9

Deep Learning for Image Matching and Co-registration

Maria Vakalopoulou, Stergios Christodoulidis, Mihir Sahasrabudhe, and Nikos Paragios

9.1 Introduction

Image matching and registration are some of the most important and popular problems for many communities, including Earth observation. Efficient and robust algorithms that can address such topics are essential for several other tasks including, but not limited to, optical flow, stereo vision, 3D reconstruction, image fusion, and change detection. Deep learning algorithms are becoming more and more popular, providing state-of-the-art performance for various problems including image matching and registration. These algorithms prove very efficient running time and robustness with variety of studies reporting their success in supervised and unsupervised settings.

Given a pair of images depicting the same area, image matching is the process of comparing the two images (source image S and target (or reference) image R) to obtain a measure of their similarity, while image registration is the process that aligns or maps these images by finding a suitable transformation between S and R. In particular, both image matching and registration are measuring or mapping identical pixels in the pair of images with the focus of the second to align S to R as accurately as possible. Although these problems seem to be easy conceptually, they are still an open research area for a variety of communities, considered as ill-posed problems that suffer from many uncertainties. This is mainly due to the nature of algorithms and images on which small changes in translation, illumination, or viewpoint can significantly affect these algorithms' performance even if the depicted areas are exactly identical. Therefore, numerous approaches have been proposed to address these problems and have been summarized in different surveys (Zitova and Flusser 2003; Sotiras et al. 2013; Leng et al. 2018; Burger and Burge 2016; Weng and He 2018). Nowadays, however, with the recent advances of deep learning, more and more techniques integrate these technologies for both matching and image registration, offering better performances especially in time requirements.

Some of the most common problems that image matching and registration algorithms need to address, especially for earth observation applications can be grouped in four main categories, namely (i) radiation distortions; (ii) geometric changes; (iii) areas including changes; and (iv) multimodal, large-scale sources of data. Starting with the first group,

Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences, First Edition. Edited by Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein.

radiation distortions refer to the difference between the real emissivity of the ground objects and the one that is represented in the image level. This difference can be caused mainly by the imaging properties of the sensor itself or radiation transmission errors caused by the atmosphere during the acquisition of the objects' emissions. The later is also known as Bidirectional Reflectance Distribution Function (BRDF) with a lot of algorithms being proposed for its modeling (Montes and Ureña 2012). The second group refers to the geometric differences in the ground objects that have to do with differences in viewpoints of the sensors and the height of objects influencing mainly high-resolution images (with spatial resolution higher than 10 meters). The next group refers to the difficulty of these methods to work on places that contain changed areas. Both image matching and registration problems assume that the depicted regions are mainly identical, being unable by default to work properly on regions that contain changes, something that is quite common on earth observation and remote sensing datasets. Finally, it is also very challenging to create algorithms that are working robustly for multimodal datasets such as Synthetic-Aperture Radar (SAR) and multispectral or hyperspectral optical sensors or images produced by sensors with significantly different spatial resolutions. A variety of different sensor characteristics are available for earth observation and there is a significant need for matching and registration algorithms that can fuse their multitemporal information. All these cases make the problems of matching and co-registration very challenging and the use of algorithms proposed by other communities such as computer vision or robotics really challenging to be adapted to satellite data (Le Moigne et al. 2011; Nicolas and Inglada 2014).

Traditionally, image matching and image registration are two closely related problems, with the first providing usually reliable information for the second (Figure 9.1). Different sub-regions or pixels from the source and reference images are matched and used to define the best transformation model, G, to map S to R, thus resulting in a warped image D. Depending on the implemented strategy the matching algorithm can be applied globally, searching for the most similar regions in the entire image, or it can be applied locally

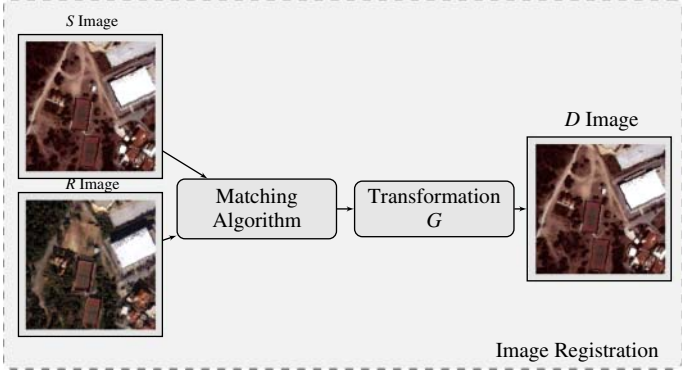

Figure 9.1 A schematic diagram of the image matching and image registration techniques for Earth observation. Image registration has two main components: the matching algorithm that is measuring how similar are different regions in the image; and the definition of the transformation G, which will be applied to the S image to generate the warped image D.

by searching for the best matching on predefined regions. The choice of the matching strategy also depends on the choice of transformation model used for the registration algorithm. Nowadays, various methods based on deep learning approaches are proposed from both the computer vision and earth observation fields. Most of the techniques that originate from the earth observation field are focusing on high-resolution datasets (Ma et al. 2019) due to the more challenging nature of this kind of images and their need for dense and more complex registration models.

