

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Spatiotemporal Fusion in Remote Sensing

Hessah Albanwan and Rongjun Qin

Abstract

Remote sensing images and techniques are powerful tools to investigate earth's surface. Data quality is the key to enhance remote sensing applications and obtaining clear and noise-free set of data is very difficult in most situations due to the varying acquisition (e.g., atmosphere and season), sensor and platform (e.g., satellite angles and sensor characteristics) conditions. With the increasing development of satellites, nowadays Terabytes of remote sensing images can be acquired every day. Therefore, information and data fusion can be particularly important in the remote sensing community. The fusion integrates data from various sources acquired asynchronously for information extraction, analysis, and quality improvement. In this chapter, we aim to discuss the theory of spatiotemporal fusion by investigating previous works, in addition to describing the basic concepts and some of its applications by summarizing our prior and ongoing works.

Keywords: spatiotemporal fusion, satellite images, depth images, pixel-level spatiotemporal fusion, feature-level spatiotemporal fusion, decision-level spatiotemporal fusion

1. Introduction

1.1 Background

Obtaining a high-quality satellite image with a complete representation of earth's surface is crucial to get clear interpretability of data, which can be used for monitoring and managing natural and urban resources. However, because of the internal and external influences of the imaging system and its surrounding environment, the quality of remote sensing data is often insufficient. The internal imaging system conditions include the spectral characteristics, resolution and other factors of the sensor, algorithms used to calibrate the images, etc. The surrounding environment refers to all external/environmental influences such as weather and season. These influences can cause errors and outliers within the images; for instance, shadow and cloud may cause obstructions in the scene and may occlude part of the information regarding an object. These errors must be resolved in order to produce high-quality remote sensing product (e.g., land-cover maps).

With the rapid and increasing development of satellite sensors and their capabilities, studies have shown that fusion of data from multisource, multitemporal images, or both is the key to recover the quality of a satellite image. Image fusion is known as the task of integrating two or more images into a single image [1–3]. The fusion of data essentially utilizes redundant information from multiple images to resolve or

minimize uncertainties associated with the data, with goals such as to reject outliers, to replace and fill missing data points, and to enhance spatial and radiometric resolutions of the data. Fusion has been used in a wide range of remote sensing applications such as radiometric normalization, classification, change detection, etc. In general, there are two types of fusion algorithms: spatial-spectral [4–7] and spatiotemporal fusion [8–10]. Spatial-spectral fusion uses the local information in a single image to predict the pixels' true values based on spectrally similar neighboring pixels. It is used for various types of tasks and applications such as filling missing data (also known as image inpainting) and generating high-resolution images (e.g., pan-sharpening [11] and super-resolution [12]). It can include filtering approaches such as fusing information within a local window using methods such as interpolation [13, 14], maximum a posteriori (MAP), Bayesian model, Markov random fields (MRFs), and Neural Networks (NN) [4, 12, 15–18]. Although spatial-spectral fusion is efficient, it is not able to incorporate information from temporal images, which produce dramatic radiometric differences such as those introduced by meteorological, phenological, or ecological changes. For instance, radiometric distortions and impurities in an image due to meteorological changes (e.g., heavy cloud cover, haze, or shadow) cannot be entirely detected and suppressed by spatial-spectral fusion since it only operates locally within a single image. To address this issue, researchers suggested spatiotemporal fusion, which encompasses spatial-spectral fusion and offers a filtering algorithm that is invariant to dynamic changes over time, in addition to being robust against noise and radiometric variations. Identifying spatiotemporal patterns is the core to spatiotemporal fusion, where the patterns are intended to define a correlation between shape, size, texture, and intensity of adjacent pixels across images taken at different times, of different types, and from different sources.

Spatiotemporal fusion has been an active area of study over the last few decades [9]. Many studies have shown that maximizing the amount of information through integrating the spatial, spectral, and temporal attributes can lead to accurate stable predictions and enhance the final output [8, 9, 19–21]. Spatiotemporal fusion can be applied within local and global fusion frameworks, where locally it can be performed using weighted functions and local windows around all pixels [22–24], and globally using optimization approaches [25, 26]. Additionally, spatiotemporal fusion can be performed on various data processing levels depending on the desired techniques and applications to be used [3]. It also can depend on the type of data used; for instance, per-pixel operations are well suited for images acquired from the same imaging system (i.e., same sensor) since they undergo similar calibration process and minimum spectral differences in terms of having the same number of bands and bandwidth ranges in the spectrum, whereas feature- or decision-level fusion is more flexible and able to handle heterogeneous data such as combining elevation data (e.g., LiDAR) with satellite images [27]. Fusion levels include:

Pixel-level image fusion: This is a direct low-level fusion approach. It involves pixel-to-pixel operation, where the physical information (e.g., intensity values, elevation, thermal values, etc.) associated with each pixel within two or more images is integrated into a single value [2]. It includes methods such as spatial and temporal adaptive reflectance fusion model (STARFM), Spatial and Temporal Reflectance Unmixing Model (STRUM), etc. [22–24].

Feature-level image fusion: It involves extracting and matching distinctive features from two or more overlapping images using methods such as dimensionality reduction like principal component analysis (PCA), linear discriminant analysis (LDA), SIFT, SURF, etc. [2, 28]. Fusion is then performed using the extracted features and the coefficients corresponding to them [2, 29]. Some other common methods that include spatiotemporal fusion on feature-level are sparse representation and deep learning algorithms [10, 30–38].

Decision-level image fusion is a high-level of fusion method that requires each image to be processed individually until an output (e.g., classification map). The outputs are then postprocessed using decision-level fusion techniques [2, 39]. This level of fusion can include the previous two levels of fusion (i.e., per-pixel operations or extracted features) within its operation [40, 41].

In this chapter, we will focus on the concept, methods, and applications of the spatiotemporal-based fusion at all levels of fusion. We will discuss all aspects of spatiotemporal fusion starting from its concepts, preprocessing steps, the approaches, and techniques involved. We will also discuss some examples that apply spatiotemporal fusion for remote sensing applications.

1.2 Contributions

This book chapter introduces the spatiotemporal analysis in fusion algorithms to improve the quality of remote sensing images. We will explore spatiotemporal fusion advantages and limitations, as well as, their applications and associated technicalities under three scenarios:

1. Pixel-level spatiotemporal fusion
2. Feature-level spatiotemporal fusion
3. Decision-level spatiotemporal fusion

1.3 Organization

The organization of this chapter is as follows: Section 2 describes remote sensing data and acquisition and generation processes and necessary preprocessing steps for all fusion levels. Section 3 talks about spatiotemporal fusion techniques under the three levels of fusion: pixel-level, feature-level, and decision-level, which can be applied to either multisource, multitemporal, or multisource multitemporal satellite images. Section 4 describes some applications applying spatiotemporal fusion, and finally Section 5 concludes the chapter.

2. Generic steps to spatiotemporal fusion

Spatiotemporal analysis allows investigation of data from various times and sources. The general workflow for any spatiotemporal fusion process is shown in **Figure 1**. The process description toward a fused image is demonstrated in **Figure 1(a)**, where it describes the process of input acquisition, preprocessing steps, and finally the fusion. Data in remote sensing are either acquired directly from a sensor (e.g., satellite images) or indirectly generated using algorithms (e.g., depth image from dense image matching algorithms [42]) (see **Figure 1(b)**). It also includes data from single or multiple sources (see **Figure 1(b)**); however, combining multisource and multitemporal images requires preprocessing steps to assure data consistency for analyses. The preprocessing steps can include radiometric and geometric correction and alignment (see **Figure 1(a)**). The main spatiotemporal fusion algorithm is then performed using one or more of the three levels of fusion as a base for their method. In this section, we will discuss the most common preprocessing steps in spatiotemporal fusion, as well as, the importance and previous techniques used in spatiotemporal fusion in the three levels of fusion to improve the quality of images and remote sensing applications.

