A Reliable Mixed-Norm-Based Multiresolution Change Detector in Heterogeneous Remote Sensing Images

Redha Touati, Max Mignotte, and Mohamed Dahmane

***Abstract*—Analysis of heterogeneous remote sensing image is a challenging and complex problem due to the fact that the local statistics of the data to be processed can be radically different. In this article, we present a novel and reliable unsupervised change detection (CD) method to analyze heterogeneous remotely sensed image pairs. The proposed method is based on an imaging modality-invariant operator that detects at different scale levels the differences in terms of high-frequency pattern of each structural region existing in the two heterogeneous satellite images. First, this new detector is based upon a dual-norm formulation that makes our underlying CD estimation particularly robust in terms of a sensitivity/specificity tradeoff. Second, the detection process, embedded in a multiresolution framework, allows us to estimate a robust similarity or difference map that is then filtered out by a superpixel-based spatially adaptive filter to further increase its reliability against noise. Finally, changes are then identified from this similarity map by a simple binary clustering process that also takes into account the spatial contextual information around each pixel. Experimental results involving different types of heterogeneous remotely sensed image pairs confirm the robustness of the proposed approach.**

***Index Terms*—Change detection (CD), change detection operator, dual-norm estimation, heterogeneous images, heterogeneous remote sensing, multiresolution, spatial–temporal gradient, superpixel.**

I. Introduction

IN REMOTE sensing imagery, heterogeneous images generally refer to a combination of two or several satellite images that can be used to represent an area of interest over the time, and which are acquired by different satellite sensors, either with the same sensor type but with two different optical, SAR, or other systems (multisensor images), or with different sensor types such as SAR/optical images (multisource images), or possibly with the same satellite sensor but with different looks or specification (mutilooking images). Thereby, pixels in heterogeneous images are represented in two distinct feature spaces that do not share the same statistical properties.

Heterogeneous (or multimodal) change detection (CD) [1] is a recent (introduced less than a decade ago) procedure seeking to identify any land cover changes (or land cover uses) that may have occurred between two heterogeneous satellite images acquired on the same geographical area at different times. It is a nontrivial and challenging task, which can be considered as the generalization of the traditional monomodal CD problem as it must take into account multiple origins and characteristics of the acquired data. On the other hand, such a procedure must be adaptive and flexible enough to adapt itself to any existing heterogeneous data types in order to solve the same problems, which are now basically well resolved by the classical monomodal CD techniques [2]–[6], namely, environmental monitoring, deforestation, urban planning, and land or natural disaster/damage monitoring and management, to name a few.

Heterogeneous (or multimodal) CD has recently generated a growing interest in the remote sensing community, and the huge amount of heterogeneous data we can now get from existing Earth observing satellites or extracted from various archives can partly explain this [1], [7]–[9]. In fact, the practical and technical advantages of such a multimodal analysis procedure are obvious both technically and practically [7], [10]. First, let us emphasize that a heterogeneous CD approach may be useful and sometimes indispensable in some emergency cases. SAR sensors can operate regardless of weather conditions, even at night, i.e., with less restrictive conditions compared to optical imaging [7]–[9]. We can give the representative case of an optical image of a given area, which is provided by available remote sensing image archive data, and only a new SAR image can be acquired for technical reasons, such as lack of time, availability, or atmospheric conditions in an emergency situation for the same area [7]–[9]. A similar example can be given in the case of specific situations, in which the area to monitor is located in a tropical or boreal forest and for which SAR imaging offer the great advantage, over its optical counterparts, of not being affected by heavy clouds, fog, haze, and also rain, or else in

Manuscript received October 8, 2018; revised January 22, 2019; accepted August 2, 2019. Date of publication August 27, 2019; date of current version September 29, 2019. This work was supported in part by the Computer Research Institute of Montréal and in part by the Ministry of Economic Science and Innovation, Government of Québec. *(Corresponding author: Redha Touati.)*

R. Touati is with the Vision Laboratory, Département d'Informatique et de Recherche Opérationnelle, Faculté des Arts et des Sciences, Université de Montréal, Montréal, QC H3C 3J7, Canada, and also with the R&D Vision Département, Centre de Recherche Informatique de Montréal, Montréal, QC H3N 1M3, Canada (e-mail: touatire@iro.umontreal.ca).

M. Mignotte is with the Vision Laboratory, Département d'Informatique et de Recherche Opérationnelle, Faculté des Arts et des Sciences, Université de Montréal, Montréal, QC H3C 3J7, Canada (e-mail: mignotte@iro.umontreal.ca).

M. Dahmane is with the R&D Vision Département, Centre de Recherche Informatique de Montréal, Montréal, QC H3N 1M3, Canada, and also with the École de Technologie Supérieure, Université du Québec, Montréal, QC H3C 1K3, Canada (e-mail: mohamed.dahmane@crim.ca).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2019.2934602

snow-covered regions of high altitudes, for which SAR is also able to penetrate a thin snow layer, or finally to monitor the progress of a fire, since SAR imaging, operating at microwave frequencies, can see (i.e., penetrate) through smoke and dust [7]–[9], [11], [12]. Let us also stress that, since multimodal CD must be adaptable to heterogeneous data with different statistics, this procedure may turn out to be more robust to natural variations in environmental variables such as soil moisture or phenological states (e.g., flowering, maturing, drying, senescence, harvesting, etc.) or shading effects, which should not be detected as land cover changes and which is sometimes taken into account and corrected in the preprocessing step of a classical monomodal CD approach. Finally, let us add that two different imaging modalities may be complementary (as it is especially the case of SAR and optical or multispectral sensors), and this complementarity could be exploited (not only in geoscience imaging [13]) for further improving the CD and analysis of complex land cover types or for sensors operating in extreme conditions.