Starting with image matching, traditionally two main components are usually applied – the feature extraction of keypoints or sub-regions of the images and the establishment of the proper correspondences. Feature extraction methods typically rely either on intensity-based methods using the raw image's intensities directly or on higher-level representations extracted from the pair of images. These representations are typically produced using either classical image descriptors or they are obtained using deep learning approaches. After feature extraction, optimal correspondences are found using a similarity function. Typical choices for this similarity function are mean squared error, normalized cross-correlation, and mutual information. The implementation of the similarity function in a deep learning framework is usually achieved using Siamese or triplet networks that share their weights (Kaya and Bilge 2019).

As far as the image registration task is concerned, depending on the transformation used, the methods can be categorized into: (i) rigid or linear; and (ii) deformable or elastic. Rigid methods define maps with transformations that include, e.g., rotations, scaling, and translations. They are global, and hence cannot model local geometric differences between images, which is usually the case in high-resolution datasets. However, they are very efficient and robust for the co-registration of satellite imagery. On the other hand, the deformable methods rely on a spatially varying model by associating the observed pair of images through non-linear dense transformations. After obtaining the optimal transformation G, the source image is resampled to construct the warped image which is in the same coordinate system as the reference image. Deep learning and convolutional neural networks have been used for image registration (Kuppala et al. 2020) while methods also based on generative models (Mahapatra and Ge 2019) and deep reinforcement learning are proposed in the literature for both 2D and 3D registration (Liao et al. 2016) mainly for the medical and computer vision communities.

This chapter focuses on recent advances in image matching and registration for earth observation tasks with emphasis on emerging methods in the domain and the integration of deep learning techniques. To study these recent advances we analyse their key components independently. The rest of the chapter is organized as follows. In section 9.2, we present a detailed overview of existing literature for both image matching and image registration focusing on the recent deep learning techniques. In section 9.3, we discuss and present an unsupervised deep learning technique applied to high-resolution datasets and compared with conventional image registration techniques. We describe the dataset used for this study in section 9.3.4, followed by experiments and results in 9.3.5. Finally, in 9.4 we summarize the chapter and enumerate future research directions for these algorithms.

9.2 Literature Review

9.2.1 Classical Approaches

Image matching has been dominated for long by hand-engineered and feature-based methods, with SIFT (Lowe 1999) being one of the most commonly used feature descriptor applied also in remote sensing with or without small variations (Vakalopoulou and Karantzalos 2014; Chen et al. 2018a). Additionally, descriptors such as SURF (Bay et al. 2008), DAISY (Tola et al. 2010), BRIEF (Calonder et al. 2012), the very recently proposed HOSS (Sedaghat and Mohammadi 2019), and other variations were equally popular. These descriptors are then used with intensity-based or more complex similarity functions such as mutual information (Viola and Wells 1995) or correlation-based methods (Pratt 1978) to establish correspondences. RANdom SAmple Consensus (RANSAC) (Fischler and Bolles 1981), a model-based technique, was also very commonly used to filter and establish proper matching of images or points. Although these features have alleviated the influences of radiometric and geometric deformations to some extent, their performance was significantly lower in the case of multi-sensor data, while their detection repeatability was still low. Tuytelaars and Mikolajczyk (2008) reports repeatability rates that were below 50% for datasets with three band image pairs.

Additionally, image registration techniques were based on the correspondences that were defined by the image matching algorithms to obtain the optimal transformation parameters, G that maps in the most optimal way S to R. Starting with the rigid methods, in remote sensing the transformations that are commonly used are translations, rotations, scaling, and shearing. These transformations can be captured by affine and homography mappings, which are used frequently in remote sensing applications (Zitova and Flusser 2003). In practice, these transformations can be described by a 3×3 matrix having 6 and 8 degrees of freedom respectively. To obtain the parameters of the transformation used, a number of correspondences have to be established (minimum 3 or 4 per dimension, respectively). The resulting system can be solved using least squares method to find the optimal values (Szeliski 2010). Numerous techniques (Wu et al. 2012; Vakalopoulou and Karantzalos 2014; Li and Leung 2007) fall into the category of methods using rigid transformations and have been tested using different spectral and spatial resolution satellite imagery.

While rigid methods are simple and efficient, they do not have the capacity to produce more complex transformations that can vary locally. To capture such locally-linear or non-linear transformations, deformable methods creating a dense transformation G are instead employed. Such methods construct a deformation grid G that defines transformations that locally express the correlation between the observations. Sotiras et al. (2013) presents a detailed survey of deformable registration methods and their categorization based on the geometric model they are using. These methods are commonly used in medical imaging and remote sensing datasets where the deformations between the images are not homogeneous (Karantzalos et al. 2014; Marcos et al. 2016). Some of the deformable strategies that have been proposed are based on correlation of objects (Marcos et al. 2016), contours (Hui Li et al. 1995), or intensity- and area-based similarity methods (Karantzalos et al. 2014; Vakalopoulou et al. 2016).

Table 9.1 Grouping of image matching techniques depending on the type of imagery they have been applied to.