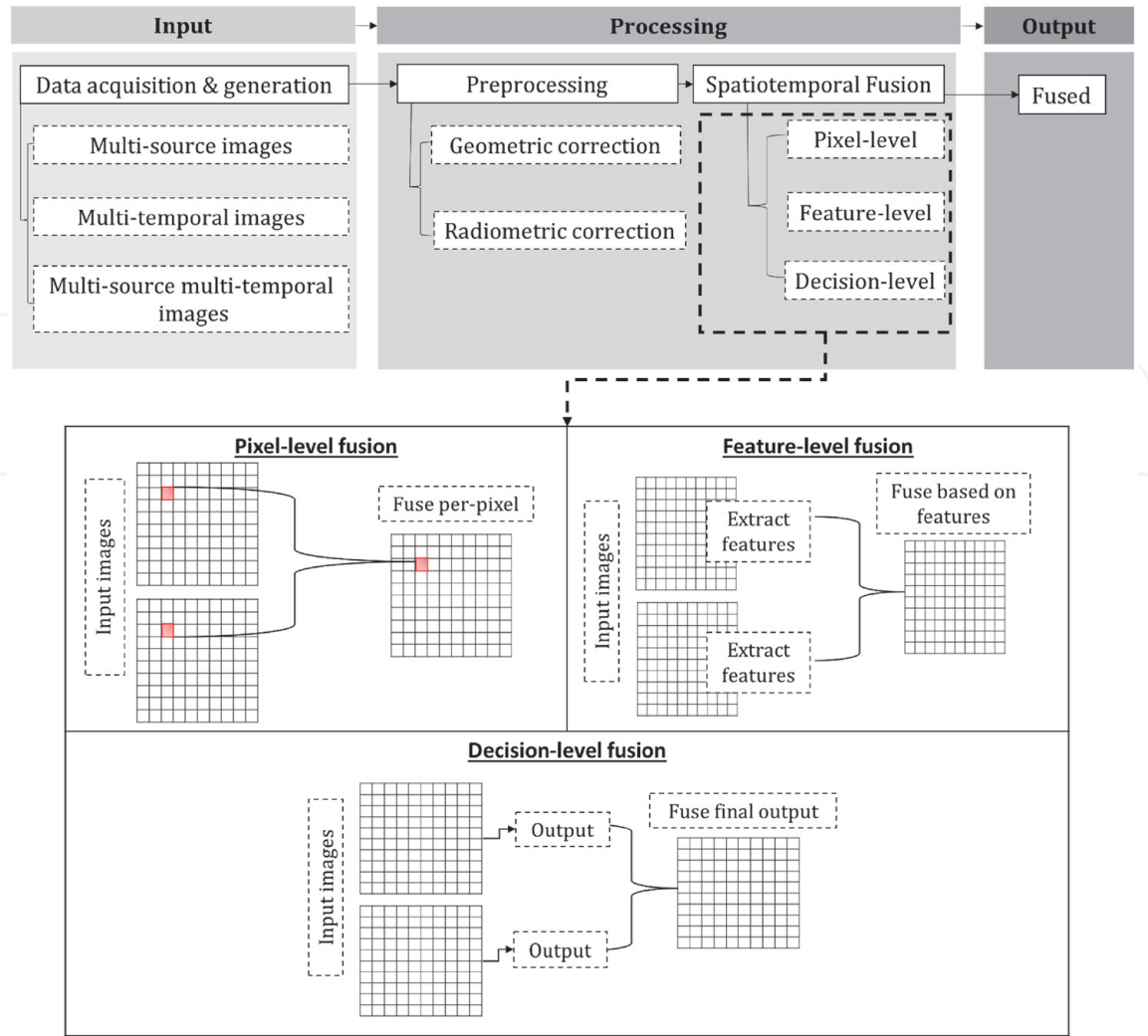


Figure 1. The general workflow for spatiotemporal fusion. (a) The generic steps in spatiotemporal fusion, and (b) fusion based on type of data.

2.1 Data acquisition and generation

Today, there exists a tremendous number of satellite sensors with varying properties and configurations providing researchers with access to a large amount of satellite data. Remote sensing images can be acquired directly from sensors or indirectly using algorithms. It is also available with a wide range of properties and resolutions (i.e., spatial, spectral, and temporal resolutions), which are described in detail in **Table 1**.

2.1.1 Data acquisition

Generally, there exist two types of remote sensing sensor systems: active and passive sensors [43]. **Active sensors** record the signal that is emitted from the sensor itself and received back when it reflects off the surface of the earth. They include sensors like Light Detection and Ranging (LiDAR) and Radar. **Passive sensors** record the reflected signal off the ground after being emitted from a natural light source like the Sun. They include satellite sensors that produce satellite images such as Landsat, Satellite Pour l’Observation de la Terre (SPOT), MODIS, etc.

2.1.2 Data generation

Sometimes in remote sensing, the derived data can be also taken as measurements. Examples include depth images with elevation data derived through

Type of resolution	Spatial resolution	Spectral resolution	Temporal resolution
Definition	Describes the ground area covered by a single pixel in the satellite images. It is also known as the ground sampling distance (GSD) and can range from a few hundreds of meters to sub-meters. Satellite sensors like Moderate Resolution Imaging Spectroradiometer (MODIS) produce coarse-resolution images with 250, 500, and 1000 meters, while fine-resolution images are produced by satellites like very high-resolution (VHR) satellites at the sub-meter level [43].	Refers to the ability of satellite sensors to capture images with wide ranges of the spectrum. It includes hyperspectral (HS) images with thousands of bands or multispectral (MS) images with few numbers of bands (up to 10–15 bands) [43]. It may also include task-specific bands that are beneficial to study the environment and weather, like the thermal band as in Landsat 7 thematic mapper plus (ETM+) [43]. Spectral resolution also refers to the wavelength interval in the spectral signal domain; for instance, MODIS has 36 bands falling between 0.4 and 14.4 μm , whereas Landsat 7 (ETM+) has 7 bands ranging from 0.45 to 0.9 μm .	It is the ability of satellite sensors to capture an object or phenomena in certain periods of time, also known as the revisiting time of sensor at a certain location on the ground. Today, modern satellite systems allow monitoring earth's surface over short and regular periods of time; for instance, MODIS provides almost a daily coverage, while Landsat covers the entire earth surface every 16 days.

Table 1.
Satellite sensors' characteristics and resolutions.

photogrammetric techniques on satellite stereo or multi-stereo images [42], classification maps, change detection maps, etc. In this section, we will discuss two important examples of the commonly fused remote sensing data and their generation algorithms:

2.1.3 *Depth maps (or digital surface model (DSM))*

3D geometric elevation information can either be obtained directly using LiDAR or indirectly using dense image matching algorithms such as Multiview stereo (MVS) algorithms. However, because LiDAR data are expensive and often unavailable for historic data (before 1970s when LiDAR was developed), generating depth images using MVS algorithms is more convenient and efficient. MVS algorithms include several steps:

Images acquisition and selection to perform MVS algorithm requires having at least a pair or more of overlapping images captured from different viewing angles that assure selecting an adequate number of matching features. Specifically, this refers to the process of feature extraction and matching, where unique features are being detected and matched in pairs of images using feature detectors and descriptors methods such as Harris, SIFT, or SURF [44].

Dense image matching and depth map generation: Dense image matching refers to the process of producing dense correspondences between two or among multiple images, and with their pre-calculated geometrical relationship, depth/height information can be determined through ray triangulation [45]. The dense correspondences problem, with pre-calculated image geometry, turns to a 1-D problem in rectified image (also called epipolar image) [46], called disparity computation, which is basically the difference between the left and right views as shown below:

$$\text{Disparity} = \Delta x = x_l - x_r = \frac{f T}{z} \quad (1)$$

where x_l and x_r are distance of pixel in the left and right images accordingly, f is the focal length, T is the distance between the cameras, and z is the depth. The depth (z) is then estimated from Eq. [1] by taking the focal length times the distance between the cameras divided by the disparity as follows:

$$\text{Depth} = Z = \frac{fT}{|x_l - x_r|} \quad (2)$$

In addition, it is noted that assessing and selecting good pairs of images can improve the dense image matching and produce a more accurate and complete 3D depth map [47, 48].

2.1.4 Classification maps

Image classification can be divided into two categories: **1) Supervised classification** is a user-guided process, where classification depends on a prior knowledge about the data that are extracted from the predefined training samples by the user; some popular supervised classification methods include support vector machine (SVM), random forest (RF), decision trees DT, etc. [49–51]. **2)**

Unsupervised classification is a machine-guided process, where the algorithms classify the pixels in the image by grouping similar pixels to come up with specific patterns that define each class. These techniques include segmentation, clustering, nearest neighbor classification, etc. [49].

2.2 Preprocessing steps

2.2.1 Geometric correction

Image registration and alignment is an essential preprocessing step in any remote sensing application that processes two or more images. For accurate analyses of multisource multitemporal images, it is necessary that overlapping pixels in the images correspond to the same coordinates or points on the earth's surface. Registration can be performed manually by selecting control points (CPs) between a pair of images to determine the transformation parameters and warp the images with respect to a reference image [52]. An alternative approach is an automated CP extraction that operates based on mutual information (MI) and similarity measures of the intensity values [52]. According to [53], there are a few common and sequential steps for image registration including the following steps:

Unique feature selection, extraction, and matching refers to the process where unique features are detected using feature extraction methods, then matched to their correspondences in a reference image. A feature can be a shape, texture, intensity of a pixel, edge, or an index such as vegetation and morphological index. According to [54, 55], features can be extracted based on the content of a pixel (e.g., intensity, depth value, or even texture) using methods such as SIFT, difference of Gaussian (DOG), Harris detection, and Histogram of oriented gradient (HOG) [53, 56–58] or based on patch of pixels [59–61] like using deep learning methods (e.g., convolutional neural networks (CNNs)), which can be used to extract complete objects to be used as features.

Transformation refers to the process of computing the transformation parameters (e.g., rotation, translation, scaling, etc.) necessary to convolve an image to a

coordinate system that matches a reference image. The projection and transformation methods include similarity, affine, projective, etc. [53].

Resampling is the process where an image is converted into the same coordinate system as the reference image using the transformation parameters; it includes methods such as interpolation, bilinear, polynomial, etc. [53].

2.2.2 Radiometric correction

Radiometric correction is essential to remove spectral distortion and radiometric inconsistencies between the images. It can be performed either using absolute radiometric normalization (ARN) or relative radiometric normalization (RRN) [62–64]. ARN requires prior knowledge of physical information related to the scene (e.g., weather conditions) for normalization [63, 65–67], while, RRN radiometrically normalizes the images based on a reference image using methods such as dark object subtraction (DOS), histogram matching (HM), simple regression (SR), pseudo-invariant features (PIF), iteratively re-weighted MAD transformation, etc. [62, 64, 68].