Up to now, relatively few research works have been developed in heterogeneous CD [7], [10], [11], [14], but, generally, we can divide them into three main categories: parametric, nonparametric, or invariant similarity measure or operator-based models.

In parametric techniques, a mixture or a set of parametric multivariate distributions is generally used to directly or indirectly model the joint statistics or the dependencies between the two imaging modalities. In this category, we can mention the copula-based approach proposed in [10], in which the dependence between the two satellite images, in unchanged areas, is modeled by a quantile regression applied according to the copula theory (a powerful tool for tackling the problem of how to describe a joint distribution) and Kullback–Leibler-based comparisons on local statistical measures to generate a similarity map, which is then finally analyzed by thresholding to detect between change and no change areas. An interesting two-step multivariate statistical approach has also been proposed in [7]–[9], whose first step aims at estimating a physical model, based on a mixture of multidimensional distributions (taking into the noise model the relationships between the sensor responses to the objects and their physical properties), with the expectation–maximization (EM) algorithm [15]. A statistical test based on this model then allows us to estimate the changes. In the same spirit, the authors of [11] also propose to first estimate a multidimensional distribution mixture estimation based on a new family of multivariate distributions with different shape parameters and especially well suited for detecting changes in SAR images acquired by different sensors having different numbers of looks.

The problem with these parametric techniques is that they have been especially designed (*via* specific distribution types) for a type of multimodal sensors (optical/SAR in [7], [8], and [10] or SAR with different numbers of looks [11]), and consequently, they are not easily generalizable for another pair of different sensors. Besides, these methods are in fact semisupervised, since they generally require (as training set) that two training images (sometimes manually selected and carefully chosen) associated with an unchanged area are available [7], [8], [10]. Let us finally add that these methods also require a maximum likelihood parameter estimation step of the distribution laws considered, which can be complex and computationally expensive.

Among nonparametric methods, an energy minimization model has been specifically designed in [12] for satisfying an overdetermined set of constraints, expressed for each pair of pixels existing in the before-and-after satellite images acquired through different modalities. An estimation of this overconstrained problem, formulated as a pairwise energy-based model, is then carried out in the least squares sense, by a fast linear-complexity algorithm based on a multidimensional scaling (MDS) mapping technique leading to a similarity feature map, which is then binarized into two classes to distinguish changes of interest of the land cover. In [16], a method is presented, in which the original pair of temporal images is transformed into a new feature space or representation, especially designed to be invariant to imaging modality and aiming at highlighting the changes. In the same spirit, Volpi *et al.* [17] find joint projections of the paired input images by maximizing the correlation between the projected data with a canonical correlation analysis. Another representation that turns out to be invariant to imaging modality can be given by a segmentation of the before-and-after images. In this optic, Liu *et al.* [18] propose a general multidimensional evidential reasoning (MDER) approach using the segmentation results of the pre- and postevent satellite images with an extension of the fuzzy c-means clustering under the belief function framework and whose result is directly used as a basic belief assignment in their MDER approach. A similar strategy is also proposed in [19]. In the same vein, a presegmentation strategy based on the normalized difference spectral index is described in [20]. Let us note that machine-learning-based methods are also nonparametric (in the sense that they do not assume a specific parametric distribution for the data), and deep learning methods through conditional adversarial networks [21], convolutional coupling networks [14], or method based on deep feature representation [22], binary support vector classifier [23], multiclassifier systems [24], or based on a simple K-nearest neighbors technique [25] have also been recently proposed and turn out to be valuable for the multimodal CD problem.

In fact, these nonparametric methods have also the defect of its main quality. Their ability to process a wide variety of imaging modalities (with different noise types and levels) explains why they are possibly less accurate than a specific heterogeneous CD model dealing with a specific type of multimodality, which is modeled by a particular joint (or mixture of) distribution(s) whose shape has a clear physical and statistical justification. For the machine-learning-based heterogeneous CD models, the efficiency of these algorithms heavily depends on the availability of an adequate massive amounts of representative training data.

Finally, in the third family of methods, Alberga [26] proposes to use a technique closed to the coregistration and based on the use of a combination of different invariant similarity measures (such as correlation ratio, mutual information, etc.) in order to estimate the correspondence between the same points in the two images and finally to detect eventual changes existing between two heterogeneous data acquisitions. Also, in [27], a CD method is presented to quantify the damages caused by an earthquake

to each individual building, using preevent optical image and postevent SAR images. To this end, the parameters of each building, estimated from the optical scene and combined with the acquisition parameters of the actual postevent SAR scene, are used to predict (*via* simulation) the expected SAR signature of the building, which is then subsequently compared, with a similarity measure, to the actual SAR scene in order to quantify the damages caused to each building. The main interest of this family of methods relies on the fact that they do not have the disadvantages of the first two above-mentioned categories of models (parametric and nonparametric) and are also more flexible in the sense that they are not closely related to a specific mathematical framework (Bayesian or multivariate analysis in the first and regression analysis for the second category).