Type of Imagery	Methods applied on Earth Observation
Optical to Optical	Altwaijry et al. (2016); Zhu et al. (2019a); En et al. (2018); Liu et al. (2018a); Chen et al. (2017b); Yang et al. (2018); He et al. (2018); Dong et al. (2019); Zhu et al. (2019); Jia et al. (2018); He et al. (2019a); Tharani et al. (2018); Wang et al. (2018)
SAR to SAR	Quan et al. (2016); Wang et al. (2018)
SAR to Optical	Merkle et al. (2017); Bürgmann et al. (2019); Merkle et al. (2018); Hughes et al. (2018); Quan et al. (2018); Ma et al. (2019); Merkle et al. (2017); Zhang et al. (2019)
Other	Ma et al. (2019); Zhang et al. (2019)

9.2.2 Deep Learning Techniques for Image Matching

Several deep learning methods for image matching in remote sensing images have been explored recently in the community. Most of the methods are based on Siamese architectures, extracting features from CNNs, and providing similarity scores for the input patches. Similar approaches were proposed by a variety of studies in the computer vision community for both the supervised and unsupervised settings (Revaud et al. 2016; Zagoruyko and Komodakis; Han et al. 2015). Even though these techniques are quite recent, deep learning-based methods have been shown to outperform traditional ones. A summary of deep learning-based methods applied to remote sensing is presented in Table 9.1. Methods have been grouped depending on the type of data they use, including supervised and unsupervised techniques, with the first being more dominant.

We begin with methods for matching optical to optical imagery. In Altwaijry et al. (2016), the authors propose an attention-based deep learning architecture trained with weak labels. In particular, the method provides local correspondence by framing the problem as a classification task, integrating an attention mechanism to produce a set of probable matches. To train their model, the authors used a dataset of urban high-resolution patches consisting of the labels “same” and “different”. A similar weak annotation for different types of optical data is also used in En et al. (2018). A similar strategy is also presented in He et al. (2019a) for matching medium resolution multi-temporal imagery. In Zhu et al. (2019a) the authors proposed the use of densely-connected CNNs in a Siamese architecture to match RGB images with infrared images reporting very promising results for image pair matching. In Chen et al. (2017b) a deep hashing network is proposed to search for feature point matching. Concerning unsupervised methods for matching, adversarial networks (see Chapter 3) are mainly used for similar types of optical data. More specifically, in Tharani et al. (2018) an encoder-decoder architecture combined with a deep discriminator network to replace distance metrics is proposed. The authors report very promising results, while they also provide a comparative study of different commonly used convolutional architectures for the accurate registration of different land cover classes.

Siamese architectures are also popular for SAR to Optical image matching. In Merkle et al. (2017) a Siamese architecture is proposed to generate reliable matching points between TerraSAR-X and PRISM images. For the same type of images, a conditional generative adversarial network (see Chapter 3) is trained in Merkle et al. (2018) to generate SAR-like image patches from optical images to enhance the performance of known classical matching approaches. Moreover, a (pseudo-) Siamese network is proposed in Hughes et al. (2018). Medium- and high-resolution SAR and optical data are evaluated in Bürgmann et al. (2019), presenting an approach for matching ground control points (GCPs) from SAR to optical imagery. The training of conditional generative adversarial networks (see Chapter 3) is also proposed in Merkle et al. (2017) to generate artificial templates and the matching of optical to SAR data. Finally, a combination of deep and local features is used in Ma et al. (2019) to match and register multimodal remote sensing data.

Deep learning methods are also used to match images from completely different sources of data such as satellite images with maps. In Ma et al. (2019) the authors evaluated their methods using a pair of an optical image and the corresponding Tencent Map providing very promising results. A similar approach based on Siamese architectures is proposed in Zhang et al. (2019), evaluating its performance on multimodal data including optical to map matching.

9.2.3 Deep Learning Techniques for Image Registration

In computer vision, there are works that include ideas of modeling transformation (Hinton 1981) directly, learning transformation invariant representations (Kanazawa et al. 2014), and attention/detection mechanisms for feature selection (Gregor et al. 2015). The study presented in Jaderberg et al. (2015) was one of the first to introduce the idea of using a deep learning-based architecture to eliminate intra-object variance by transforming intermediate feature maps. The spatial transformer network proposed in this study is a trainable module that can be integrated and trained together with any deep learning architecture. The module estimates an optimal transformation of intermediate feature maps to remove variance due to intra-object shape differences, object placement, and object size, thus aiding in recognition by mapping objects to a canonical space. The estimated transformations can include rigid deformations such as scaling, cropping, rotations, as well as non-rigid deformations.

In remote sensing, the literature for methods that can obtain the parameters of the transformation directly from the developed models is not very vast. Most methods use deep learning-based models to match the images (as mentioned in the previous sub-section) and then they use these matching to obtain the parameters of the registration model independently. Recently, in Vakalopoulou et al. (2019) the authors proposed the use of the spatial transformer to regress rigid and non-rigid deformations directly from source and target images under a deep learning framework for the registration of high-resolution satellite datasets.

Moreover, a deep learning-based method that could output the displacement field directly from the networks was proposed in Zampieri et al. (2018) to register optical imagery to cadastrial maps of buildings and road polylines. The method was based on a fully convolutional architecture that learned scale-specific features predicting the deformations directly.