3. Data analysis and spatiotemporal fusion

Pixels in remote sensing data are highly correlated over space and time due to earth’s surface characteristics, repeated patterns (i.e., close pixels belong to the same object/class), and dynamics (i.e., season). The general algorithm for spatiotemporal fusion is demonstrated in **Figure 2**, where all levels of fusion follow the same ideology. The minimum image requirement for spatiotemporal fusion is a pair of images whether they are acquired from multiple sources or time, the input images are represented with t_1 to t_n in **Figure 2**. The red square can be either a single

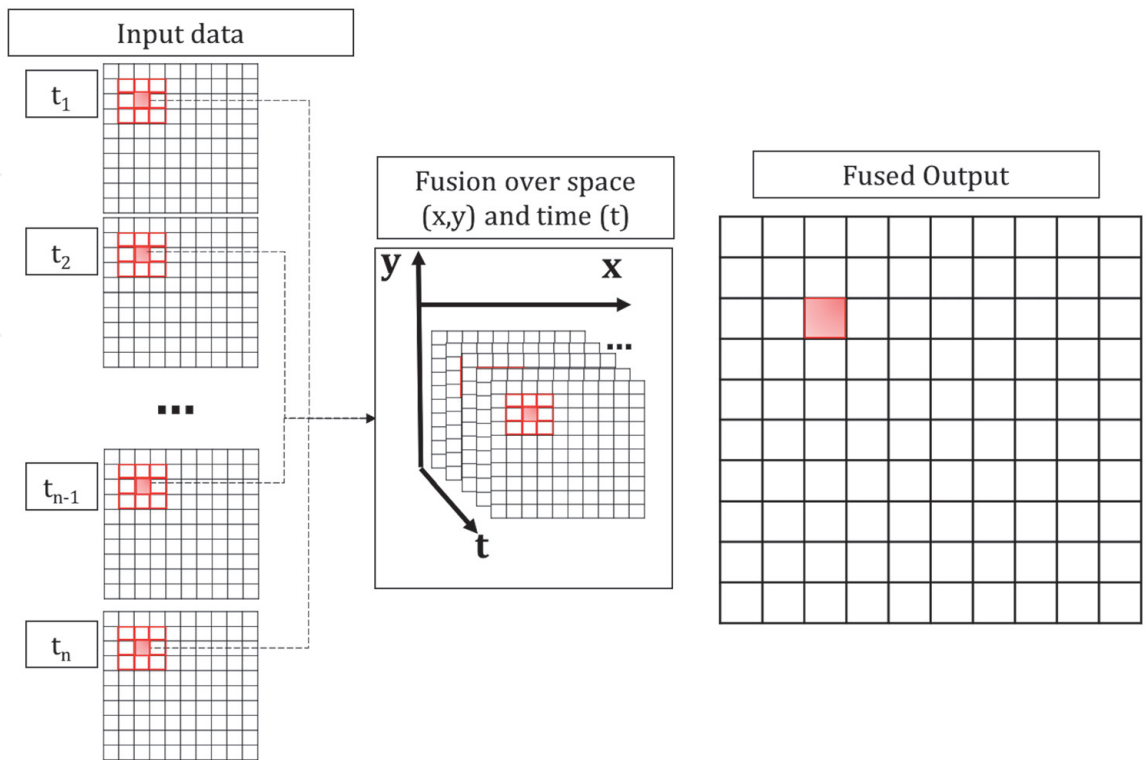


Figure 2.
The general concept of spatiotemporal fusion to process patch of pixels (the red square) spatially across different times (t).

raw pixel, an extracted feature vector, or processed pixel with valuable information (e.g., probability value indicating the class of a pixel). The fusion algorithm then finds the spatiotemporal patterns over space (i.e., the coordinates (x, y) , and pixel content) and time (t) to predict accurate and precise values of the new pixels (see **Figure 2**). In this section, we will provide an overview of some previous works regarding spatiotemporal image fusion that emphasize on the importance of space-time correlation to enhance image quality and discuss this type of fusion in the context of three levels of fusion: pixel-level, feature-level, and decision-level.

3.1 Pixel-level spatiotemporal fusion

As mentioned in the introduction, pixel-based fusion is the most basic and direct approach to fuse multiple images by performing pixel-to-pixel operations; it has been used in a wide range of applications and is preferred because of its simplicity. Many studies performing pixel-level fusion algorithms realized the power of spatiotemporal analysis in fusion and used it in a wide range of applications such as monitoring, assessing, and managing natural sources (e.g., vegetation, cropland, forests, flood, etc.), as well as, urban areas [9]. Most of the pixel-level spatiotemporal fusion algorithms operate as a filtering or weighted-function method; they process a group of pixels in a window surrounding each pixel to compute the corresponding spatial, spectral, and temporal weights (see **Figure 3**). A very popular spatiotemporal fusion method that set the base for many other fusion methods is spatial and temporal adaptive reflectance fusion model (STARFM); it is intended to generate a high-resolution image with precise spectral reflectance by merging multisource fine- and coarse-resolution images [22]. Their method resamples the coarse-resolution MODIS image to have a matching resolution as the Landsat TM image, after that it computes the overall weight by calculating the spectral and temporal differences between the images. STARFM is highly effective in detecting phenological changes, but it fails to handle heterogeneous landscapes with rapid land-cover changes and around mixed pixels [22]. To address this issue, [20] have proposed Enhanced STARFM (ESTARFM); it applies a conversion coefficient to assess the temporal differences between fine- and coarse-resolution images. In [69], Hilker also addressed the problem of drastic land-cover change by proposing Spatial Temporal Adaptive Algorithm for mapping Reflectance Change (STAARCH), which applies Tasseled cap transformation [70] to detect the seasonal changes over

Pixel-level fusion

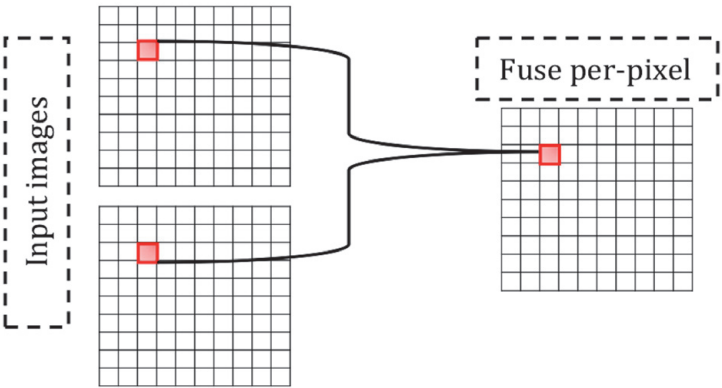


Figure 3.
Pixel-based fusion process diagram.

a landscape. For further improvement of these algorithms, studies have suggested using machine learning methods to identify similar pixels by their classes [71]. The authors also show an example on using machine learning unsupervised classification within the spatiotemporal fusion to enhance its performance. They used clustering on one of the images using the ISODATA method [72], where pixels are considered similar if the difference between the current and central pixel in the window is less than one standard deviation of the pixels in the cluster. Other methods use filtering algorithms to enhance the spatial and spectral aspects of images, in addition to embedding the temporal analysis to further enhance the quality and performance of an application. For instance, [73] proposed a method that combines the basic bilateral filter with STARFM to estimate land surface temperature (LST). In [19], they proposed a 3D spatiotemporal filtering as a preprocessing step for relative radiometric normalization (RRN) to enhance the consistency of temporal images. Their idea revolves around finding the spatial and spectral similarities using a bilateral filter, followed by assessing the temporal similarities for each pixel against the entire set of images. The temporal weight, which assesses the degree of similarity, is computed using an average Euclidean distance using the multitemporal data. In addition to the weighted-based functions, approaches such as unmixing-based and hybrid-based methods are also common in spatiotemporal fusion [74]. The unmixing-based methods predict the fine-resolution image reflectance by computing the mixed pixels from coarse-resolution image [75], while hybrid-based methods use a color mapping function that computes the transformation matrix from the coarse-resolution image and apply it on the finer resolution image [76].

3.2 Feature-level spatiotemporal fusion

Feature-level fusion is a more complex level of fusion, unlike pixel-based operations, it can efficiently handle heterogeneous data that vary in modality and source. According to [2], feature-based fusion can either be conducted directly using semantically equivalent features (e.g., edges) or through probability maps that transform images into semantically equivalent features. This characteristic allows fusion to be performed regardless of the type and source of information [27]. Fusion can then be performed using arithmetic (e.g., addition, division, etc.) and statistical (e.g., mean, median, maximum, etc.) operations; the general process of feature-based fusion is shown in **Figure 4**. The approach in [27] demonstrates a

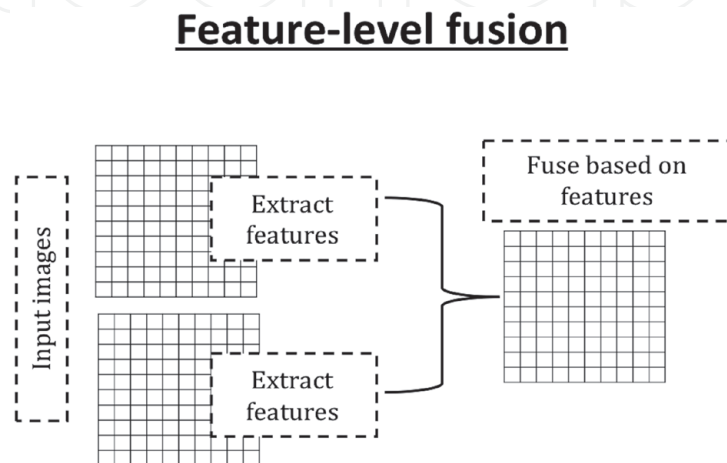


Figure 4.
Feature-based fusion diagram.

simple example on feature-level spatiotemporal fusion to investigate and monitor deforestation; in their method, they combined data from medium-resolution synthetic aperture radar (SAR) and MS Landsat data, they extracted features related to vegetation and soil location (using scattering information and Normalized Difference Fraction Index (NDFI) respectively), finally, fusion was performed through decision tree classifier. Both [26, 62] point out to the most popular methods in feature-level fusion, which include Laplacian pyramid, gradient pyramid, morphological pyramid, high-pass filter, and wavelet transform methods [77–81]. A very famous fusion example in this category is inverse discrete wavelet (IDW) transform, which is a wavelet transform fusion approach; it uses temporal images with varying spatial resolutions to down-sample the coarse-resolution image. It basically extracts the wavelet coefficients from the fine-resolution image and uses them to down-sample the coarse-resolution image [82]. Sparse representation is another widely used learning-based method in feature-level fusion due to its good performance [8, 10, 30, 31]. All sparse representation algorithms share the same concept and core idea, where the general steps include: 1) dividing the input images into patches, 2) extracting distinctive features from the patches (e.g., high-frequency feature patches), 3) generating coefficients from the feature patches, 4) training jointly using dictionaries to find similar structures by extracting and matching feature patches, and finally, 5) fusion using the training information and extracted coefficients [8, 10, 30, 79, 83, 84].

