based threshold and the probability of TP/TN/FP/FN.

	NCC	MI	HOG	SIFT	HOPC
λ	*0.19*	*1.13*	*-0.71*	*-0.86*	*-0.90*
TP	12.2%	6.5%	17.6%	13.7%	21.1%
TN	47.5%	47.8%	48.4%	47.5%	48.2%
FP	2.5%	2.2%	1.6%	2.5%	1.8%
FN	37.8%	43.5%	32.4%	36.3%	28.9%

4. APPLICATION TO SAR-OPTICAL STEREOGRAMMETRY

For a more detailed and application-oriented evaluation of similarity measures for SAR-optical image matching, we use these similarity measures in the SAR-optical stereogrammetry framework proposed in [8]. Similar to the experiments described in [8], two very high resolution spaceborne datasets acquired by TerraSAR-X and Worldview-2, respectively, over the city of Munich, Germany, are used. The height interval for constructing the IMBLS window was set to $[h_0 - 5m, h_0 + 20m]$, where h_0 was taken from the SRTM DEM of the study area. A ±1 pixel pre-defined buffer in the row direction was used to form the final IMBLS search window. The patch size used for similarity calculation was set to 111 × 111 pixels.

For quantitative evaluation of the stereogrammetic 3D reconstruction result, point distances to a dense LiDAR reference point cloud are analyzed. In order to ensure as unbiased results as possible, we calculate the median of the euclidian distances between the stereogrammetrically reconstructed 3D points and a least square plane fitted through its 10 nearest neighbors in the LiDAR dataset.

The evaluation results of the five similarity measures investigated in this paper, both using 80% of the points with highest similarity and using the NP-based threshold, are shown in Fig. 2 and Fig. 3, respectively.

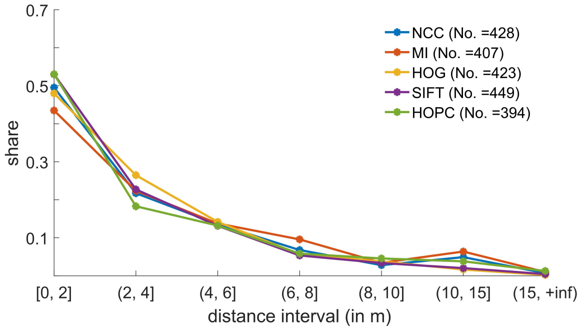


Fig. 2. Distance distribution of the reconstructed points corresponding to 80%-quantile of most similar points.

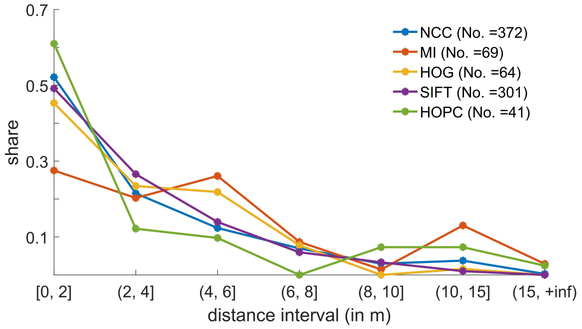
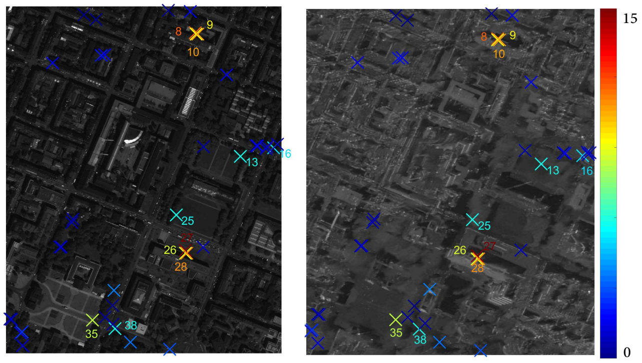


Fig. 3. Distance distribution of the reconstructed points corresponding to the NP threshold with $P_{F,max}$ set to 5%).