Additionally, the authors proposed an improvement to their previous work in Girard et al. (2019) by developing a multi-task scheme for simultaneous registration and segmentation, which improved the performance of the reported registration.

9.3 Image Registration with Deep Learning

In this section, we describe the approach presented in Vakalopoulou et al. (2019) in detail, which is one of the methods proposed recently for predicting the deformation maps in an end-to-end deep neural network. As discussed in the previous section, end to end deep learning architectures are more commonly used for matching than registration problems. It is for this reason that we have chosen to focus on the latter in this chapter. The method presented here is a modification of a recent work on accurate and efficient registration of 3D medical volumes (Christodoulidis et al. 2018). Both implementations are available online at https://github.com/stergioc/smooth-transformer.

The main advantages of this method are fourfold: (i) a completely unsupervised technique for regressing the dense deformation grid G from a pair of images, (ii) a modular formulation that couples rigid and deformable registration within a single optimization, (iii) a framework that is independent of the CNN architecture, (iv) fast inference allowing real-time applications even for very large-scale remote sensing datasets. The proposed framework can be divided into three different components – the transformation strategy, the CNN architecture, and the optimization procedure.

9.3.1 2D Linear and Deformable Transformer

The main component of the proposed CNN architecture is the 2D transformer layer, which enables the architecture to regress the spatial gradients. This layer warps the image S under a dense deformation G to create the warped image D that best matches R. This operation is defined by the equation

$$D = \mathcal{W}(S, G), \tag{9.1}$$

where $\mathcal{W}(\cdot, G)$ indicates a sampling operation $\mathcal{W}$ under the deformation G.

In our implementation, the deformation is fed to the transformer layer as sampling coordinates, which uses a backward bilinear interpolation sampler as $\mathcal{W}$, adapting a strategy similar to Shu et al. (2018). The backward sampling indicates that for every pixel of the warped image a coordinate in the original image S is computed indicating where the intensity value originates. Often backward sampling is preferred compared to forward due to the discrete nature of the images. The backward bilinear interpolation sampler is defined as

$$D(\vec{p}) \;= \mathcal{W}(S, G)(\vec{p}) = \sum_{\vec{q}} S(\vec{q}) \prod_{d} \max\left(0, 1 - \left|[G(\vec{p})]_d - \vec{q}_d\right|\right), \tag{9.2}$$

where $\vec{p}$ and $\vec{q}$ denote pixel locations, $d \in \{x, y\}$ denotes an axis, and $[G(\vec{p})]_d$ denotes the d-component of $G(\vec{p})$.

The formulation in this case consists of two different components – one which calculates a linear/affine transformation, and another that calculates a dense transformation. Depending on the application and the type of data, these two terms can be used and trained together

or separately. In case, that these two operations are trained at the same time, they are applied one after the other, by integrating first the affine component and then the deformable for more fine transformations. Such scheme can be described by

$$\mathcal{W}(S, G) = \mathcal{W}(S, \mathcal{W}(G_N, G_A)) \tag{9.3}$$

where G_A represents the affine deformation grid, while G_N represents the deformable one. Here it should be mentioned that the network is trained end-to-end, optimizing both linear and deformable parts simultaneously. G_A is computed from six regressed affine transformation components, denoted by a 2×3 matrix A. For the deformable part G_N, an approach similar to Shu et al. (2018) is adopted. Instead of regressing the components of G_N directly, we regress instead a matrix Φ of spatial gradients along x- and y- axes. As is discussed in Shu et al. (2018), this approach helps generate smoother grids that render the deformable component, making it easier to train. The actual grid G_N can then be obtained by applying an integration operation on Φ along x- and y-axes, which is approximated by the cumulative sum in the discrete case. Adopting such a strategy enables us to draw some conclusions on the relative position of adjacent pixels in the warped image based on Φ. Concretely, two pixels $\vec{p}$ and $\vec{p} + 1$ will have moved closer, maintained distance, or moved apart in the warped image, if $\Phi(p)$ is respectively less than 1, equal to 1, or greater than 1. Such an approach avoids self-crossings, while allows the control of maximum displacements among consecutive pixels.

9.3.2 Network Architectures

Such formulation is independent of the network architecture and according to the application and dataset used, different ones can be incorporated. To test this modular nature of the proposed approach, two different architectures were employed – one based on dilated filters and one based on maxpooling. The two architectures are presented in detail in Figure 9.2.

The network architecture is based on an encoder-decoder scheme. For the encoder part two different sets of experiments were constructed. The first is very similar to the one presented in Anthimopoulos et al. (2018) adopting for the encoder dilated convolutional kernels together with feature merging, while the decoder employs non-dilated convolutional layers. Specifically, a kernel size of 3×3 was set for the convolutional layers while LeakyReLU activation was employed for all convolutional layers. Each of the encoder-decoder parts contains four of these layer blocks with the feature maps starting from 16 and being doubled for each block, resulting in a 128 feature map. Before the decoder, all the feature maps were concatenated in order to create a more informative, multi-resolution feature space for the decoder.

