Object-related alignment of heterogeneous image data in remote sensing

Stefan Auer and Peter Reinartz
Remote Sensing Technology Institute
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
Email: stefan.auer@dlr.de, peter.reinartz@dlr.de

Michael Schmitt
Signal Processing in Earth Observation
Technical University of Munich
Arcisstr. 21, 80333 Munich, Germany
E-mail: m.schmitt@tum.de

***Abstract*—The fusion of heterogeneous image data, in particular optical images and synthetic aperture radar (SAR) images, is highly worthwhile in the context of remote sensing tasks as it allows to exploit complementary information – such as spectral and distance measurements or different observation perspectives – of the two data sources while diminishing their individual weaknesses (e.g. cloud cover, difficulty of image interpretation, limited sensor revisit). However, relating the heterogeneous data on the signal level requires a data alignment step, which cannot be realized without auxiliary knowledge. This paper addresses and discusses this fundamental fusion problem in remote sensing in the context of a framework named SimGeoI, which solves the multi-sensor alignment task based on geometric knowledge from existing digital surface models. Sections of optical and SAR images are related to individual objects using interpretation layers generated with ray tracing techniques. Results of SimGeoI are presented for a test site in London in order to motivate an object-related fusion of remote sensing images.**

I. Introduction

Any data fusion endeavor needs to solve the tasks of data alignment and data/object correlation before the actual fusion step, which is particularly challenging when heterogeneous remote sensing data are involved [1]. One prominent example in this context is the combination of synthetic aperture radar (SAR) and optical satellite imagery [2]. This is mainly caused by the strongly different imaging modalities of these two sensor principles: While optical images basically collect angular measurements and information about the chemical characteristics of the observed environment, radar images result from range measurements based on emitted signals and the observation of physical information of the illuminated surfaces (e.g. roughness or moisture). Thus, scenes imaged by both sensors typically appear vastly different so that homogeneous areas can only be matched under favorable circumstances [3], [4]. One approach for SAR-optical multi-sensor alignment is based on the introduction of prior knowledge about the three-dimensional structure of the scene, which is used to connect the different images supported by a simulation framework [5]. In this paper, we will describe this simulation-based alignment approach, named SimGeoI, and demonstrate how it can be used to relate and analyze multi-sensor remote sensing images in an object-based manner. As main contribution, the context of the proposed approach and results of a case study in London are used to address and discuss the difficulty of aligning the multi-modal remote sensing data.

II. SimGeoI for Object-Based Multi-Sensor Remote Sensing Image Alignment

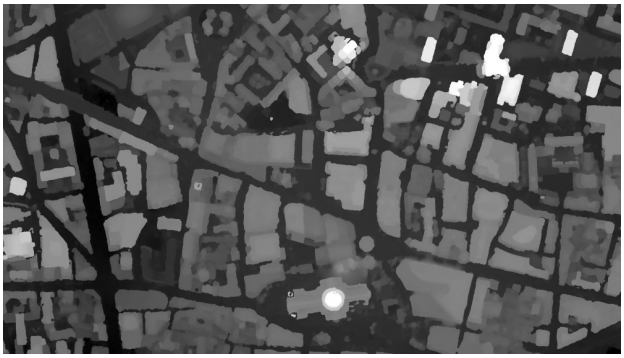
The basic challenge of fusing optical and SAR images becomes apparent in Fig. 1, which shows an optical and a SAR image for an urban scene in London. It is obvious that both image types exhibit completely different radiometric appearances, which is caused by the sensors collecting information from different fields of the electromagnetic spectrum. Looking closely to the image details reveals that furthermore severe image differences are caused by the different imaging geometries: Structural details are mapped differently from 3D to the two 2D image spaces and hence not found at the same image positions. Thus, image content with equal spatial location cannot be jointly interpreted or analyzed in a straight-forward manner. Two main reasons are responsible for that: on the one hand, the images have been acquired with different imaging modes: While the optical sensor is of the push-broom type, the SAR sensor carries out distance-dependent imaging. This leads to contradictory geometric distortion effects (see example shown in Figs. 2 and 3). On the other hand, differences of the sensor perspectives impose their own effect on the geometry of above-ground objects, which is also experienced when mono-sensor images are acquired from different viewing perspectives. All this shows that the fusion of multi-modal data on the signal level, which is typical for the discipline of remote sensing, is a difficult task and hardly possible without auxiliary information.