Another state-of-the-art approach in feature- and decision-level fusion is deep learning or artificial neural networks (ANNs). They are currently a very active area of interest in many remote sensing fields (especially image classification) due to their outstanding performance that surpasses traditional methods [32–38, 76, 82–84]. They are also capable of dealing with multi-modality like images from varying sources and heterogeneous data, for instance, super-resolution and pan-sharpening images from different sensors, combining HS and MS images, combining images with SAR or LiDAR data, etc. [32–38]. In feature-level fusion, the ANN is either performed on the images for feature extraction or to learn from the data itself [38]. The extracted features from the temporal images or classification map are used as an input layer, which are then weighted and convoluted within several intermediate hidden layers to result in the final fused image [32–35, 37, 79]. For instance, [85] uses neural networks (CNN) to extract features from RGB image and a DSM elevation map, which are then fed into the SVM training model to generate an enhanced semantic labeling map. ANNs have also been widely used to solve problems related to change detection of bi-temporal images such as comparing multi-resolution images [86] or multisource images [87], which can be solved in a feature-learning representation fashion. For instance, the method in [87] directly compares stacked features extracted from a registered pair of images using deep belief networks (DBNs).

3.3 Decision-level spatiotemporal fusion

The decision-level fusion operates on a product level, where it requires images to be fully and independently processed until the meaningful output is obtained (e.g., classification or change detection maps) (see **Figure 5**). Decision-level fusion can adapt to different modularities like combining heterogeneous data such as satellite and depth images, which can be processed to common outputs (e.g., full/partial classification maps) for fusion [88]. Additionally, the techniques followed by this fusion type are often performed under the umbrella of Boolean or statistical operations using methods like likelihood estimation, voting (e.g., majority voting, Dempster-Shafer's estimation, fuzzy Logic, weighted sum, etc.) [88–90]. In [88],

Decision-level fusion

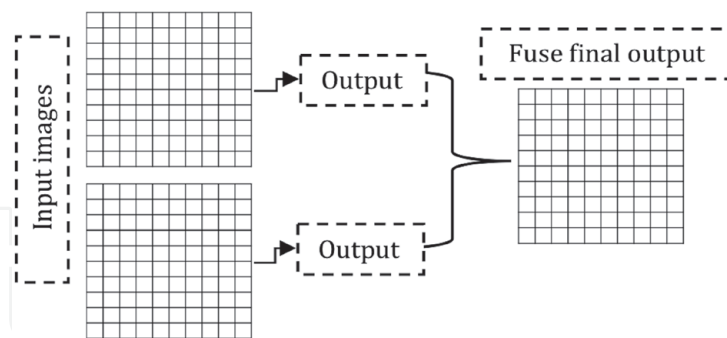


Figure 5.
Feature-based fusion diagram.

they provide an example on the mechanism of decision-level fusion; they developed a fusion approach to detect cracks and defects on ground surface, they first convert multitemporal images into spatial density maps using kernel density estimation (KDE), then, fused the pixels density values using a likelihood estimation method. In general, most of the decision-level fusion techniques rely on probabilistic methods, where they require generating an initial label map with each pixel upholding a probability value and indicating its belonging to a certain class, which can be generated using traditional classification methods like the supervised (e.g., random forest) or unsupervised (e.g., clustering or segmentation) classification (see Section 2.1.2.). Another advantage of the decision-level fusion is that it can be implemented while incorporating both levels of fusion, the pixel- and feature-level fusion. The method in [41] shows a spatiotemporal fusion algorithm that includes all levels of fusion, where they propose a post-classification refinement algorithm to enhance the classification maps. First, they generate probability maps for all temporal images using random forest classifier (as an initial classification map); then they use a recursive approach to iteratively process every pixel in the probability maps by fusing the multitemporal probability maps with the elevation from the DSMs using a 3D spatiotemporal filtering. Similarly, [40] have also proposed fusion of probability maps for building detection purposes, where they first generate the probability maps, then fuse them using a simple 3D bilateral filter.

Recently, more focus has been driven toward using spatiotemporal fusion to recover the quality of 3D depth images generated from MVS (e.g., DSM fusion). Median filtering is the oldest and most common fusion approach for depth images; it operates by computing the median depth of each pixel from a group of pixels at the same location in the temporal images [91]. The median filtering is robust to outliers and is efficient in filling missing depth values. However, the median filter only exploits the temporal domain; to further enhance its performance and the precision of the depth values, studies suggest spatiotemporal median filtering. In [92], the authors have proposed an adaptive median filtering that operates based on the class of the pixels; they use an adaptive window to isolate pixels belonging to the same class, then choose the median pixel based on the location (i.e., adaptive window) and temporal images. In [93], the authors also show that spatiotemporal median filtering can be improved by adopting an adaptive weighing-filtering function that involves assessing the uncertainty of each class in the spatial and temporal domains in the depth images using standard deviation. The uncertainty will then be used as the bandwidth parameter to filter each class individually. The authors in [47] also suggested a per-pixel fusing technique to select the depth value for each

pixel by using a recursive K-median clustering approach that generates one to eight clusters until it reaches the desired precision.

Other complex yet efficient methods used in decision-level fusion are deep learning algorithms as mentioned previously in Section 2.3.2. [94]. They are either used as postprocessing refinement approaches or to learn end-to-end from a model [38]. For example, the method in [95] used a postprocessing enhancement step for semantic labeling, where they first generate probability maps using two different methods, RF and CNN using multimodal data (i.e., images and depth images), then they fused the probability maps using Conditional random fields (CRFs) as postprocessing approach. In [96], on the other hand, the authors used a model learning-based method, where they first semantically segment multisource data (i.e., image and depth image) using a SegNet network, then fuse their scores using a residual learning approach.

4. Examples on spatiotemporal fusion applications

4.1 Spatiotemporal fusion of 2D images

4.1.1 Background and objective

A 3D spatial-temporal filtering algorithm is proposed in [19] to achieve relative radiometric normalization (RRN) by fusing information from multitemporal images. RRN is an important preprocessing step in any remote sensing application that requires image comparison (e.g., change detection) or matching (e.g., image mosaic, 3D reconstruction, etc.). RRN is intended to enhance the radiometric consistency across set of images, in addition to reducing radiometric distortions that result due to sensor and acquisition conditions (as mentioned in Section 1.1.). Traditional RRN methods use a single reference image to radiometrically normalize the rest of the images. The quality of the normalized images highly depends on the reference image, which requires the reference image to be noise-free or to have minimum radiometric distortions. Thus, the objective of [19] is to generate high-quality radiometrically consistent images with minimum distortion by developing an algorithm that fuses the spatial, spectral, and temporal information across a set of images.

4.1.2 Theory

The core of the 3D spatiotemporal filter is based on the bilateral filter, which is used to preserve the spectral and spatial details. It is a weighting function that applies pixel-level fusion on images from multiple dates (see **Figure 4(a)**). The general form of this filter is as follows:

$$\bar{I}_i = \int_{\Omega} w_{j,i} \cdot I_j \cdot dj \quad (3)$$

where the original and filtered images are indicated using I and \bar{I} . The weight for every pixel at point j into the fused pixel i is indicated using $w_{j,i}$. The filtering is carried out on the entire space of the set of images Ω including all domains, that is, the spatial (i.e., pixels' coordinates (x, y)), the spectral (i.e., intensity value), and temporal (i.e., intensity of temporal images). The spatial and spectral weights are described by [97] and are indicated in Eqs. (4) and (5) respectively

$$w_{\text{spatial}} = \exp \left(\frac{|j_x - i_x|^2}{\sigma_x} + \frac{|j_y - i_y|^2}{\sigma_x} \right), j, i \in \Omega \quad (4)$$

$$w_{\text{spectral}} = \exp \left(-\frac{|I_j - I_i|^2}{\sigma_I} \right), j, i \in \Omega \quad (5)$$

where, I is the pixel value at x and y locations, and σ_x and σ_I are the spatial and spectral bandwidths respectively that set the degree of filtering based on the spatial and spectral similarities between the central pixel and nearby pixels. The novelty of this filter is in the design of the temporal weight, where it computes the resemblance between every image and the entire set of images using an average Euclidean distance as the follows

$$w_{\text{spectral}} = \exp \left(-\frac{|j_t - i_t|^2}{\sigma_T} \right), j, i \in \Omega \quad (6)$$

where $(j_t - i_t)$ are the difference between the current image being processed and all other images and σ_T is the degree of filtering along the temporal direction. Eq. (6) allows all images to contribute toward each other in enhancing the overall radiometric characteristics and consistency without requiring a reference image for the RRN process.

4.1.3 Experimental results and analysis

The 3D spatial-temporal filter was conducted on three experiments with varying resolutions and complexities. Experiments 1 and 2 were applied on urban and sub-urban areas respectively; each experiment had five medium-resolution images from Landsat 8 satellite (with 15- to 30- m spatial resolution). Experiment 3 was on a fine-resolution image from Planet satellite (with 3-m spatial resolution).

Figure 6(b) and **(c)** shows an example of the input and results of the filter using the data from experiment 1 (i.e., the urban area). The input images show a significant discrepancy in the radiometric appearance (see **Figure 6(b)**); however, the heterogeneity between multitemporal images is reduced after the filtering process (see **Figure 6(c)**). By comparing the original and filtered images in **Figure 6(c)**, we can notice that the land covers are more similar in the filtered images than in the original images. For instance, the water surface (shown in **Figure 6(c)** in blue bold dashed line) used to have a clear contrast in intensity in the original images, but after the filtering process, they become more spectrally alike in terms of intensity looks and ranges.