The second architecture follows a U-Net like architecture (Ronneberger et al. 2015b) adopting consecutive layers of 2D convolutional layers with kernel size 3×3 followed by instance normalization, LeakyReLU and max-pooling that reduce the dimension of the input by half at each layer. The encoder part consists of four of these layer blocks with features maps from 16 to 128. The decoder part consists of a symmetric part where the max-pooling is replaced by upsampling to return the input to the initial dimensions. Moreover, skip connections flow information from the encoder to the decoder part.

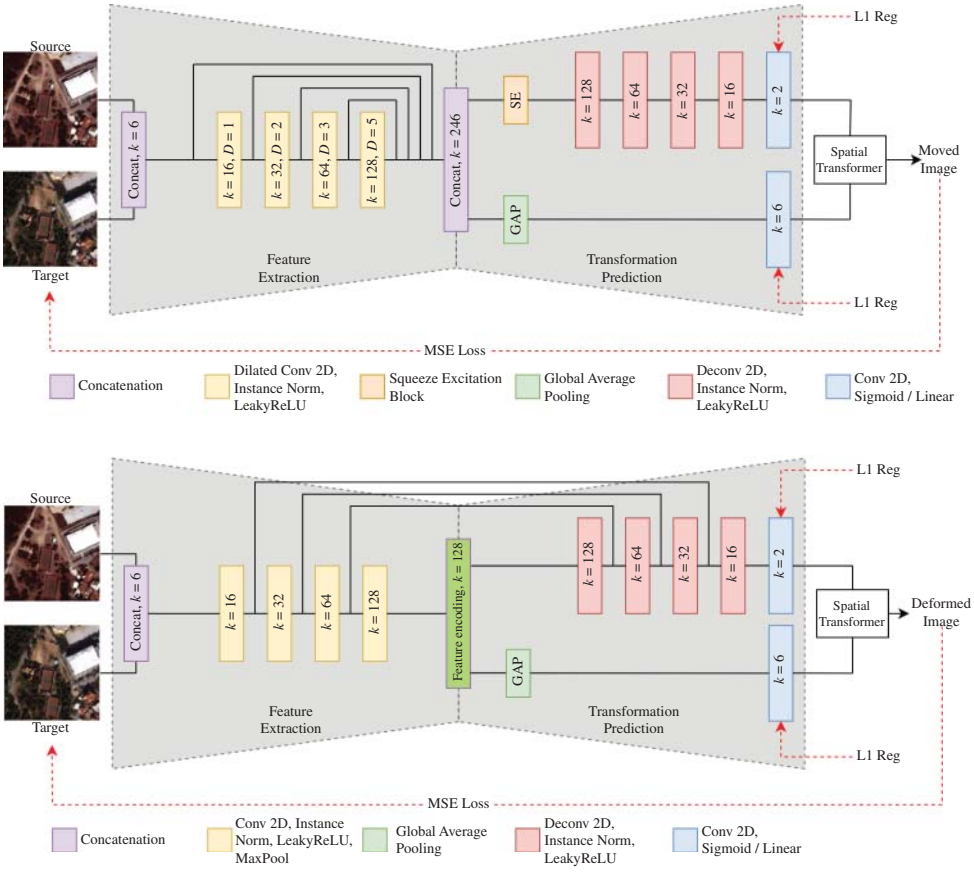

Figure 9.2 A schematic diagram of two different architectures presented in this chapter. The architecture consists of two different parts following an autoencoder scheme: the feature extraction part and the part for the prediction of the transformation.

The decoder part has two different branches, one that calculates the affine parameters and one the deformable ones. For the linear/affine parameters A, a linear layer was used together with a global average pooling to reduce the spatial dimensions, while for the spatial gradients Φ a sigmoid activation was employed. Finally, the output of the sigmoid activation was scaled by a factor of 2 to allow consecutive pixels to have larger displacements than the initial.

9.3.3 Optimization Strategy

The entire framework is trained in a completely unsupervised way, in that it does not require registered pairs of images to be trained. As similarity function between the R and D images, the mean squared error (MSE) is used and the overall loss is defined as

$$\text{Loss} = \|R - \mathcal{W}(S, G)\|^2 + \alpha\left\|A - A_I\right\|_1 + \beta\left\|\Phi - \Phi_I\right\|_1, \tag{9.4}$$

where A_I represents the identity affine transformation matrix, Φ_I the spatial gradients of the identity deformation, and α and β are regularization weights, controlling the influence of the regularization terms on the obtained displacements. The higher the values of α and β, the closer the deformation is to the identity. The regularization parameters are essential for the joint optimization, as they ensure that the predicted deformations will be smooth for both components. Moreover, the regularization parameters are very important in the regions of change, as they do not allow the deformations to become very large.

The most commonly employed reconstruction loss is the mean-squared error (MSE). However, MSE suffers from several drawbacks. Firstly, it cannot account for changes in contrast, brightness, tint, etc. Secondly, MSE tends to produce smooth images. Thirdly, MSE does not account for the perceptual information in the image (Wang et al. 2004). Recent papers have hence reported the use of other types of similarity functions, either instead of MSE or in combination with it, to construct more descriptive loss functions. One of this type of losses is the local cross correlation (LCC) presented in Balakrishnan et al. (2019).