In this paper, the challenges and opportunities of signal-level data fusion in multi-sensor remote sensing scenarios are exemplified. In order to solve the necessary multi-sensor data alignment problem, we make use of a method named SimGeoI [5], which provides a framework for aligning optical and SAR, SAR and SAR or optical and optical images – regardless of the viewing perspective. The link between the heterogeneous image sources is defined based on geometric prior knowledge of the scene.

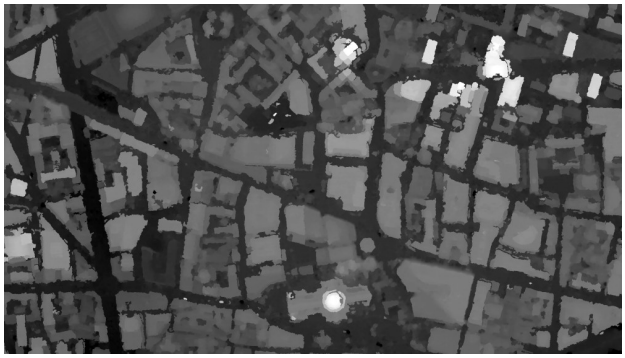
Figure 4 shows the basic work flow (see [5] for details). The geometry of the scene of interest, given as a digital surface model in Universal Transverse Mercator (UTM) coordinates,

(a) WorldView-2 image (pan channel)

(b) TerraSAR-X image (spotlight mode)



(c) Digital surface model (generated from 5 images), triangulated to closed surface



(d) Digital surface model (generated from 2 images), triangulated to closed surface

Fig. 1: Data of London test site. Meta information of the image data and the digital surface model (DSM) are required as input for SimGeoI.

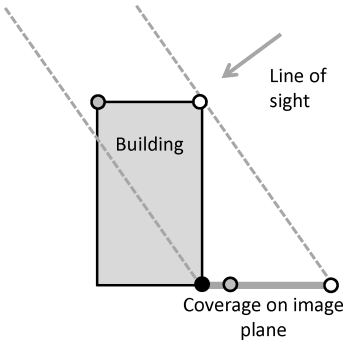


Fig. 2: Layover effect in SAR imaging due to the measurement of distance along line-of-sight. Signal contributions from roof, facade, and ground are merged for the building example.

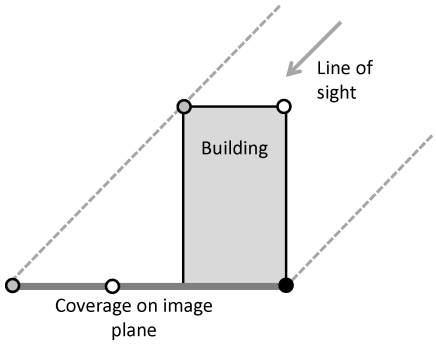


Fig. 3: Geometric projection in optical image for off-nadir angle. Facade and roof parts remain separated but are geometrically distorted.

and image meta information (image parameters, surface illumination, sensor type) are exploited to predict the appearance of scene objects. The combination of simulated images allows for the generation of interpretation layers in world coordinates (i.e. in the UTM coordinate system) which are used to extract semantically meaningful parts of the remote sensing images. The impact of signal illumination, sensor perspective, spatial distance, signal-surface intersection, and occlusion effects is covered by ray tracing methods.

The digital surface models (DSM; geometric prior knowledge) are derived from high-resolution optical data using semi-global matching [6], which requires a set of optical images for reconstructing the scene geometry (see examples in Figs. 1c and 1d). In a pre-processing step, the DSM is filtered for vegetation and noise [7]. Thereafter, elevated parts are extracted from the DSM using an adapted version of the method reported in [8] for detecting terrain (digital terrain model, DTM). The subtraction of terrain from the DSM leads to the identification of elevated objects in the DSM, referred to as normalized digital surface model (nDSM) in the remainder