In this article, we propose a new imaging modality invariant change detector, which belongs to the third family of above-cited methods. Compared to our preliminary model [28], this operator is defined at three resolution scales and made scale invariant. In addition, this operator is estimated according to two different and complementary norms, for complementarity reasons and better detection results in terms of self-balancing the precision and recall of the considered changed/unchanged detection problem. Finally, the information provided by these dual operators at different scales is combined, thanks to the MDS mapping method, to generate a similarity feature map, which turns out to be especially well suited to estimate the differences existing in the land cover change between heterogeneous images coming from different imaging modalities or sensors involved in remote sensing imagery. Once a similarity feature map is estimated by this change detector, changed and unchanged areas are then finally identified by a final unsupervised binary clustering approach based on the K-means procedure.

The major advantage of the proposed model lies in its flexibility to process a wide variety of heterogeneous images without requiring the main drawbacks of parametric models that require an explicit knowledge of the data distribution (and also a complex parameter estimation step of these distribution laws) or again the drawbacks of nonparametric models that require a large and representative training set (and heavy supervised training procedure).

The validation of the proposed approach is done by a series of tests conducted on different real heterogeneous datasets chosen to reflect the different CD problems in the multimodal case, namely, multisensor image pairs with: 1) heterogeneous optical images and multisource image pairs; 2) SAR+optical or optical+SAR images; and, finally, 3) multilooking SAR images.

The remainder of this article is organized as follows. Section II describes the proposed multiscale change detector, which allows us to estimate the similarity-feature map, from which changed and unchanged areas are then identified. Section III presents a set of experimental results and comparisons with existing multimodal CD algorithms. Finally, Section IV concludes this article.

II. Proposed CD Model

The proposed model takes as input two bitemporal heterogeneous remote sensing images (in our case either heterogeneous optical or multisource SAR/optical or multilooking SAR images). The proposed CD model is based on the following four-step procedure.

1) We first estimate a set of multiscale features aiming at detecting the structural difference in terms of high-frequency components of each local region (two-dimensional signal) existing in the before-and-after satellite images. This detector is based on a multiresolution framework that makes it somewhat scale invariant and also exploits a dual-norm relationship that makes it robust to the eventual context of unbalanced data (which is typically our case since the majority of pixels belongs to the *unchanged* class), and consequently, our estimation model could estimate a degenerate overfit solution to this problem by classifying all pixels to be *unchanged* (see Sections II-A and II-B).
2) In order to both reduce the noise and remove redundant information, provided by the previous estimation step, the multiscale feature vector is reduced to one dimension, to get of a similarity (change/no-change) map, by using a fast (linear-complexity) version of the MDS mapping technique (see Section II-C).
3) To further reduce the noise of this similarity map, we then apply a spatially adaptive filters based on the superpixel representation of the before-and-after satellite images (see Section II-D).
4) Finally, to increase the class (change/no-change) separability of each pixel of this similarity map, we transform the local region, in the neighborhood of each pixel, into a point in a discriminant textural feature space, where an unsupervised binary ($K = 2$) clustering algorithm (K-means) is applied (see Section II-E). More precisely, the different steps of our approach are the following.

A. Imaging Modality Invariant Change Feature

Let us consider two (previously coregistered) bitemporal remote sensing (N pixel size) images, y^{t_1} and y^{t_2}, acquired from different sensors or sources at two times (before and after a given event) in the same geographical area.

In the classical monomodal (or homogeneous) CD case, the two coregistered temporal images at two different times are usually first compared pixel by pixel in order to generate a *difference image* by differencing (with a simple subtraction or a temporal gradient operator) or (log-)rationing (i.e., with a log temporal gradient) [2]–[5] [6]. This latter *difference image* is such that the pixels associated with land cover changes present gray-level values significantly larger from those of pixels associated with unchanged areas. A binary segmentation is then finally achieved on this temporal gradient image to distinguish between the changed and no-changed areas. In the heterogeneous or multimodal case, this temporal gradient is not effective [28] particularly when the input images are acquired by different sensor types. Indeed, the gray or color value of each pixel is not useful information, since the gray levels of the same region, in the before and after a given event, may be radically different according to the characteristics of the two input (possibly highly) heterogeneous imaging modalities. Conversely, two distinct regions, at two different times, may be locally coded with

the same (gray or color) value, since two different textures may have the same mean or similar local intensity/color value. Consequently, the classical temporal (or log temporal) gradient operator is thus irrelevant in the heterogeneous case for estimating an accurate *difference image*, which will be subsequently used for identifying land cover change.