9.3.4 Dataset and Implementation Details

To validate the method, a multimodal high-resolution dataset from Quickbird and WorldView-2 satellites was used. This dataset is a multitemporal dataset acquired in 2006 and 2007, covering a 14km^2 region in the East Prefecture of Attica in Greece. This particular dataset was challenging due to the very large size of the high-resolution satellite images, their complexity due to different acquisition shadows, angles, important height differences, numerous terrain objects and the sparse multitemporal acquisitions. To train the framework, patches of size 256×256 were created. In particular, 1350 patches were selected randomly for training, 150 for validation, and 150 for testing the proposed framework. Regions that are spatially independent were selected from the image to generate the training, validation and testing sets of pairs.

Concerning the implementation details of the framework, the initial learning rate was 10^{-3} and was divided by a factor of 10 if the performance on the validation set did not improve for 50 epochs, while the training procedure was stopped when there was no observed improvement for 100 epochs. The regularization weights α and β were both set to 10^{-6}. For the optimization, the Adam optimizer was selected, while the entire framework was implemented in tensorflow and keras. For all the experiments we used a GeForce GTX 1080Ti GPU. We noted that the training converges after around 140 epochs for the dilated architecture and 100 epochs for the maxpooling one. The overall training time was approximately 4 hours.

9.3.5 Experimental Results

Extensive experiments compare and benchmark the performance of the proposed method with other state-of-the-art algorithms that perform both rigid and deformable registration.

In particular, a framework similar to the one presented in Vakalopoulou and Karantzalos (2014) using SIFT, SURF and ASIFT descriptors, RANSAC and an affine transformation is developed to evaluate the performance of the non deep learning-based rigid transformation. Moreover, the method presented in Karantzalos et al. (2014) that applies a graph based method to obtain deformable transformations is also evaluated. Different similarity functions were used for comparison namely the Sum of Absolute Difference (SAD), the Sum of Absolute of Differences plus Gradient Inner Products (SADG), the Normalized Cross Correlation (NCC), and the Normalized Mutual Information (NMI).

Moreover, for the presented completely unsupervised CNN framework we benchmark each of its different components together with two different CNN architectures. In particular, experiments using only the linear A or deformable Φ components and also their ensemble were constructed. This enabled to better examine the framework and find the most optimal configuration for earth observation datasets. The performance with different network architectures are also benchmarked in this chapter.

Starting with the quantitative evaluation, 5 different landmarks, mainly on the buildings' corners have been selected and their average errors in each of the axes are reported in Table 9.2. It should be noted that for all the methods the same landmarks have been selected and around 20 image pairs were used to extract the landmarks. These landmarks contained mainly roofs of buildings as they were the ones presenting the highest registration errors.

Both deep learning-based and classical techniques are evaluated in this study. All the methods recover the geometry of the pairs, achieving better performance than the

Table 9.2 Errors measured as average Euclidean distances between estimated landmark locations. *dx* and *dy* denote distances along *x*-, *y*-, respectively, while *ds* denotes the average error along all axes per pixel.

Method		***dx***	***dy***	***ds***	**Time (sec)**
Unregistered		**7.3**	**6.3**	**9.6**	**–**
Rigid (Vakalopoulou and Karantzalos 2014)	(SIFT)	3.0	2.8	4.1	∼2
	(ASIFT)	4.0	3.5	5.3	∼2.5
	(SURF)	4.7	3.0	5.6	∼3
Deformable (Karantzalos et al. 2014)	(SAD)	1.5	2.7	2.9	∼2
	(SADG)	1.5	2.6	2.8	∼2
	(NCC)	1.3	2.3	2.6	∼2
	(NMI)	1.4	2.4	2.7	∼2
Dilated (Vakalopoulou et al. 2019)	A	2.5	2.8	3.7	∼0.02
	Φ	1.2	2.0	2.3	∼0.02
	A **&Φ**	**0.9**	**1.8**	**1.9**	**∼0.02**
Maxpool (Vakalopoulou et al. 2019)	A	2.6	2.8	3.7	∼0.02
	Φ	1.3	2.1	2.4	∼0.02
	A &Φ	1.0	1.9	2.0	∼0.02

unregistered case; however, the superiority of deformable methods compared to rigid ones is visible in this study, showing the need for more complex deformations for high-resolution imagery. The *ds* error for all the rigid methods is around 4 pixels with the ASIFT and SURF descriptor showing the highest errors. On the other hand, the deformable methods report *ds* errors around 2.5 pixels while the combination of rigid and deformable models reach an error of 2 pixels indicating that the combination of these two methods can boost even more the accuracy of the registration systems. Moreover, the use of different similarity metrics does not affect the performance a lot (Karantzalos et al. 2014); however, one can observe that area-based metrics like NCC and NMI perform slightly better. A similar conclusion was drawn for the influence of the deep neural network architecture in Vakalopoulou et al. (2019). The type of convolution did not really influence the performance of the algorithm, however, we should mention that the architecture with the maxpooling was converging slightly faster than the one with the dilated filters. Finally, it should be noted that the deep learning-based method is faster, with inference time for an image pair of size 256×256 less than half a second in comparison to the $2,3$ seconds that the other methods need, giving a big advantage for large datasets such as the remote sensing ones, and allowing even real-time applications.