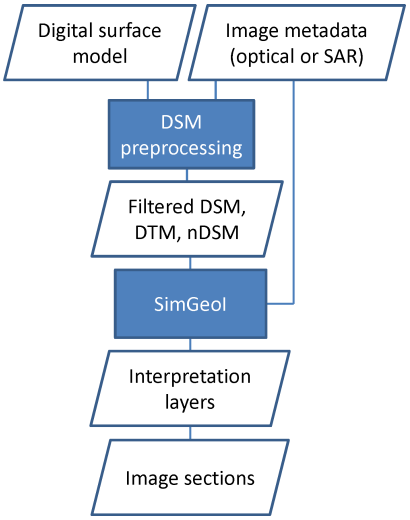


Fig. 4: Basic flow chart of SimGeoI.

of this paper. As vegetation has been filtered beforehand, the nDSM mainly contains man-made structures for urban areas. Individual building blocks can be identified in the nDSM, leading to a second type of input model for SimGeoI.

In SimGeoI, the following main steps are conducted:

1) A scene model is defined, considering the sensor perspective / signal source (known from image meta data), triangulated surface models (DSM, nDSM, DTM) and surface parameters. Direct and strong signal response is assigned to surfaces to ensure the visibility of objects in the resulting images.
2) The appearance of the DSM, nDSM, and DTM is simulated, either in the optical or SAR image plane. To this end, a ray tracer detects signal contributions throughout the modeled scene. The image pixel size is derived from the meta data.
3) The simulated images are geocoded in UTM coordinates to match the coordinate system of the satellite data.
4) Binary interpretation layers are generated by combining simulated images, marking, e.g., building extents (see examples below), shadow regions or ground. For instance, the extent of ground is marked by the difference of the simulated DSM (all parts responding) and nDSM (elevated parts responding, ground missing in scene model; see [5]). The extent of buildings is derived by simulating the extent of the nDSM, where ground and vegetation is filtered out (see examples in Fig. 6).

The basic motivation for the development of SimGeoI was to relate parts of heterogeneous remote sensing images. As a consequence, SimGeoI provides the basis for subsequent object-related image analysis, e.g., in the context of city monitoring, machine learning, or change detection. The potentials of optical images (object shape, spectral characteristics, familiar perspective) and SAR images (distance information, slant view, physical characteristics) become combinable although the images cannot be matched in image space directly.

III. Test data

The data alignment in this paper is realized for images of two sensors, WorldView-2 (optical, properties summarized in table I) and TerraSAR-X (SAR, properties summarized in table II, see [9] for more information on the sensor). Figure 1 shows the image data (London city center) to be interpreted and the digital surface model as geometrical description of the scene. The optical image in Fig. 1a corresponds to the *pan* channel of WorldView-2 which is sensitive to the full range of visible light. The TerraSAR-X image in Fig. 1b relies on signals in the X-Band of the electromagnetic spectrum (wave length: $3.1cm$) and is captured in high-resolution spotlight mode, where the time for image capturing is increased with focus on a local scene. For the case study presented below, two triangulated DSMs are used as input, being generated from two or five WorldView-2 images. The WorldView-2 images have been captured on one pass of the satellite, i.e., subsequent data takes with slightly varying perspective along the orbit (optical image in Fig. 1a is one of those). As shown in Figs. 1c and 1d, the DSM derived from five images is less noisy and better describes the geometric shape of buildings. Nevertheless, a DSM created from just two images can still be an important data source as data access is often limited in realistic scenarios. The spatial resolution of the DSM is $0.5m$ horizontally and $1m$ in height.

The image data of TerraSAR-X and WorldView-2 belong to the class of high-resolution sensors and represent a renowned and accessible source of information in the remote sensing community. The data varies in sensor perspective, signal source, and acquisition date.

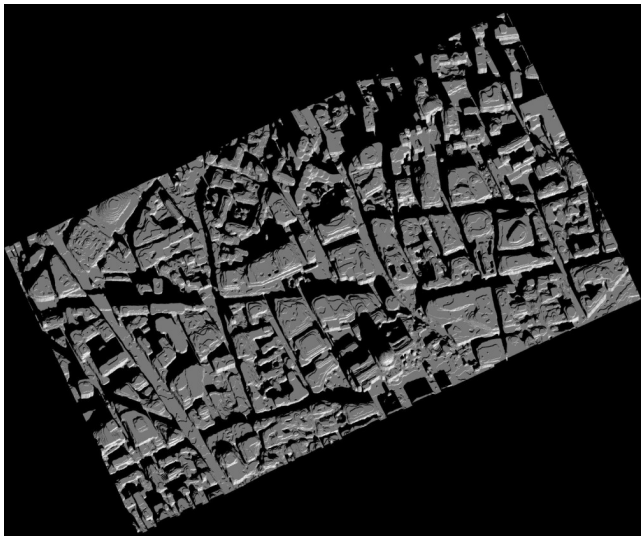
The urban scene of London has a spatial extent of $1150m \times 650m$ and contains different types of buildings / building blocks with variable height between $10m$ and $85m$. The visual interpretation of the optical image is straightforward except for the impression of height whereas identifying and interpreting geometric structures in the SAR image is a hard task.