Nevertheless, for the same region, represented by two different imaging modalities, there is a feature, which remains relatively invariant between different types of imaging and thus can be herein efficiently exploited and captured by an operator. This feature is the magnitude and orientation distribution of the spatial edges and/or contours existing in the considered region. Indeed, each specific homogeneous region generally exhibits a unique geometric high-frequency pattern. For example, an urban region exhibits a specific directional edge or gradient magnitude distribution (due to the presence of rectangular regions defined by the roads/streets, building roofs, parking lots, electric field lines, residential houses, etc.) which is, more or less, well preserved in the two imaging modalities in the high spatial frequencies of the texture pattern.[1] It is also the case of an agricultural region, where the intrinsic regular location of crops produces edges and contours, which are also fairly well conserved in the two kinds of imagery. This remains true for the other homogeneous regions in a satellite image, even for the water region where the absence or the presence of waves (or wavelets at a finer spatial scale) can be detected and localized (and analyzed, as proposed in [29], for SAR and radar images) in the two different heterogeneous modalities by a high-frequency filter or a simple edge detection algorithm for texture. Let us note that physical features such as normalized difference vegetation index (NDVI) [30] in multispectral imagery or the polarization ratio of SAR data [31] in SAR imagery can also describe the physical properties (size, shape, orientation, etc.) of agricultural areas (in addition to estimating the dielectric properties of the plants for the polarization ratio and the photosynthetic capacity and hence energy absorption of plant canopies for the NDVI). These features have already been used in (monomodal) remote sensing and have been proved to be reliable for segmentation and classification tasks and more precisely for retrieving live green plant canopies or for estimating the different agricultural crop growth stages and some vegetation phenology metrics. Nevertheless, these physical features cannot be straightforwardly used and exploited in a multimodal CD system except in the specific multispectral/optical case introduced in [25], in which an NDVI image, combined with an optical (SPOT) image, is projected in a common feature space for the convenience of CD.

Consequently, since the edge at different spatial scales, or more precisely the specific high-frequency pattern of each textural region, is fairly well preserved, in spite of the difference in the imaging modality between the two heterogeneous temporal images, we propose to base the estimation of our *difference image* z^D on a temporal gradient applied on a local spatial gradient. In our case, this spatial–temporal gradient is approximated using a first-order temporal and spatial finite-difference approximation (in the L_1 norm). More precisely, the similarity map z^{D_1} is computed by estimating at each pixel site s by

$$z_s^{D_1} = \sum_{<s,s'>\in W_n} \left| |\mathbf{y}_s^{t_1} - \mathbf{y}_{s'}^{t_1}|_1 - |\mathbf{y}_s^{t_2} - \mathbf{y}_{s'}^{t_2}|_1 \right| \qquad (1)$$

where the summation is done over all pairs of pixels at location $\langle s, s' \rangle$ contained in an $N_n \times N_n$ squared window W_n, including the central pixel located at site s. This summation allows us to render this temporal–spatial gradient operator invariant to rotation and also less sensitive to noise (due to the averaging process). Hence, we compute a spatial gradient for a (possible) texture region, where the difference $\mathbf{y}_s^{t_1} - \mathbf{y}_{s'}^{t_1}$ is achieved by considering $\mathbf{y}_s$ and $\mathbf{y}_{s'}$ as being two vectors (respectively, at locations $\langle s, s' \rangle \in W_n, s \neq s'$) obtained by gathering together all the gray (or color) values contained in an $N_{s'} \times N_{s'}$ squared window $W_{s'}$ centered on pixel s (for $\mathbf{y}_s$) and centered on pixel s' (for $\mathbf{y}_{s'}$). (Let us note that, instead of gathering the pixel values in the vector $\mathbf{y}_s$, we could also compute local statistics estimated from the values contained around s.) Finally, this temporal–spatial local finite difference between these two (feature) vectors are computed in the L_1-norm sense ($|.|_1$).

A simple way to improve our CD result accuracy consists of considering and estimating the dual and complementary version of the previously expressed [in (1)] similarity map by considering the same local spatiotemporal gradient operator but expressed in terms of the infinity norm (which is the dual norm of the L_1 norm [32], [33]). In this regard, a second similarity map z^{D_2} is estimated, at every pixel s of the image, by the following operator:

$$z_s^{D_2} = \sum_{<s,s'>\in W_n} \max_{1\le i\le N_{s'}\times N_{s'}} \left| |y_i^{t_1}(s) - y_i^{t_1}(s')| - |y_i^{t_2}(s) - y_i^{t_2}(s')| \right| \qquad (2)$$

where $y_i^{t_1}(s)$ is the ith component of the pixel vector $\mathbf{y}_s^{t_1}$ or the ith pixel value taken within the $N_{s'} \times N_{s'}$ squared window $W_{s'}$ (i.e., considering that $\mathbf{y}_s^{t_1} = (y_1^{t_1}(s), \ldots, y_i^{t_1}(s), \ldots, y_{N_{s'}\times N_{s'}}^{t_1}(s))$. In our application, we take $N_n = 7$ and $N_{s'} = 3$ for $z_s^{D_1}$ and $z_s^{D_2}$.

Let us mention that the strategy of combining or mixing different norms, for complementarity reasons and better results, has already been investigated and observed recently in machine learning theory for improving feature selection techniques or for finding a support-vector-machine-based classification rule with a minimal generalization error [34], as well as in image processing, where the quality of the estimation has been found to be improved in the framework of optimization-based regularization in image restoration [35], denoising [36], image deconvolution [37], or fluorescence diffuse optical tomographic reconstruction [38], etc., to name a few. More generally and in summary, it is established, in these works, that estimations based on the $L_{p=1}$ norm generally encourage sparsity contrary to $L_{p>2}$ (and especially L_∞ norm) that favors diversity [39]. This is what we have observed in our multimodal CD or two-class segmentation problem; the spatiotemporal gradient operator based on the L_1 norm favors sparse segmentation result contrary to the one based on the L_∞ norm, which rather encourages diversity.