The experiments are also validated numerically using transfer learning classification (using SVM) to test the consistency between the normalized filtered images. The transfer learning classification uses a reference training data from one image and applies it to the rest of the images. The results in **Table 2** indicate that the filtered images have higher accuracy than the nonfiltered original images, where the average improvement in accuracy is ~6%, 19%, and 2% in all three experiments respectively. Reducing the uncertainty in the filtering process by not requiring a reference image for normalization was the key to this algorithm. The algorithm was formulated to take advantage of the temporal direction by treating all images in the

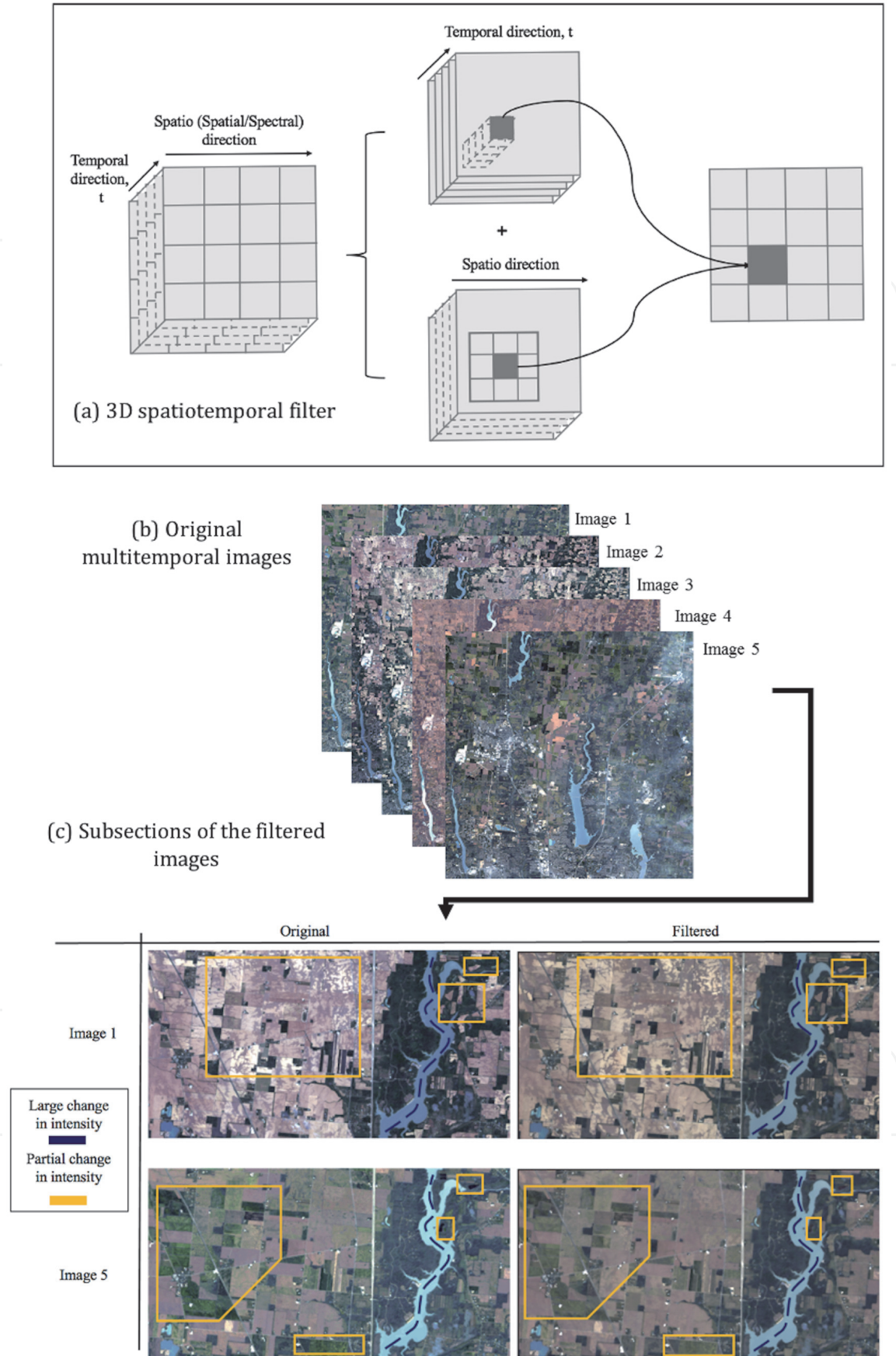


Figure 6. Pixel-level fusion using 3D spatiotemporal bilateral filter to combine multitemporal images [19].

dataset as a reference. Therefore, it will have higher confidence to distinguish between actual objects and radiometric distortions (like clouds) in the scene when processing each pixel.

Transfer learning classification					
Image	1	2	3	4	5
Exp. I Suburban					
Without filter	80.88	74.14	93.30	93.59	91.97
With filter	91.99	91.96	93.95	86.08	94.30
Exp. I - Urban					
Without filter	72.61	67.91	81.00	50.21	93.75
With filter	89.29	90.60	91.35	77.82	93.10
Exp. II					
Without filter	66.48	74.14	68.35	67.17	73.30
With filter	66.75	76.20	72.06	65.10	78.54

The bold numbers indicate an increase in the accuracy, and the numbers highlighted in gray indicate the reference image used for the training in the transfer learning classification.

Table 2.
The accuracy results for the 3D spatial-temporal filter [19].

4.2 Spatiotemporal fusion of multisource multitemporal images

4.2.1 Background and objective

Multitemporal and multisource satellite images often generate inconsistent classification maps. Noise and misclassifications are inevitable when classifying satellite images, and the precision and accuracy of classification maps vary based on the radiometric quality of the images. The radiometric quality is a function of the acquisition and sensor conditions as mentioned in the background in Section 1.1. The algorithm can also play a major role in the accuracy of the results; some classification algorithms are more efficient than others, while some can be sensitive to the spatial details in the images like complex dense areas and repeated patterns, which lead objects of different classes to have similar spectral reflectance. The acquisition time, type of algorithm, and distribution of objects in the scene are huge factors that can degrade the quality and generate inconsistent classification maps across different times. To address these issues, the authors in [41] proposed a 3D iterative spatiotemporal filtering to enhance the classification maps of multitemporal very high-resolution satellite images. Since the 3D geometric information is more stable and is invariant to spectral changes across temporal images, [41] proposed combining the 3D geometric information in the DSM with multitemporal classification maps to provide spectrally invariant algorithm.

4.2.2 Theory

The 3D iterative spatiotemporal filter is a fusion method that combines information from various types, sources, and times. The algorithm is a combination of feature and decision levels of fusion; it is described in detail in Algorithm 1. The first step is to generate initial probability maps for all images using random forest classification. The inference model is then built to recursively process every pixel in the probability maps using a normalized weighing function that computes the total weight $W_{3D}(x_j, y_j, t_n)$ based on the spatial ($W_{spatial}$), spectral ($W_{spectral}$), and temporal ($W_{temporal}$) similarities. The temporal weight is based on the elevation values in the DSMs. The probability value for every pixel is computed and updated

using $W_{3D}(x_j, y_j, t_n)$ and the previous iteration until it satisfies the convergence condition, which requires the difference between the current and previous iterations to be under a certain limit.

Algorithm 1: Pseudo code of the proposed 3D iterative spatiotemporal filter [41]

Input: Initial probability maps $P_c^0(x_i, y_i, t_n)$, ortho photos I , and the band widths: σ_s and σ_r Output: Final probability maps $P_c^f(x_i, y_i, t_n)$
For every category/class c do <div style="margin-left: 20px;"> While not converge do <div style="margin-left: 20px;"> For every pixel (x, y) in the window w do <div style="margin-left: 20px;"> $W_{spatial} \rightarrow \exp(\frac{\ x_i - x_j\ ^2 + \ y_i - y_j\ ^2}{2\sigma_s^2})$ $W_{spectral} \rightarrow \exp(\frac{\ I(x_i, y_i) - I(x_j, y_j)\ ^2}{2\sigma_r^2})$ Compute σ_h for class c $W_{nDSM} \rightarrow \exp(\frac{\ nDSM(x_i, y_i, t_m) - nDSM(x_j, y_j, t_n)\ ^2}{2\sigma_h^2})$ $W_{3D} = W_{spatial} * W_{spectral} * W_{nDSM}$ Update the probability distribution map $P_c^k(x_i, y_i, t_n) = \frac{1}{N * T} \sum \sum_{j \in N, n \in T} W_{3D}(x_j, y_j, t_n) * P_c^{k-1}(x_j, y_j, t_n)$ </div> </div> </div>

End For
Check convergence

$$\frac{P_c^k(x_i, y_i, t_n) - P_c^{k-1}(x_i, y_i, t_n)}{P_c^k(x_i, y_i, t_n)} * 100\% \rightarrow \begin{cases} \leq 5\% & \text{Stop} \\ > 5\% & \text{Continue} \end{cases}$$
End While
End For

4.2.3 Experimental results and analysis

The proposed filter was applied to three datasets that include an open area, residential area, and school area. The input data include multisource and multitemporal very high-resolution images and DSMs; the probability maps were created for six types of classes: buildings, long-term or temporary lodges, trees, grass, ground, and roads (see **Figure 7(a)** for more details about the input data). **Figure 7(b)** shows a sample of the filtering results. We can see that the initial classification of the building (circled with an ellipse) is mostly incorrectly classified to long-term lodge; however, it keeps improving as the filtering proceeds through the iterations.