TABLE I: WorldView-2 data properties for test site London (pan channel, geo-referenced, level 2A).

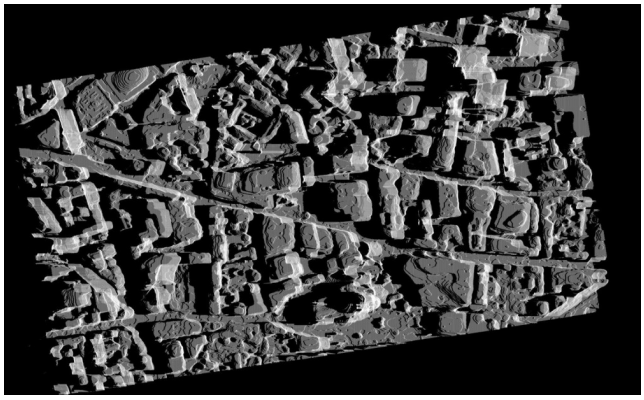
Pixel spacing (east, north)	0.5 m
Off-nadir angle (at scene center)	10.8°
Scene azimuth angle	208.7°
Sun azimuth angle	177.2°
Sun elevation angle	27.6°
Acquisition date	2011-10-22

TABLE II: TerraSAR-X data properties for test site London (geo-referenced, level 1B).

Azimuth resolution	1.14 m
Ground range resolution	1.0 m
Pixel spacing (east, north)	0.5 m
Signal incidence angle (at scene center)	41.0°
Orbit	ascending
Acquisition date	2008-05-05



(a) Optical Image (black: sun shadow)



(b) SAR-Image (black: sensor shadow), distance to the sensor increases from right to left

Fig. 5: Simulated images using SimGeoI. Geometric model: DSM based on 5 WorldView-2 images.

(a) Layer for WorldView-2 image, DSM based on 5 images

(b) Layer for WorldView-2 image, DSM based on 2 images

(c) Layer for TerraSAR-X image, DSM based on 5 images

(d) Layer for TerraSAR-X image, DSM based on 2 images

Fig. 6: Building layers, geocoded in UTM coordinates. Building areas marked with white color.

IV. From scene interpretation to image alignment

The difficulty of aligning optical and SAR images is related to object height. Elevated parts of the buildings are mapped differently due to the contrary imaging concepts of the sensors. With increasing height, building parts are mapped towards the SAR sensor due to decreasing spatial distance. In the optical case, building parts are mapped in the opposite direction in the image plane. Hence, corresponding structures appear at different UTM coordinates. One may solve the alignment task by geometrically transforming one image to the other. However, this would mean a loss of information in one of both data sources. Moreover, this way of image fusion is obsolete for all surfaces not visible from both sensors. SimGeoI performs image alignment in the original image planes. Image pixels are not matched geometrically but linked to individual scene objects derived from so-called interpretation layers.

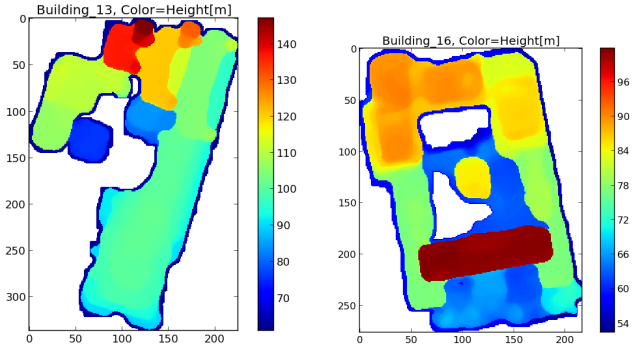
The generation of interpretation layers is based on simulating images with ray tracing techniques (extended version of POV-Ray [10]). Figure 5 shows the simulated optical and SAR image for the urban scene in London. The definition of the sensor and signal position allows for the consideration of sensor perspective and shadowing effects (sun shadow in optical image; signal shadow in SAR image). Assigning high roughness to the DSM, i.e., strong diffuse reflection, helps to

distinguish the spatial extent of responding surfaces. SimGeoI does not focus on realistic image radiometry but the separation of foreground (bright color) and background (black color) in order to generate layers for automated interpretation.