Comparing quantitatively the performance of the different methods, three different cases from the test set are presented in Figure 9.3 using checkerboard visualizations between the target R and warped image D before and after the registration. Regions of interest are indicated with red rectangles. For the different architectures in Vakalopoulou et al. (2019) and the different similarity metrics in Karantzalos et al. (2014) there were no differences in the visualizations and therefore the configuration with the lowest reported error is presented, namely the architecture with the dilated filters and the NCC metric, respectively. Even if the initial displacements were quite important all the methods recover the geometry and register the pair of images. However, the proposed method using only the A deformations fails to register accurately high buildings which have the largest deformations, due to the global nature of the transformation. On the other hand, all the deformable based methods report a good performance, registering very accurately the pair of images. One thing that should be mentioned is that the method in Vakalopoulou et al. (2019) reports easier convergence in the case that both the A and Φ parts are trained simultaneously, proving that the additional linear component is a valuable part of the proposed framework.

Moreover, in Figure 9.4 a comparison between the registration performance is provided for the explored methods after the application of the rigid (Vakalopoulou and Karantzalos 2014) using the SIFT descriptor, deformable (Karantzalos et al. 2014) using the NCC metric, and deep learning-based with deformable and affine (Vakalopoulou et al. 2019) methods. Again, the indicated problems are the same, with Vakalopoulou and Karantzalos (2014) failing to recover the local deformations, while the other two methods report very similar performance. In addition, Vakalopoulou et al. (2019) seem not to have a problem creating proper displacement fields for the different sensors, proving its power and potentials. However, experiments with sensors that have higher spectral and spatial differences should be performed.

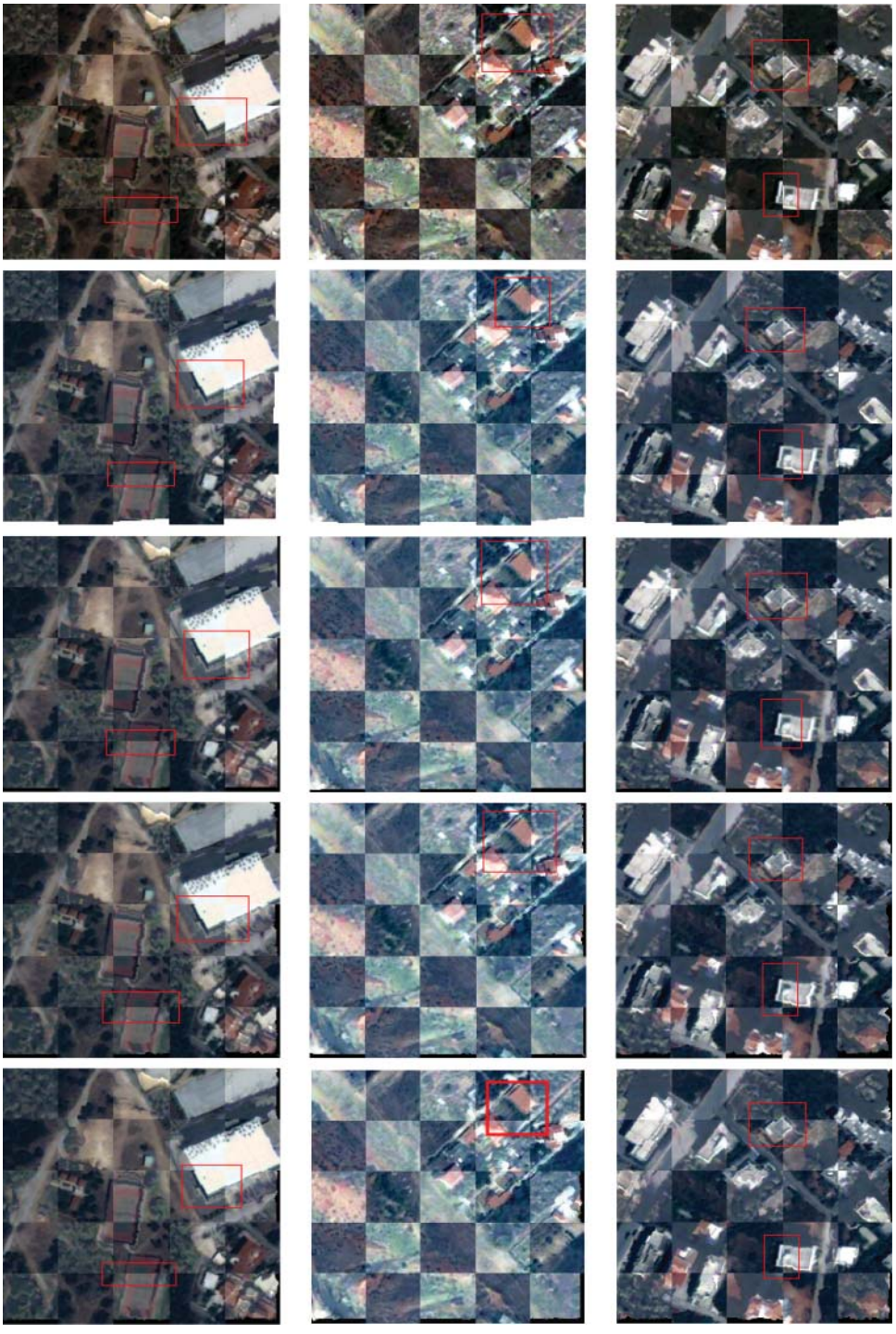

Figure 9.3 Qualitative evaluation for three different pairs of images. From top to bottom: unregistered, (Karantzalos et al. 2014) with NCC, dilated (Vakalopoulou et al. 2019) only A, dilated (Vakalopoulou et al. 2019) only Φ, dilated (Vakalopoulou et al. 2019) A and Φ.