The overall accuracy was reported, and it indicates that the overall enhancement in the accuracy is about $\sim 2\text{--}6\%$ (see **Table 3**). We can also notice that dense areas such as the residential area have the lowest accuracy range (around 85%), while the rest of the study areas had accuracy improvement in the 90% range. It indicates that the filtering algorithm is dependent on the degree of density and complexity in the

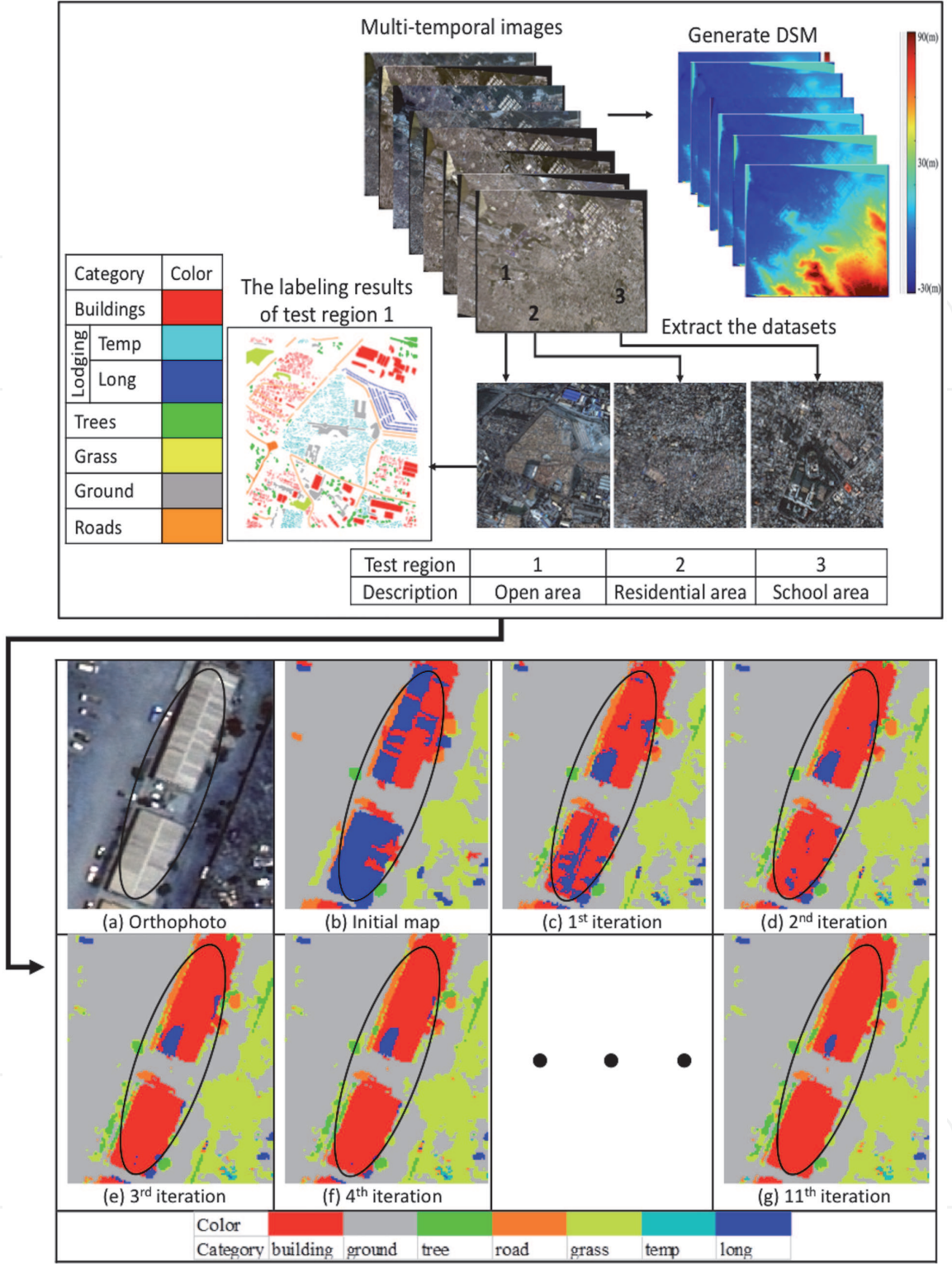


Figure 7.
3D iterative spatiotemporal filtering for classification enhancement [41].

scene, where objects are hard to distinguish in condensed areas due to mixed pixel and spectral similarity of different objects.

4.3 Spatiotemporal fusion of 3D depth maps

4.3.1 Background and objective

Obtaining high-quality depth images (also known as depth maps) is essential for remote sensing applications that process 3D geometric information like 3D

Date	Test region 1			Test region 2			Test region 3		
	Before (%)	After (%)	Δ (%)	Before (%)	After (%)	Δ (%)	Before (%)	After (%)	Δ (%)
2007	91.04%	95.21	+4.17	83.47	88.14	+4.67	91.12	95.85	+4.73
2010/1	93.21	96.45	+3.24	81.50	85.67	+4.17	93.06	96.82	+3.76
2010/6	91.93	96.26	+4.33	83.52	89.79	+6.27	88.82	94.87	+6.05
2010/12	89.08	95.57	+6.49	80.81	87.59	+6.78	88.58	94.86	+6.28
2012/3	92.19	95.92	+3.73	81.43	86.92	+5.49	91.44	97.08	+5.64
2013/9	90.40	96.56	+6.16	81.03	87.29	+6.26	94.99	97.54	+2.55
2014/7	95.11	97.27	+2.16	82.19	88.90	+6.71	90.39	96.58	+6.19
2015	92.74	96.35	+3.61	83.22	85.69	+2.47	94.61	97.19	+2.58
Average	92.09	96.20	4.24	82.15	87.50	5.29	91.63	96.35	4.72

Table 3.
The overall accuracy for classification results using the method in [41].

reconstruction. MVS algorithms are widely used approaches to obtain depth images (see Section 2.1.2.); however, depth maps generated using MVS often contain noise, outliers, and incomplete representation of depth like having missing data, holes, or fuzzy edges and boundaries. A common approach to recover the depth map is by fusing several depth maps through probabilistic or deterministic methods. However, most fusion techniques in image processing focus on the fusion of depth images from Kinect or video scenes, which cannot be directly applied on depth generated from satellite images due to the nature of images. The difference between depth generated from satellite sensors and Kinect or video cameras include:

1. Images captured indoor using Kinect or video cameras have less noise, since they are not exposed to external environmental influences like atmospheric effects.
2. Kinect or video cameras generate a large volume of images, which can improve dense matching, while the number of satellite images is limited due to the temporal resolution of the satellite sensor.
3. The depth from satellite images is highly sensitive to the constant changes in the environment and the spatial characteristics of the earth surface like the repeated patterns, complexity, sparsity, and density of objects in the scene, which can obstruct or create mismatching errors in the dense image matching process.

Most depth fusion algorithms for geospatial data focus on median filtering (see Section 4.3.), but it still needs some improvement in terms of robustness and adaptivity to the scene content. To address the aforementioned problems, [90] proposed an adaptive and semantic-guided spatiotemporal filtering algorithm to generate a single depth map with high precision. The adaptivity is implemented to address the issue of varied uncertainty for objects of different classes.

4.3.2 Theory

The adaptive and semantic-guided spatiotemporal filter is a pixel-based fusion method, where the depth of the fused pixel is inferred using multitemporal depths and a prior knowledge about the pixel class and uncertainty. A reference orthophoto

is classified using a rule-based classification approach that uses normalized DSM (nDSM) with indices such as normalized difference vegetation index (NDVI). The uncertainty is then measured for all four classes (trees, grass, buildings, and ground and roads) using the standard deviation. The uncertainty is measured spatially using the classification map and also across the temporal images. The adaptive and semantic-guided spatiotemporal filter is intended to enhance the median filter, thus it uses height $h(i,j,t)_{med}$ as the base to the fused pixel, where the general form of the filter is expressed as

$$DSM_f(i,j) = \frac{1}{W_T} * \sum_{i=1}^{Width} \sum_{j=1}^{Height} W_r * W_s * W_h * h(i,j,t)_{med} \tag{7}$$

where DSM_f is the fused pixel; i, j are the pixel's coordinates; h_{med} is the median height value from the temporal DSMs; and the spectral, spatial, and temporal height weights are expressed as W_r , W_s , and W_h respectively. The W_r and W_s are described in Eqs. (4) and (5) that measure the spectral and spatial components from the orthophoto. The W_h is a measure of similarity for the height data across temporal images, and it can be computed using the following formula:

$$W_h(i,j) = \exp \frac{-||h_{med}-h(i,j,t)||^2}{2 \sigma_h^2} \tag{8}$$

where σ_h is the adaptive height bandwidth, which varies based on the class of pixel as follows:

$$\sigma_h = \begin{cases} \sigma_{Building} \rightarrow \text{if pixel } (i,j) \text{ is building} \\ \sigma_{Ground/road} \rightarrow \text{if pixel } (i,j) \text{ is ground/road} \\ \sigma_{tree} \rightarrow \text{if pixel } (i,j) \text{ is tree} \\ \sigma_{grass} \rightarrow \text{if pixel } (i,j) \text{ is grass} \\ \sigma_{water} \rightarrow \text{if pixel } (i,j) \text{ is water} \end{cases} \tag{9}$$

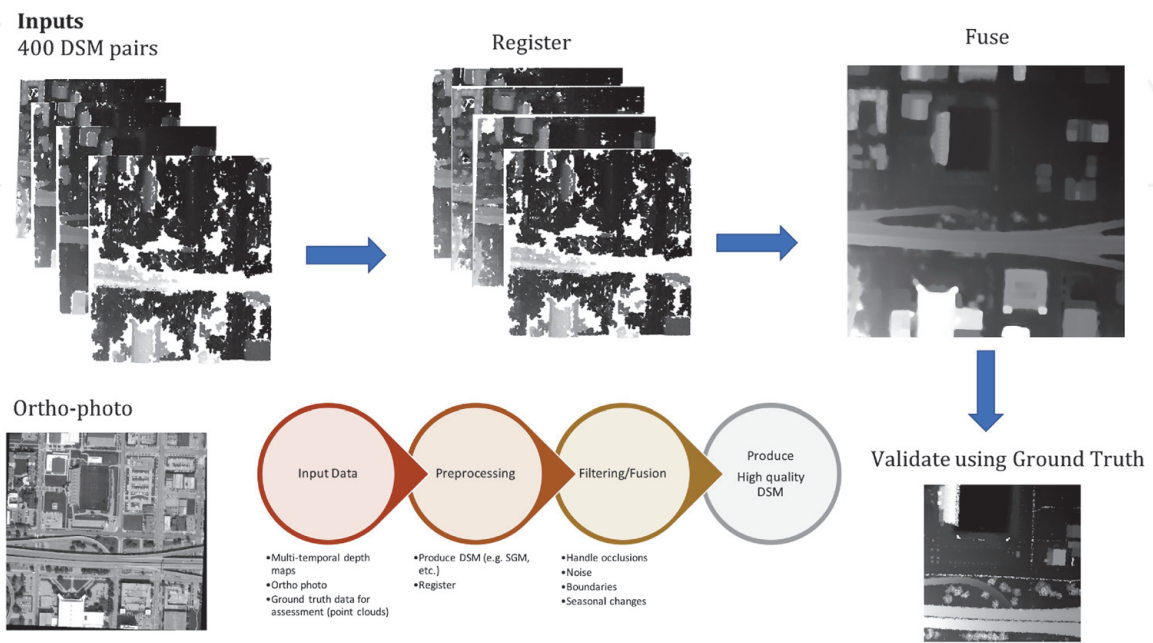


Figure 8.
Process description of adaptive and semantic-guided spatiotemporal filtering [93].