[1] In fact, more precisely, the local texture pattern created by a given imaging modality is (a mixture of) characteristic(s) of both the region that is being imaged and the imaging system (at medium- or high-frequency levels). This explains why, thanks to its natural bandpass capabilities, the human visual system can recognize, even in a complex SAR image with strong correlated speckle, the specific high-frequency spatial textural pattern created by an urban area.

We can also understand this complementarity in the context of estimation from noisy image data. L_{inf} norm is more sensitive to noise than L_1 and thus less efficient when the image is noisy. Conversely, L_{inf} norm is more discriminant than L_1 if there is not much noise in the image. For different levels of noise (thus regardless of the imaging modality), L_{inf} norm produces a complementary version of L_1, and taking into account of these two norms thus gives a compromise CD estimation, whose distribution (given by the confusion matrix) is well balanced with no bias in favor of one class.

B. Scale-Invariant Change Detector

An appealing hierarchical framework for our CD problem is to consider a multiresolution representation of the input bitemporal satellite images. This multiresolution representation (which can be simply achieved by Gaussian low-pass filtering each previous scale of the input image and decimation by a factor of 2 in the horizontal and vertical directions) has the intrinsic capability to represent and reorganize image information into a set of details (i.e., high-frequency patterns) appearing at different spatial resolution levels. Conceptually, this strategy will allow us to detect and integrate relevant information at different frequencies (which are only represented at a specific resolution scale or pyramid level), and it also makes our change detector robust against noise and somewhat scale invariant.

To this end, we construct two three-level pyramidal representations, resulting from the application (at each resolution level) of, respectively, the first ($z_s^{D_1}$) and second ($z_s^{D_2}$) CD operators [see (1) and (2)] on the two temporal heterogeneous satellite images. For each pixel of the coordinate s=(i, j)=(ROW, COLUMN), a multiscale feature vector v_s is then based on the concatenation of $(z_s^{D_1}, z_s^{D_2})$ obtained at first or finer resolution level, with the two estimations obtained at the second resolution scale, i.e., $(z_{s^{[2]}}^{D_1}, z_{s^{[2]}}^{D_2})$ at pixel coordinate $s^{[2]}$=$(\lceil i/2 \rceil, \lceil j/2 \rceil)$, and, finally, those obtained at the third resolution scale, i.e., $(z_{s^{[3]}}^{D_1}, z_{s^{[3]}}^{D_2})$, at pixel coordinate $s^{[3]}$=$(\lceil i/2^2 \rceil, \lceil j/2^2 \rceil)$ (with $\lceil i \rceil$ being the ceiling function and with $N_n = 7$ and $N_{s'} = 3$ for each operator applied and each scale).

C. Similarity Feature Map Estimation

Finally, in order to further reduce both the noise of the estimation and the redundant information provided by our two operators at different resolution scales, while reducing the dimensionality of the data to be analyzed (and thus also the complexity of the subsequent clustering process described in Section II-E), we reduce the dimensionality of each N_f-size ($N_f = 3(\text{levels}) \times 2(\text{operators})$) multiscale feature vector $(z_s^{D_1}, z_s^{D_2}, z_{s^{[2]}}^{D_1}, z_{s^{[2]}}^{D_2}, z_{s^{[3]}}^{D_1}, z_{s^{[3]}}^{D_2})$, to one dimension with the linear-complexity version of the MDS mapping method, called the FastMap technique[2] [40]. This allows us to obtain a robust similarity feature map y^D with two classes of gray-level values corresponding to change and no-change areas.

[2]The first step of the FastMap algorithm is to select two objects (or feature vector), the most dissimilar to form the projection line. These two objects are selected by using a deterministic procedure called choose distant objects [40]. The second step is to project any other object onto this orthogonal axis (called a pivot line) by employing the cosine rule (see Algorithm 1). The FastMap C++ codes are freely available on the web.

Algorithm 1: FastMap.

Input:
k: Dimensionality of target space
N_p: Number of objects (vectors) in database

Output:
$X_{N_p \times k}$: Number of objects in target space

Initialization:
$d \leftarrow 0$

FASTMAP ALGORITHM $(k, D(), \mathcal{O})$

- **if** $k \leq 0$ **then** return X
- $d \leftarrow d + 1$
- Choose *pivot* objects O_a and O_b such that the distance $D(O_a, O_b)$ is maximized

foreach *object* i *from* $\mathcal{O}$ **do**
- Project O_i on the line (O_a, O_b)
  Compute : $X[i, d] = x_i$
  $x_i = \frac{D^2(O_a,O_i)+D^2(O_a,O_b)-D^2(O_b,O_i)}{2\,D^2(O_a,O_b)}$

end

foreach *object* i *from* $\mathcal{O}$ **do**
- Project O_i on an hyper-plane perpendicular to the line (O_a, O_b)
  $(D')^2(O'_i, O'_j) = D^2(O_i.O_j) - (x_i - x_j)^2$

end

call FASTMAP$(k - 1, D'(), \mathcal{O})$

D. Superpixel-Based Filtering Step

Once the feature similarity map y^D is estimated, thanks to our above-presented scale and rotation-invariant temporal–spatial gradient operator for texture, we decide to filter y^D with an original superpixel-based filtering strategy in order to make y^D less noisy and thus to make its subsequent classification into change and no-change areas (see Section II-E) more robust.