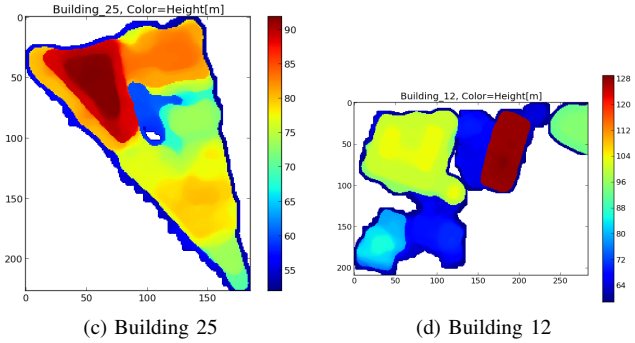
Using the DSM and both a DTM and an nDSM derived thereof as input to SimGeoI, the spatial distribution of scene content can be predicted, e.g., ground parts, buildings, and shadow. As an example, Fig. 6 shows simulated interpretation layers for building areas (see [5] for further types of interpretation layers).

When comparing the masks for the optical (Figs. 6a and 6b) and SAR image (Figs. 6c and 6d), it is obvious that the building shapes do not match, which is due to the different imaging concepts of the sensors. Therefore, image pixels with equal UTM coordinates are not comparable for most areas of the scene. SimGeoI helps to link parts of images to the same object (here: building parts extracted from full scene). The geometry of the object is the connecting element.

Substituting the connecting element, we can move one step further. To this end, individual building models are extracted from the nDSM. Pixels with height information are grouped to a new model if the identified segment exceeds a size threshold. In case of the London example, 40 building models are derived using a threshold of 4000 pixels and the DSM based on five WorldView-2 images. Figure 7 shows four selected models with variable shape and height intervals.

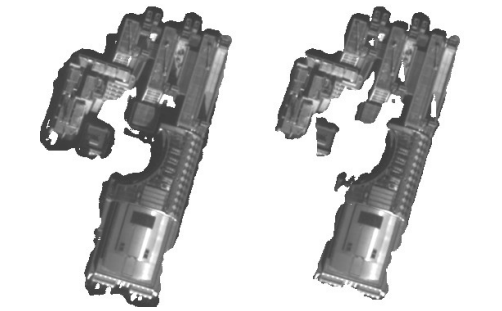


(a) Building 13 (b) Building 16

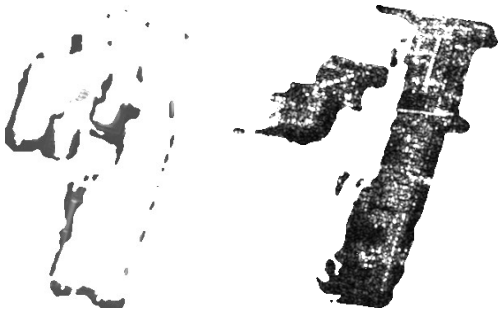


(c) Building 25 (d) Building 12

Fig. 7: Building models for processing depth level 2, color indicating height. DSM based on five WorldView-2 images.



(a) Full extent (WV-2) (b) Illuminated parts (WV-2)

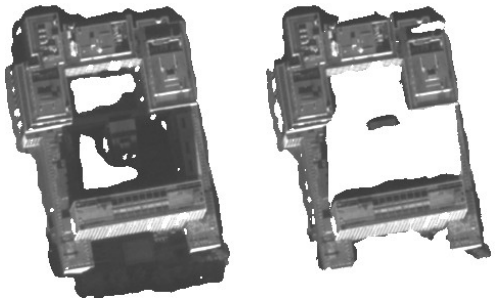


(c) Shadowed building parts (WV-2) (d) Layover (TS-X)

Fig. 8: Extracted image parts from WorldView-2 and TerraSAR-X images for building 13. WV-2: WorldView-2; TS-X: TerraSAR-X.