4.3.3 Experimental results and analysis

The method in [90] was experimented on three datasets with varying complexities. The satellite images are taken from the World-View III sensor, and depth is generated using MVS algorithm on every image pair using RSP (RPC Stereo Processor) software developed by [95] and semi-global matching (SGM) algorithm [42]. **Figure 8** describes the procedures followed by the fusion algorithm, in addition to the visual results where it shows that noise and missing elevation points were recovered in the fused image. The validation of three experiments shows that this fusion technique can achieve up to 2% increase in the overall accuracy of the depth map.

5. Conclusions

Spatiotemporal fusion is one of the powerful techniques to enhance the quality of remote sensing data, hence, the performance of its applications. Recently, it has been drawing great attention in many fields, due to its capability to analyze and relate the space-interaction on ground, which can lead to promising results in terms of stability, precision, and accuracy. The redundant temporal information is useful to develop a time-invariant fusion algorithm that leads to the same inference from the multitemporal geospatial data regardless of the noise and changes that occur occasionally due to natural (e.g., metrology, ecology, and phenology) or instrumental (e.g., sensor conditions) causes. Therefore, incorporating spatiotemporal analysis in any of the three levels of fusion can boost their performance, where it can be flexible to handle data from multiple sources, types, and times. Despite the effectiveness of spatiotemporal fusion, there are still some issues that may affect the precision and accuracy of the final output. These considerations must be taken into account while designing the spatiotemporal fusion algorithm. For example, spatiotemporal analysis for per-pixel operations is highly sensitive to mixed pixels especially for coarse-resolution images where one pixel may contain the spectral information of more than one object. The accuracy of the spatiotemporal fusion can also be sensitive to the complexity of the scene, where in densely congested areas such as cities the accuracy may be less than open areas or sub-urban areas (as mentioned in the examples in Section 4.). This is due to the increase in the heterogeneity of the images in these dense areas. This issue can be solved using adaptive spatiotemporal fusion algorithms, which is a not widely investigated area of study in current practices. Feature and decision levels of fusion can partially solve this problem by learning from patches of features or classified images, but their accuracy will also be under the influence of the feature extraction algorithm or the algorithm to derive the initial output. For instance, mismatching features can result in fusing unrelated features or data points, thus produce inaccurate coefficients for the feature-level fusion model. Another observation is the lack of studies that relates the number of temporal images and the fusion output accuracy, which is useful to decide the optimal number of input images for fusion. Additionally, it is rarely seen that the integrated images are picked before fusion, where assessing and choosing good images can lead to better results. Spatiotemporal fusion algorithms are either local or global approaches, the local algorithms are simple and forward like pixel-level fusion or local filtering like the methods in [19, 22], while global methods tend to perform extensive operations for optimization purposes like in [25]. In future works, we aim to explore how these explicitly modeled spatiotemporal fusion algorithms can be enhanced by the power of more complex and inherent models such as deep learning-based models to drive more important remote sensing applications.

Acknowledgments

The authors would like to express their gratitude to Planet Co. for providing them with the data; to sustainable institute at the Ohio state university, Office of Naval Research (Award No. N000141712928) for partial support of the research; and to the Johns Hopkins University Applied Physics Laboratory and IARPA and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for making the benchmark satellite images available.

Author details


Hessah Albanwan¹ and Rongjun Qin^{1,2*}

¹ Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio, USA

² Department of Electrical and Computer Engineering, The Ohio State University, USA

*Address all correspondence to: qin.324@osu.edu

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. Distributed under the terms of the Creative Commons Attribution - NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited. 

References

- [1] Li SZ, Jain AK. Encyclopedia of Biometrics. 1st ed. Boston, MA: Springer US; 2015. DOI: 10.1007/978-1-4899-7488-4 [Accessed: 31 March 2020]
- [2] Mitchell HB. Image Fusion. Berlin, Heidelberg: Springer Berlin Heidelberg. Epub ahead of print; 2010. DOI: 10.1007/978-3-642-11216-4
- [3] Pohl C, Genderen JLV. Review article multisensor image fusion in remote sensing: Concepts, methods and applications. *International Journal of Remote Sensing*. 1998;**19**:823-854
- [4] Dian R, Li S, Fang L, et al. Multispectral and hyperspectral image fusion with spatial-spectral sparse representation. *Information Fusion*. 2019;**49**:262-270
- [5] Fauvel M, Tarabalka Y, Benediktsson JA, et al. Advances in spectral-spatial classification of Hyperspectral images. *Proceedings of the IEEE*. 2013;**101**:652-675
- [6] Wang L, Zhang J, Liu P, et al. Spectral-spatial multi-feature-based deep learning for hyperspectral remote sensing image classification. *Soft Computing*. 2017;**21**:213-221
- [7] Kang X, Li S, Benediktsson JA. Spectral-spatial Hyperspectral image classification with edge-preserving filtering. *IEEE Transactions on Geoscience and Remote Sensing*. 2014;**52**:2666-2677
- [8] Chen B, Huang B, Xu B. A hierarchical spatiotemporal adaptive fusion model using one image pair. *The International Journal of Digital Earth*. 2017;**10**:639-655
- [9] Chen B, Huang B, Bing X. Comparison of spatiotemporal fusion models: A review. *Remote Sensing*. 2015;**7**:1798-1835
- [10] Song H, Huang B. Spatiotemporal satellite image fusion through one-pair image learning. *IEEE Transactions on Geoscience and Remote Sensing*. 2013;**51**:1883-1896
- [11] Ehlers M, Klonus S, Johan Åstrand P, et al. Multi-sensor image fusion for pansharpening in remote sensing. *International Journal of Image and Data Fusion*. 2010;**1**:25-45
- [12] Shen H, Ng MK, Li P, et al. Super-resolution reconstruction algorithm to MODIS remote sensing images. *The Computer Journal*. 2008;**52**:90-100
- [13] Nguyen H, Cressie N, Braverman A. Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*. 2012;**107**:1004-1018
- [14] Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting. In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*. ACM Press; 2000. pp. 417-424
- [15] Kasetkasem T, Arora M, Varshney P. Super-resolution land cover mapping using a Markov random field based approach. *Remote Sensing of Environment*. 2005;**96**:302-314
- [16] Pathak D, Krahenbuhl P, Donahue J, et al. Context Encoders: Feature Learning by Inpainting. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. NV, USA: Las Vegas; 2016, pp. 2536-2544
- [17] Eismann MT, Hardie RC. Application of the stochastic mixing model to hyperspectral resolution enhancement. *IEEE Transactions on Geoscience and Remote Sensing*. 2004;**42**:1924-1933
- [18] Wei Q, Dobigeon N, Tourneret J-Y. Bayesian fusion of hyperspectral and

multispectral images. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE; 2014. pp. 3176-3180

[19] Albanwan H, Qin R. A novel spectrum enhancement technique for multi-temporal, multi-spectral data using spatial-temporal filtering. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2018;**142**:51-63

[20] Zhu X, Chen J, Gao F, et al. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment*. 2010;**114**:2610-2623

[21] Gómez C, White JC, Wulder MA. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2016;**116**:55-72

[22] Gao F, Masek J, Schwaller M, et al. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*. 2006;**44**:2207-2218

[23] Gevaert CM, García-Haro FJ. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sensing of Environment*. 2015;**156**:34-44

[24] Jia K, Liang S, Zhang N, et al. Land cover classification of finer resolution remote sensing data integrating temporal features from time series coarser resolution data. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2014;**93**:49-55

[25] Tang Q, Bo Y, Zhu Y. Spatiotemporal fusion of multiple-satellite aerosol optical depth (AOD) products using Bayesian maximum entropy method: merging satellite AOD products using BME. *Journal of*

Geophysical Research-Atmospheres. 2016;**121**:4034-4048

[26] Melgani F, Serpico SB. A markov random field approach to spatio-temporal contextual image classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2003; **41**:2478-2487

[27] Reiche J, Souza CM, Hoekman DH, et al. Feature level fusion of multi-temporal ALOS PALSAR and Landsat data for mapping and monitoring of tropical deforestation and Forest degradation. *The IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2013; **6**:2159-2173

[28] Ross A. Fusion, feature-level. In: Li SZ, Jain AK, editors. *Encyclopedia of Biometrics*. Boston, MA: Springer US; 2015. pp. 751-757

[29] Sasikala M, Kumaravel N. A comparative analysis of feature based image fusion methods. *Information Technology Journal*. 2007;**6**:1224-1230

[30] Huang B, Song H. Spatiotemporal reflectance fusion via sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*. 2012; **50**:3707-3716

[31] Wu B, Huang B, Zhang L. An error-bound-regularized sparse coding for spatiotemporal reflectance fusion. *IEEE Transactions on Geoscience and Remote Sensing*. 2015;**53**:6791-6803

[32] Palsson F, Sveinsson JR, Ulfarsson MO. Multispectral and Hyperspectral image fusion using a 3-D-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*. 2017;**14**:639-643

[33] Masi G, Cozzolino D, Verdoliva L, et al. Pansharpening by convolutional neural networks. *Remote Sensing*. 2016; **8**:594