Figure 9.4 Qualitative evaluation for the different methods ((Vakalopoulou and Karantzalos 2014), (Karantzalos et al. 2014), (Vakalopoulou et al. 2019) respectively). From top to bottom: Quickbird 2006 - WorldView-2 2011, Quickbird 2007 - WorldView-2 2011, Quickbird 2009 - WorldView 2011, WorldView-2 2010 - WorldView-2 2011.

9.4 Conclusion and Future Research

Image matching and registration are two problems of utmost importance that have been extensively studied in various communities, including the Earth observation community. A significant amount of research has been devoted to providing the theory and the tools to address these two challenging problems properly. In the last years, both computer vision and earth observation communities have proposed ways to standardize the procedures and create generic methods that can properly address the challenges depending on the problem and the applications. Currently, with the development of deep learning techniques, multiple works in the literature propose approaches based on these techniques. However, these methods are not yet completely exploited for these two problems compared to other problems such as segmentation, classification, and change detection. This indicates the need to focus more towards this direction, especially if one considers that these two problems are very important for a variety of other problems such as image fusion, change detection (Vakalopoulou et al. 2016), 3D reconstruction, or even few shot learning (Sung et al. 2018).

In this chapter, we made an effort to provide a comprehensive survey of the recent advances in both fields focusing on deep learning-based methods. Our approach was structured around the key components of the problems, and in particular we focused on (i) the formulation of the two problems and the main ways proposed in the literature for addressing them, (ii) the presentation of the most recent and important deep learning-based methods that the earth observation community has proposed, and (iii) the comparison and benchmark of different registration methods summarizing their advantage and disadvantages using a challenging high-resolution dataset depicting a peri-urban region. Finally, based on these developments and state-of-the-art methods the present study highlighted certain issues and insights for future research and development.

9.4.1 Challenges and Opportunities

The current challenges of applying the deep learning techniques to image matching and registration of earth observation data are summarized in the following, divided into topics that have the most potential as future directions. To conclude, we believe that deep learning-based methods could provide very good solutions to the image registration problem and the future development in the field should be moved towards addressing the following challenges applied to Earth Observation applications.

9.4.1.1 Dataset with Annotations

Even if deep learning-based methods provide very promising directions for both image matching and registration problems, most of the methods that exist in the literature need annotations to train their models. Especially in the case of multimodal registration, annotations are most of the time essential. However, currently there are no available datasets to generate and evaluate these models, making the proper use and development of these algorithms very slow. It is very important to generate matching and registration datasets for earth observation while the community should also investigate unsupervised techniques

for these two problems. Generative Adversarial Networks can provide very valuable tools towards developments in that direction.

9.4.1.2 Dimensionality of Data

One of the main problems of Earth observation data is their dimensionality both on the spatial and spectral domains, which is also an important difference compared to computer vision datasets. This is also one of the main problems for traditional algorithms to use all the available spectral information and efficiently provide matching or registration results. In particular, for rigid methods the computational time is not significant; however, for deformable methods the calculation of complex transformations is a bottleneck. Nowadays, with deep learning-based approaches the computational time even for deformable methods has considerably decreased, opening new opportunities for the community to design efficient solutions that exploit all the available information. Such methods will also be easily applicable for hyperspectral data, exploiting all their spectral information, something that currently is not easily achievable.

9.4.1.3 Multitemporal Datasets

Moreover, with the recent developments on satellites and the adaptation of open policies for many main space missions access to Earth observation data became easier. Moreover, due to these advances the community currently has access to multitemporal datasets with high temporal resolution that was unavailable earlier. Image matching and registration are the two methods that enable the use of these data for various applications and problems such as land monitoring, damage managing, environmental changes, and many others. There is a need for the community to propose methods and tools that efficiently solve these problems for two or even more images. The idea of group-wise registration has already been proposed in medical imaging (Kornaropoulos et al. 2016) and, with the recent deep learning advantages, it can provide a solution for registration of multitemporal datasets in the same coordinate system in an efficient way. The development of these algorithms will further boost the effectiveness and applicability of earth observation methods in large-scale monitoring of earth and environmental applications.

9.4.1.4 Robustness to Changed Areas

Earth observation datasets and especially the imagery based on optical sensors suffer from high radiometric, atmospheric changes and cloud coverage, making the application of matching and registration techniques really challenging. The use of deep learning-based methods could provide more robust algorithms on these changes as they rely on higher-level representations. Moreover, the development of algorithms that can handle regions with semantic changes would be an interesting direction for the future.