Figures 8, 9, 10, and 11 provide examples for image sections extracted for these individual building blocks depicted in Fig. 7. Looking at the image sections, the basic aim of the SimGeoI method becomes obvious: The consideration of sensor characteristics / perspectives and the impact of scene height allows for the prediction of building shapes in the optical and SAR images and subsequently for the extraction of related image parts.

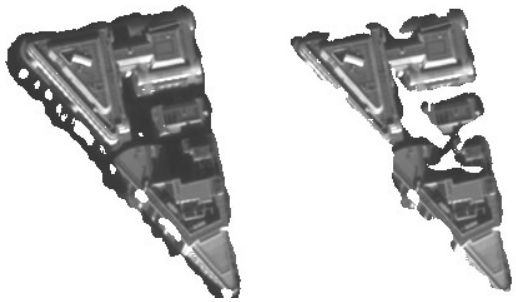
The impact of scene heights increases with growing off-nadir angle when geometric distortions in the images become more prominent. In case of SAR imaging, big off-nadir angles are standard to obtain sensitivity with respect to differences in distance. Due to lower distance, elevated parts of the buildings are mapped towards the sensor and merged with ground contributions with equal distance. This effect is called ”layover” and leads to the challenge of interpreting SAR images of urban areas. The results for buildings shown in this paper (Figs. 8d, 9d, 10d, and 11d) exemplify that is hardly possible to define building outlines manually in SAR images (see also visual impression of Fig. 1b). Due to signal emission and detection in the microwave domain, most urban structures are characterized by specular multiple reflections, which lead to salient point-like signatures in the images. Information on the extent of surfaces is hard to identify due to weak diffuse signal response from flat building surfaces (as a result, buildings are partly invisible) and the layover effect



(a) Full extent (WV-2) (b) Illuminated parts (WV-2)

(c) Shadowed building parts (WV-2) (d) Layover (TS-X)

Fig. 9: Extracted image parts from WorldView-2 and TerraSAR-X images for building 16.



(a) Full extent (WV-2) (b) Illuminated parts (WV-2)

(c) Shadowed building parts (WV-2) (d) Layover (TS-X)

Fig. 10: Extracted image parts from WorldView-2 and TerraSAR-X images for building 25.

(different building structures mapped to the same image pixels due to equal distance). SimGeoI allows for the automatic identification of building outlines in the SAR image which eases follow-up steps of image interpretation significantly.

Optical images are mostly captured with nadir perspective or small off-nadir angles. However, bigger off-nadir angles are likely for near-real time image acquisitions on demand for particular urban areas of interest. As an example, the WorldView-2 image of London was captured with an off-nadir angle of 10.8°, which is moderate but already leads to a noticeable geometric projection effect in the image. In contrast to the SAR image, elevated building parts are mapped away from the sensor. The view on facade structures is opened.

Besides the difference of imaging modes, the impact of sensor perspective leads to secondary geometric projections. The optical image has been captured from the south, leading to mapping towards the north. The SAR image has been captured on a descending orbit, i.e., the sensor line-of-sight was approx. from east to west. As a consequence, the off-nadir angles of the sensors lead to imaging of different facades of the buildings. As geometric projections of one pixel from one image to the other are not possible, this constitutes one further reason to conduct image alignment on the object-level. SimGeoI marks and extracts groups pixels which are not comparable in the image plane but linked to the same object.

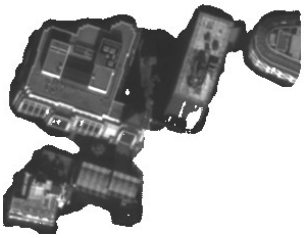
In the optical case, further types of interpretation layers are reasonable, e.g., considering the impact of sun illumination (Figs. 8b, 9b, 10b, and 11b). In terms of the impact of building geometry, radiometric changes in the images can be related to sun shadows (Figs. 8c, 9c, 10c, and 11c).

The main advantage of SimGeoI lies in aligning image sections of optical images and SAR images in the context of individual (building) objects. Accordingly, selected pixels of the image sources can be connected despite different imaging modes, sensor perspectives / resolutions, and signal types. Information extracted from image parts remains connected to scene models. The alignment of the heterogeneous images is highly relevant for the remote sensing community in the context of city monitoring or change detection tasks, where methods have to be flexible for incoming data. Moreover, methods for object-related analysis of heterogeneous image data are rare due to missing strategies for the compensation of geometric distortion effects.