A superpixel is a perceptually meaningful collection of pixels, obtained from some low-level grouping process. Fundamentally, it is the result of an oversegmentation, in which the pixels inside each superpixel form a consistent, perceptually meaningful, unit or atomic region, e.g., in terms of color, texture, intensity, and so on. In addition to estimating a set of homogeneous regions (of nearly similar size) allowing to preserve the important structures in the image, this low-level process is also representationally and computationally efficient. By replacing the rigid structure of the pixel grid, it reduces the complexity of images from hundreds of thousands of pixels to only a few hundred superpixels. Recently, an interesting superpixel algorithm called simple linear iterative clustering (SLIC) [41] has been proposed, which, compared to the state-of-the-art superpixel methods, turns out to be superior for both efficiency and boundary preservation. SLIC is a two-step procedure, which first estimates

Algorithm 2: SLIC Segmentation.

Input:
Image with N pixels
K: Desired number of Superpixels

Output:
Image segmented

Initialization:

- $S = \sqrt{N/K}$
- Choose K Cluster (superpixel) centers $C_k = [l_k, a_k, b_k, x_k, y_k]^T$ in LAB space color (or gray level L; $C_k = [l_k, 0, 0, x_k, y_k]^T$, where the l_k component is calculated directly from the grayscale value) with position (x, k) by sampling pixels at regular grid steps S
- Perturb cluster centers in an $n \times n$ neighborhood, to the lowest gradient position

while $E \leqslant threshold$ **do**
  foreach *each cluster center* C_k **do**
    - Assign the best matching pixels from a $2S \times 2S$ square neighborhood around the cluster center
  end
  - Compute new cluster centers and residual error E (L_1 distance between previous centers and recomputed centers)
end
- Enforce connectivity

superpixels by grouping pixels with a local k-means clustering method and second exploits a connected component algorithm to remove the generated small isolated regions by merging them into the nearest large superpixels.

In our application, SLIC is applied on y^{t_1} and y^{t_2} in order to detect the different consistent structural regions (*land uses*) existing in these images. The intersection between these two SLIC segmented images[3] allows us to define a third oversegmented map y^S (with thus smaller superpixels), in which the set of pixels inside each new superpixel has the appealing property to exhibit homogeneous structural regions (in terms of land uses) in the *before* and *after* images. At this stage, a possible strategy is to exploit the collection of superpixel belonging to y^S (and $\{y^{t_1}, y^{t_2}\}$ or y^D) to individually classify each superpixel into *changed* or *no-changed* class. This approach is algorithmically complex and, in practice, does not perform as well as the second strategy used in this article that consists of averaging each pixel value of y^D, inside each superpixel of y^S, between them. Conceptually, this latter strategy can be interpreted as a segmentation-based spatially adaptive filter, which averages the values given by our CD operator within each individual homogeneous *changed* or *no-changed* small region previously estimated (see Fig. 1 and Algorithm 3, which simply average out each y^D values of each segment).

[3]if x^{t1} denotes the segmentation or the subdivision of the image y^{t1} into a set of superpixels or regions: $x^{t1} = \{R_1^{t1}, R_2^{t1}, \ldots, R_{N_1}^{t1}\}$ and x^{t2} is the subdivision of y^{t2}, *i.e.*, $x^{t2} = \{R_1^{t2}, R_2^{t2}, \ldots, R_{N_2}^{t2}\}$. Every pixel of the image pair (y^{t1}, y^{t2}) is thus associated with a unique region in the set x^{t1} and a unique region in the set x^{t2}. Each unique pair of regions defines a new individual region in the segmentation map y^S, which is defined as the intersection of x^{t1} and x^{t2}. Conceptually, each generated superpixel in y^S corresponds to a group of connected pixels belonging to the same region in x^{t1} and the same region in x^{t2}.

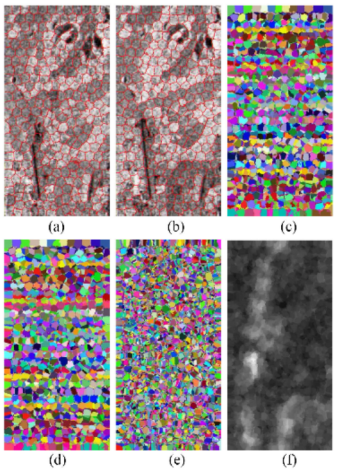

Fig. 1. Superpixel-based filtering step on SAR three-look/SAR five-look dataset (sixth dataset; cf., Section III-A). (a) and (b) Superpixel contour superimposed on y^{t_1} (before) and y^{t_2} (after) satellite image. (c) and (d) Segmentation into superpixel regions (on images y^{t_1} and y^{t_2}). (e) Segmentation intersection y^S [between segmentation maps (c) and (d)]. (f) Filtered similarity feature map $\overline{z}^D$ [by spatial averaging all the values of the similarity feature map over each superpixel estimated in (e)].

E. Two-Class Clustering

Finally, in order to automatically separate the change and no-change areas from the previously filtered feature similarity map $\overline{y}^D$, we use the following unsupervised clustering approach, whose aim is to increase the separability of the two classes or clusters; we apply a small overlapping sliding window over the image of size 7×7, in which we compute three features, namely, the empirical mean and variance of luminance and the maximum gray level for each location of the window. Each window location thus provides a three-component "sample" $\mathbf{x}_m$. The collected samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are then clustered into two classes $\{e_0, e_1\}$ using the k-means clustering procedure [42], [43]. In fact, this strategy allows us to increase the separability of the two clusters by taking into account the spatial contextual information (or the neighborhood) around each pixel in the binary clustering process (see Fig. 2 and Algorithm 4).