- [34] Shao Z, Cai J. Remote sensing image fusion with deep convolutional neural network. *The IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2018;**11**:1656-1669
- [35] Song H, Liu Q, Wang G, et al. Spatiotemporal satellite image fusion using deep convolutional neural networks. *The IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2018;**11**:821-829
- [36] Tuia D, Flamary R, Courty N. Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2015;**105**:272-285
- [37] Zhong J, Yang B, Huang G, et al. Remote sensing image fusion with convolutional neural network. *Sensing and Imaging*. 2016;**17**:10
- [38] Zhu XX, Tuia D, Mou L, et al. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*. 2017;**5**:8-36
- [39] Osadciw L, Veeramachaneni K. Fusion, decision-level. In: Li SZ, Jain AK, editors. *Encyclopedia of Biometrics*. Boston, MA: Springer US; 2015. pp. 747-751
- [40] Qin R, Tian J, Reinartz P. Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images. *International Journal of Remote Sensing*. 2016;**37**:3455-3476
- [41] Albanwan H, Qin R, Lu X, et al. 3D iterative spatiotemporal filtering for classification of multitemporal satellite data sets. *Photogrammetric Engineering and Remote Sensing*. 2020;**86**:23-31
- [42] Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. San Diego, CA, USA: IEEE; 2005. pp. 807-814
- [43] Jensen JR. *Remote Sensing of the Environment: An Earth Resource Perspective*. 2nd ed. Pearson Prentice Hall: Upper Saddle River, NJ; 2007
- [44] Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005;**27**:1615-1630
- [45] Mikhail EM, Bethel JS, McGlone JC. *Introduction to Modern Photogrammetry*. New York: Chichester: Wiley; 2001
- [46] Habib AF, Morgan MF, Jeong S, et al. Epipolar geometry of line cameras moving with constant velocity and attitude. *ETRI Journal*. 2005;**27**:172-180
- [47] Facciolo G, De Franchis C, Meinhardt-Llopis E. Automatic 3D reconstruction from multi-date satellite images. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Honolulu, HI: IEEE; 2017. pp. 1542-1551
- [48] Qin R. A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2019;**154**:139-150
- [49] Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*. 1991;**37**:35-46
- [50] Foody GM, Mathur A. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*. 2004;**42**:1335-1343

- [51] Xiong Y, Zhang Z, Chen F. Comparison of artificial neural network and support vector machine methods for urban land use/cover classifications from remote sensing images a case study of Guangzhou, South China. In: 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). 2010. pp. V13-52-V13-56
- [52] Brown LG. A survey of image registration techniques. *ACM Computing Surveys*. 1992;**24**:325-376
- [53] Goshtasby A. 2-D and 3-D Image Registration for Medical, Remote Sensing, and Industrial Applications. Hoboken, NJ: J. Wiley & Sons; 2005
- [54] Yu Y, Liu F. Dense connectivity based two-stream deep feature fusion framework for aerial scene classification. *Remote Sensing*. 2018;**10**: 1158
- [55] Boureau Y-L, Bach F, LeCun Y, et al. Learning mid-level features for recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE; 2010. pp. 2559-2566
- [56] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA: IEEE; 2005. pp. 886-893
- [57] Harris C, Stephens M. A combined corner and edge detector. In: Proceedings of the Alvey Vision Conference 1988. Manchester: Alvey Vision Club; 1988. pp. 23.1-23.6
- [58] Lowe DG. Distinctive image features from scale-invariant Keypoints. *International Journal of Computer Vision*. 2004;**60**:91-110
- [59] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06). New York, NY, USA: IEEE; 2006. pp. 2169-2178
- [60] Ranzato M Aurelio, Boureau Y-Lan, Cun YL. Sparse feature learning for deep belief networks. In: Platt JC, Koller D, Singer Y, et al, editors. *Advances in Neural Information Processing Systems* 20. Curran Associates, Inc.; 2008. pp. 1185-1192
- [61] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*. San Jose, California: ACM Press; 2010. p. 270
- [62] Yuan D, Elvidge CD. Comparison of relative radiometric normalization techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*. 1996;**51**:117-126
- [63] Young NE, Anderson RS, Chignell SM, et al. A survival guide to Landsat preprocessing. *Ecology*. 2017; **98**:920-932
- [64] Elvidge CD, Yuan D, Weerackoon RD, et al. Relative radiometric normalization of Landsat multispectral scanner (MSS) data using a automatic scattergram-controlled regression. *Photogrammetric Engineering and Remote Sensing*. 1995;**61**:1255-1260
- [65] Moran MS, Jackson RD, Slater PN, et al. Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output. *Remote Sensing of Environment*. 1992;**41**:169-184
- [66] VERMOTE E, KAUFMAN YJ. Absolute calibration of AVHRR visible and near-infrared channels using ocean

and cloud views. *International Journal of Remote Sensing*. 1995;**16**:2317-2340

[67] Slater PN, Biggar SF, Holm RG, et al. Reflectance- and radiance-based methods for the in-flight absolute calibration of multispectral sensors. *Remote Sensing of Environment*. 1987;**22**:11-37

[68] Paolini L, Grings F, Sobrino JA, et al. Radiometric correction effects in Landsat multi-date/multi-sensor change detection studies. *International Journal of Remote Sensing*. 2006;**27**:685-704

[69] Hilker T, Wulder MA, Coops NC, et al. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sensing of Environment*. 2009;**113**:1613-1627

[70] Crist EP, Cicone RC. A physically-based transformation of thematic mapper data—the TM Tasseled cap. *IEEE Transactions on Geoscience and Remote Sensing*. 1984;**GE-22**:256-263

[71] Knauer K, Gessner U, Fensholt R, et al. An ESTARFM fusion framework for the generation of large-scale time series in cloud-prone and heterogeneous landscapes. *Remote Sensing*. 2016;**8**:425

[72] Ball GH, Hall DJ. *Isodata, a Novel Method of Data Analysis and Pattern Classification*. Menlo Park, Calif: Stanford Research Institute; 1965

[73] Huang B, Wang J, Song H, et al. Generating high spatiotemporal resolution land surface temperature for urban Heat Island monitoring. *IEEE Geoscience and Remote Sensing Letters*. 2013;**10**:1011-1015

[74] Kwan C, Zhu X, Gao F, et al. Assessment of spatiotemporal fusion algorithms for planet and worldview images. *Sensors*. 2018;**18**:1051

[75] Ma J, Zhang W, Marinoni A, et al. An improved spatial and temporal

reflectance Unmixing model to synthesize time series of Landsat-like images. *Remote Sensing*. 2018;**10**:1388

[76] Kwan C, Budavari B, Gao F, et al. A hybrid color mapping approach to fusing MODIS and Landsat images for forward prediction. *Remote Sensing*. 2018;**10**:520

[77] Chen S, Wang W, Liang H. Evaluating the effectiveness of fusing remote sensing images with significantly different spatial resolutions for thematic map production. *Physics and Chemistry of the Earth, Parts A/B/C*. 2019;**110**:71-80

[78] Chavez PSJ, Sides SC, Anderson JA. Comparison of three different methods to merge multiresolution and multispectral data: LANDSAT TM and SPOT panchromatic. *AAPG Bulletin (American Association of Petroleum Geologists) (USA)*. 1990;**74**(6):265-303. Available from: <https://www.osti.gov/biblio/6165108> [Accessed: 23 April 2020]

[79] Zhang Q, Liu Y, Blum RS, et al. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*. 2018;**40**:57-75

[80] Zhang Q, Ma Z, Wang L. Multimodality image fusion by using both phase and magnitude information. *Pattern Recognition Letters*. 2013;**34**:185-193

[81] Liu Y, Jin J, Wang Q, et al. Region level based multi-focus image fusion using quaternion wavelet and normalized cut. *Signal Processing*. 2014;**97**:9-30

[82] Mallat SG. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1989;**11**:674-693

[83] Li S, Yin H, Fang L. Remote sensing image fusion via sparse representations

over learned dictionaries. IEEE Transactions on Geoscience and Remote Sensing. 2013;51:4779-4789

[84] Wei Q, Bioucas-Dias J, Dobigeon N, et al. Hyperspectral and multispectral image fusion based on a sparse representation. IEEE Transactions on Geoscience and Remote Sensing. 2015; 53:3658-3668

[85] Lagrange A, Le Saux B, Beaupere A, et al. Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Milan, Italy: IEEE; 2015. pp. 4173-4176

[86] Zhang P, Gong M, Su L, et al. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing. 2016;116:24-41

[87] Gong M, Zhan T, Zhang P, et al. Superpixel-based difference representation learning for change detection in multispectral remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 2017; 55:2658-2673

[88] Heideklang R, Shokouhi P. Decision-level fusion of spatially scattered multi-modal data for nondestructive inspection of surface defects. Sensors. 2016;16:105

[89] Du P, Liu S, Xia J, et al. Information fusion techniques for change detection from multi-temporal remote sensing images. Information Fusion. 2013;14: 19-27

[90] Nunez J, Otazu X, Fors O, et al. Multiresolution-based image fusion with additive wavelet decomposition. IEEE Transactions on Geoscience and Remote Sensing. 1999;37:1204-1211

[91] Kuschik G. Large Scale Urban Reconstruction from Remote Sensing Imager. 2013. Available form: <https://www.semanticscholar.org/paper/LARGE-SCALE-URBAN-RECONSTRUCTION-FROM-REMOTE-Kuschik/91cd21f39b27088e6a9ba8443558281074356f16> [Accessed: 27 April 2020]

[92] Qin R, Fang W. A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization. Photogrammetric Engineering and Remote Sensing. 2014; 80:873-883

[93] Albanwan H, Qin R. Enhancement of depth map by fusion using adaptive and semantic-guided spatiotemporal filtering. In: Annals. Photogramm. Remote Sens. Spatial Inf. Sci. 2020. ISPRS Congress (2020/2021). Nice, Fr: ISPRS; 2020

[94] Jing L, Wang T, Zhao M, et al. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox. Sensors. 2017;17:414. DOI: 10.3390/s17020414

[95] Paisitkriangkrai S, Sherrah J, Janney P, et al. Semantic Labeling of aerial and satellite imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2016;9:2868-2881

[96] Audebert N, Le Saux B, Lefèvre S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Lai S-H, Lepetit V, Nishino K, et al, editors. Computer Vision – ACCV 2016. Cham: Springer International Publishing; 2016. pp. 180–196

[97] Tomasi C, Manduchi R. Bilateral filtering for gray and color images. In: Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271). 1998. pp. 839-846