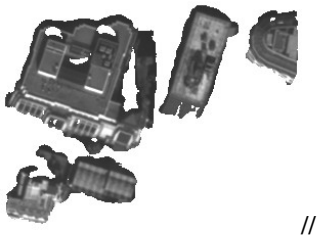
V. Impact of the Prior Knowledge on SimGeoI-based Image Alignment

As mentioned earlier, prior knowledge about the 3D structure of the scene of interest is crucial for multi-sensor remote sensing image alignment using the SimGeoI simulation framework. It is thus necessary to discuss the impact of DSM that constitutes this prior knowledge.

First and foremost, it can be seen from Figs. 1c and 1d that the input model derived from two WorldView-2 images leads to a poorer definition of building shapes and outlines in the

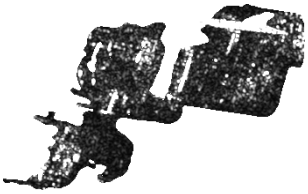


(a) Building 12: full extent (WV-2)



(b) Building 12: illuminated parts (WV-2)

(c) Building 12: shadowed building parts (WV-2)



(d) Building 12: layover (TS-X)

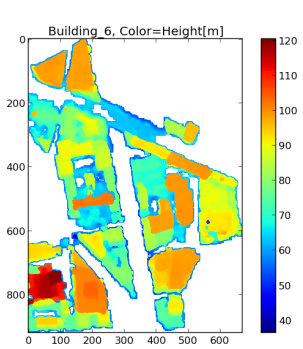
Fig. 11: Result for input model with connected buildings. WV-2: WorldView-2; TS-X: TerraSAR-X.

resulting masks. However, the result confirms that one pair of optical images is still enough to provide appropriate geometric information of the scene.

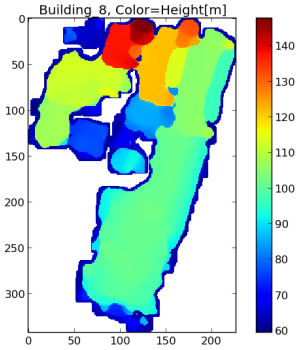
Furthermore, the examples shown in this paper indicate that the spatial resolution of the WorldView-2 based DSM is sufficient for describing the geometry of buildings. This is encouraging in the context of real scenarios, e.g. city monitoring or change detection, where access to surface models with very high spatial resolution is limited. Nevertheless, Fig. 7d exemplifies the challenge of DSM preprocessing. Several building parts are connected to one model where visual perception would favor further separation to sub-models. In addition, Fig. 11 exemplifies the impact of limited nDSM segmentation on image extraction. As the underlying geometric model contains several buildings (see Fig. 7d), the resulting image sections do not reach the level of manual interpretation. Possible improvements comprise:

- Improvement of the DSM: increasing the spatial resolution leads to a more precise definition of height steps and building outlines.
- Improvement of terrain extraction: the DSM2DTM algorithm can be substituted with alternative approaches. Methods based on neural network strategies are promising in this field.
- Improved nDSM segmentation: the identification of building models is based on a size criterion of segments in the nDSM. The strategy could be improved by learning typical building shapes to be identified in the nDSM.

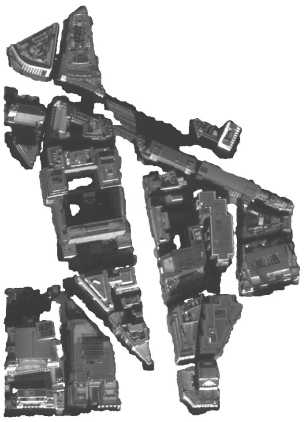
The summarized options for improvement do not influence the framework of SimGeoI whose simulation functionality is independent of the quality of the input model. The ray tracing



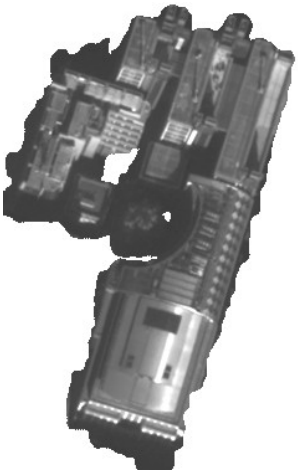
(a) Missed building separation



(b) Limited definition of building shape (compare building 13 in Fig. 7)



(c) Extracted image part (WV-2)



(d) Extracted image part (WV-2)

Fig. 12: Impact of DSM quality reduction (based on two optical images).

environment is open to highly detailed geometric models (see [10]).