Algorithm 3: Superpixel-based Filter.

Input:
y^D: Similarity map (to be filtered)
$\{y^{t_1}, y^{t_2}\}$: Image before and after
K: Desired number of superpixels

Output:
$\overline{y}^D$: Filtered similarity map

Initialization:
- $x^{t1} \leftarrow$ SLIC_SEGMENTATION $(y^{t_1}; K)$
- $x^{t2} \leftarrow$ SLIC_SEGMENTATION $(y^{t_2}; K)$
- $y^S \leftarrow$ INTERSECTION(x^{t1}, x^{t2})

foreach *superpixel* $b_i \in y^S$ **do**
  $val \leftarrow 0$
  $nb \leftarrow 0$
  foreach *pixel* p_s *(at location* s*)* $\in b_i$ **do**
    $val \leftarrow val + y_s^D$
    $nb \leftarrow nb + 1$
  end
  foreach *pixel* p_s *(at location* s*)* $\in b_i$ **do**
    $\overline{y}_s^D \leftarrow (val/nb)$
  end
end

Algorithm 4: Two-Class Clustering.

Input:
$\overline{z}^D$: Filtered similarity map (to be segmented)

Output:
x^{CD}: binary CD map (with N pixels)

foreach *pixel* p_i *(at location* i*)* $\in \overline{z}^D$ **do**
  • Compute the m_i =mean, v_i =variance and mx_i = maximum gray level contained within the 7×7 window centered on p_i.
  $\mathbf{x_i} \leftarrow (m_i, v_i, mx_i)$
end
• $x^{CD} \leftarrow$ K(=2)-MEANS ALGO $(\mathbf{x_1}, ..., \mathbf{x_N})$

III. Experimental Results

To validate our approach, in this section, we present a series of tests conducted on different real heterogeneous datasets, chosen to reflect the three possible CD conditions in the multimodal case, namely, two heterogeneous optical images and heterogeneous SAR images, and one optical and one SAR images. This allows us to compare the performance of the proposed method with different state-of-the-art multimodal CD algorithms recently proposed in this field [7], [8], [10] [9], [11], [44] in different multimodal CD conditions. In this benchmark, all the ground-truth images (or CD mask) were provided by an expert photo interpreter. We also compare the obtained results with the other change detector traditionally proposed in mono- or multimodal cases and provided by the ORFEO Toolbox [45] [46]. In our implementation, we have used the FastMap and SLIC C++ codes kindly provided by their authors and freely available on the web.

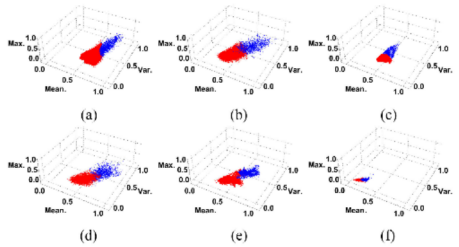

Fig. 2. (a)–(f) 3-D feature space for the local textural features (mean, variance, and maximum gray level) of the filtered feature similarity map y^D related to different heterogeneous datasets. Red and blue colors represent, respectively, the unchanged and changed clusters or classes found by the K-means algorithm.

A. Heterogeneous Dataset Description

1) The first heterogeneous dataset is a pair of SAR/optical satellite images (Toulouse, France), with a size of 4404×2604 pixels, before and after a construction. The SAR image was taken by the TerraSAR-X satellite (February 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of the Centre National d'Etudes Spatiales) satellite (July 2013). The TSX image was coregistered and resampled in [47] with a pixel resolution of 2 m to match the optical image.
2) The second one is a pair of optical/SAR satellite images (Gloucestershire region, in southwest England, near Gloucester), with a size of 2325×4135 pixels, before and after a flooding taking place in urban and rural areas. The optical image comes from the Quick Bird 02 (QB02) VHR satellite (July 15, 2006) and the SAR image was acquired by the TerraSAR-X satellite (July 2007). The TSX image presents a resolution of 7.3 m, and the QB02 image (with a resolution of 0.65 m and 0% cloud cover) was coregistered and resampled in [47] to match this resolution.
3) The third dataset shows two heterogeneous optical images acquired in Toulouse (Fr) area by different sensor specifications (size of 2000×2000 pixels with a resolution of 0.5 m). The *before* image is acquired by the Pleiades sensor in May 2012 before the beginning of the construction work, and the *after* image is acquired by WorldView2 satellite from three (Red, Green, and Blue) spectral bands (July 11, 2013) after the construction of a building. The WorldView2 VHR-image was coregistered in [47] to match the Pleiades image.
4) The fourth dataset [11] is a pair of SAR/SAR satellite images (Gloucester, U.K.) before and during a flood event caused by intense and prolonged rainfall,