5. DISCUSSION

From the analysis results shown in this paper, it becomes obvious that especially mutual information is not a suitable measure for SAR-optical image matching, as it does not provide a sufficiently high discriminative power – the similarity value distributions of both the correct and the incorrect matches are almost identical, thus leading to an accuracy of only 54.3%, and an FN rate of 43.5%. In contrast, NCC, HOG and SIFT provide accuracies of 59.7%, 66.0%, and 61.2%, respectively, with reduced FN rates of around 35%. The seemingly most discriminative similarity measure regarding the theoretical detection rates, however, is HOPC, which shows an accuracy of almost 70% at an FN rate of less than 30%. This is also confirmed by the stereogrammetry results depicted in Figs. 2 and 3, where MI gives the worst results, independent of whether the 80%-quantile of the most similar points is used for 3D-reconstruction or the NP-based threshold is applied. Similarly, HOPC shows the best performance in both cases, with more than 50% (80%-quantile) and 60% (NP-threshold) of all reconstructed points lying in a ±2 m accuracy interval. The other similarity measures lie in between, with NCC and SIFT performing slightly better than HOG. Interestingly, MI, HOG and HOPC reduce the number of matched points significantly, while both NCC and SIFT keep a much higher share of point-pairs. Given their still acceptable accuracies and FN rates, this indicates that also SIFT and NCC are useful similarity measures, albeit not as robust as HOPC. However, this will probably not hold for NCC in case the patches are not pre-processed but dissimilar with regard to scale and orientation.

Besides giving us an impression about the performance of the individual similarity measures, these results also illustrate the benefit of applying a proper threshold to exclude dissimilar patches from further processing. However, it is obvious that the NP threshold, which was trained on an independent dataset, does not lead to quasi-perfect results. While a lack of domain adaptation might be part of the explanation, another reason is the nature of the key points processed in the experiments: As Fig. 4 shows, the worst points (numbered in Fig. 4) lie in areas containing trees, which appear blurred in the despeckled SAR image, or in the surroundings of larger buildings, where mismatches can occur due to layover and shadowing. This indicates that there will always be the need for additional post-processing, e.g. employing suitable regularization techniques or support by prior knowledge about the semantic contents of the scene.



(a) Optical image.

(b) SAR image.

Fig. 4. Stereogrammetrically reconstructed points using HOPC-based similarity and the NP-based threshold. The points are colorized by the distance from the LiDAR reference (in meters).

6. CONCLUSION AND OUTLOOK

In this paper, we discussed several signal-based and descriptor-based similarity measures for the identification of homologous patches in SAR and optical imagery. We came to the conclusion that descriptor-based measures outperform signal-based measures, whereas mutual information overall performed worst, while the novel histogram-of-oriented-phase-congruency descriptor performed best. Still, none of these handcrafted similarity measures provides the perfect solution to the SAR-optical similarity determination problem, which motivates investigations towards the learning of a suitable similarity measure from sufficient annotated training data.

7. ACKNOWLEDGEMENTS

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.

8. REFERENCES

[1] M. Schmitt and X. Zhu, "Data fusion and remote sensing: An ever-growing relationship," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 4, pp. 6–23, 2016.

[2] M. Schmitt and X. Zhu, "On the challenges in stereogrammetric fusion of SAR and optical imagery for urban areas," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 41, no. B7, pp. 719–722, 2016.

[3] S. Suri and P. Reinartz, "Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 939–949, 2010.

[4] P. Schwind, S. Suri, P. Reinartz, and A. Siebert, "Applicability of the SIFT operator to geometric SAR image registration," *Int. J. Remote Sens.*, vol. 31, no. 8, pp. 1959–1980, 2010.

[5] G. Lehureau, F. Tupin, C. Tison, G. Oller, and D. Petit, "Registration of metric resolution sar and optical images in urban areas," in *Proc. EUSAR*, 2008, pp. 1–4.

[6] O. Hellwich, C. Wefelscheid, J. Lukaszewicz, R. Hänsch, M. A. Siddique, and A. Stanski, "Integrated matching and geocoding of SAR and optical satellite images," in *Proc. of PRIA*, 2013, pp. 798–807.

[7] G. Palubinskas and P. Reinartz, "Template based matching of optical and SAR imagery," in *Proc. JURSE*, 2015, pp. 1–4.

[8] C. Qiu, M. Schmitt, and X. Zhu, "A tie point matching strategy for very high resolution SAR-optical stereogrammetry over urban areas," in *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2017, in press.

[9] J. Inglada and A. Giros, "On the possibility of automatic multisensor image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2104–2120, 2004.

[10] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recogn.*, vol. 32, no. 1, pp. 71–86, 1999.

[11] R. K. McConnell, "Method of and apparatus for pattern recognition," 1986, US Patent 4,567,610.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, 2017.

[14] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychological research*, vol. 64, no. 2, pp. 136–148, 2000.

[15] 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. JURSE*, 2017.

[16] Yuanyuan Wang, Xiao Xiang Zhu, Bernhard Zeisl, and Marc Pollefeys, "Fusing meter-resolution 4-d insar point clouds and optical images for semantic urban infrastructure monitoring," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 14–26, 2017.