In contrast to these improvement perspectives, Fig. 12 indicates the impact of reducing the quality of the DSM. Here, the results for individual buildings are based on a DSM which has been derived from only two WorldView-2 images (i.e. classical stereo reconstruction). On the one hand, distinguishing individual models in the nDSM becomes more difficult (connected building block in Fig. 12a) and leads to larger image extracts provided by SimGeoI. On the other hand, the geometric representation of building outlines in the nDSM is more noisy (Fig. 12c) which leads to a poorer definition of building shapes in the image (Fig. 12d; compare to result in Fig. 8a for the nDSM derived from five WorldView-2 images). To conclude, building blocks have to be more spatially isolated to be identified as input models (applicability of SimGeoI restricted to less buildings), extracted image sections are less representative at building borders. Nonetheless, DSMs from stereo views provide a suitable source of geometric scene knowledge.

The application of SimGeoI is focused on urban areas so far. However, it may be also applied to alternative regions with uncompensated object geometry in the related sensor images.

VI. Conclusion & Outlook

The alignment of multi-modal image data in remote sensing requires geometric prior knowledge in order to consider the impact of sensor type and perspective. This paper has presented a framework, named SimGeoI, to solve the task by exploiting a ray tracing-based simulation model and a digital surface model of the scene as prior knowledge. For a case study containing the city center of London object-related image parts of an optical image and a synthetic aperture radar (SAR) image have been matched successfully using the presented framework. It has been shown that state-of-the-art digital surface models provide sufficient geometric details to link sections of remote sensing images. In the context of urban scenes, directions for possible improvements of the application of SimGeoI have been summarized and the consequences of reduced DSM quality have been discussed.

Future work will focus on exploiting the potential of aligned optical and SAR images in the context of classification by extracting multi-sensor image pairs for training classifiers, and in the context of city monitoring and change detection. Moreover, quantitative measures are required to evaluate the impact of the input scene model on the level of image alignment.

References

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

[2] Michael Schmitt, Florence Tupin, and Xiao Xiang Zhu, "Fusion of SAR and optical remote sensing data – challenges and recent trends," in *Proceedings of IGARSS Conference*, 2017, pp. 5458–5461.

[3] G. Palubinskas, P. Reinartz, and R. Bamler, "Image acquisition geometry analysis for the fusion of optical and radar remote sensing data," *Int. J. Image Data Fusion*, vol. 1, no. 3, pp. 271–282, 2010.

[4] M. Schmitt and X. Zhu, "On the challenges in stereogrammetric fusion of SAR and optical imagery for urban areas," in *The Int. Arch. Photogram. Rem. Sens. Spatial. Inform. Sci.*, 2016, vol. 41, pp. 719–722.

[5] S. Auer, I. Hornig, M. Schmitt, and P. Reinartz, "Simulation-based interpretation and alignment of high-resolution optical and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 10, no. 11, pp. 4779–4793, 2017.

[6] P. d'Angelo and G. Kuschk, "Dense multi-view stereo from satellite imagery," in *Proceedings of IGARSS Conference*, 2012, pp. 6944–6947.

[7] Rebecca Ilehag, "Exploitation of digital surface models from optical satellites for the identification of buildings in high resolution SAR imagery," M.S. thesis, KTH, School of Architecture and the Built Environment (ABE), Urban Planning and Environment, Geoinformatics, 2016.

[8] R. Perko, H. Raggam, K.H. Gutjahr, and M. Schardt, "Advanced DTM generation from very high resolution satellite stereo images," in *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2015, vol. II-3/W4, pp. 165–172.

[9] W. Pitz and D. Miller, "The TerraSAR-X satellite," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 615–622, 2010.

[10] POV-Ray, "POV-Ray," Persistence of Vision Raytracer Propriety Limited, www.povray.org [checked: (14.03.2018], 2018.