TABLE I
DESCRIPTION OF THE EIGHT HETEROGENEOUS DATASETS

Dataset	Date	Location	Size (pixels)	Common spatial resolution	Sensor
1	Feb. 2009 - July 2013	Toulouse, Fr	4404×2604	2 m.	TerraSAR-X / Pleiades
2	July 2006 - July 2007	Gloucester, UK	2325×4135	0.65 m.	TerraSAR-X / QuickBird 02
3	May 2012 - July 2013	Toulouse, Fr	2000×2000	0.52 m.	Pleiades / WorldView 2
4	Sept. 2000 - Oct. 2000	Gloucester, UK	762×292	40 m.	RADARSAT
5	Sept. 1999 - Nov. 2000	Gloucester, UK	1318×2359	10 m.	VHR Spot / ERS
6	Jan. 2002 - Jan. 2002	Central Africa, CF	400×800	10 m.	RADARSAT
7	Sept. 1995 - Jul. 1996	Sardinia, IT	412×300	30 m.	Landsat-5 (NIR band) / Landsat-5
8	Jun. 2008 - Sept. 2012	Dongying, CH	921×593	8 m.	RADARSAT-2 / QuickBird and Landsat-7

overwhelming the drainage capacity, on urban and agricultural/rural areas, with a size of 762×292 pixels, acquired by RADARSAT satellite with different numbers of looks. The number of looks for the before SAR image is 1 (September 2000), and the number of looks for the after image is 5 (October 2000). These two SAR images have a resolution of about 40 m.

5) The fifth dataset [44], [46] consists of one multispectral image and one SAR image showing the area of Gloucester (U.K.), with a size of 1318×2359 pixels. The multispectral image is taken by the Spot VHR satellite in September 1999 before a flooding event. The SAR image is captured by the European Remote Sensing (ERS) satellite (around November 2000) during the flooding event. The resolution of these two images is about 10 m [46].
6) The sixth dataset [11] shows a pair of heterogeneous satellite images (with a size of 400×800 pixels and a resolution of 10 m) acquired over the Democratic Republic of the Congo (country located in central Africa) before and after the eruption of the Nyiragongo volcano (January 2002). It consists of two SAR images captured by the RADARSAT satellite with different numbers of looks. The number of looks for the SAR image before and after change is 3 and 5, respectively.
7) The seventh dataset is composed of two heterogeneous optical images [22]. It shows the changes of the Mediterranean in Sardinia area (Italy). This dataset is acquired by different sensor specifications and consists of one TM image (optical) and one optical image. The before image is the fifth band of a TM image (near-infrared band) acquired by the Landsat-5 (September 1995) with a spatial resolution of 30 m [22]. The second optical image comes from Google Earth (July 1996) and is an RGB image with a spatial resolution of 4 m [22]. After coregistration, these two images are resampled at the same pixel resolution of 412×300 pixels [22].
8) The eighth dataset consists of one SAR image and one RGB optical image. It shows a piece of the Dongying City in China, before and after a new building construction. The SAR image is acquired by RADARSAT-2 (June 2008) with a spatial resolution of 8 m. The optical image comes from Google Earth image (September 2012), and it is a combination of aerial photography imaging with a satellite imaging (produced, respectively, by QuickBird and Landsat-7) with a spatial resolution of 4 m. After coregistration, these two images are resampled at the same pixel resolution of 921×593 pixels.

Table I summarizes a brief description of the eight heterogeneous remote sensing image datasets used in our research, which cover the possible cases that may arise in the heterogeneous CD problem.

B. Results and Evaluation

In all the experimental results, we have considered the simple gray level of the image (and not a local statistics vector around a neighborhood of s) [see (1) and (2)]. In the case of an optical image, this also requires the conversion of the possible color image to a grayscale image. Each operator results z^D (at each resolution level) are rescaled for all sites s of the image between 0 and 255. We have considered $N_p = 3$ levels of the multiresolution pyramidal structure and $N_n = 7$ and $N_{s'} = 3$ for each operator applied at each scale of this pyramid (see Section II-B). Finally, for the superpixel-based filtering step (see Section II-D), the parameters of the SLIC algorithm are $N_s = 300$.

In order to discuss and compare obtained results, a quantitative study is realized by computing the classification rate accuracy that measures the percentage of the correct changed and unchanged pixels:

$$\text{PCC} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \tag{3}$$

where TP is the true positive value that corresponds to the number of pixels that are detected as the changed area in both the ground-truth image and the obtained results. TN is the true negative value that corresponds to the pixel number belonging to the intersection of the unchanged area in both the reference image and the obtained results. FN represents the false negative value done by the number of the missed changed pixels in the obtained results, and FP represents the false positive value corresponding to the unchanged pixels wrongly classified as changed.

A comparison with different state-of-the-art approaches [7], [8], [10] [9], [11], [44] is summarized in Table II. We have also summarized in Table III the confusion matrix obtained by the proposed change detector. From Table II, we can see that the rate accuracy of our method outperforms the most other state-of-the-art approaches and shows the strength and the flexibility of our method to process both the three different heterogeneous image pairs possibly used in remote sensing (see Figs. 3–5), as well as multitemporal image pairs with different

TABLE II
ACCURACY RATE OF CD ON THE EIGHT (IN LEXICOGRAPHIC ORDER) HETEROGENEOUS DATASETS OBTAINED BY THE PROPOSED METHOD AND THE STATE-OF-THE-ART MULTIMODAL CHANGE DETECTORS (FIRST UPPER PART OF EACH TABLE) AND MONOMODAL CHANGE DETECTORS (SECOND LOWER PART OF EACH TABLE)

SAR/Optical dataset (1)	Accuracy
Proposed method	**0.881**
Prendes *et al.* [9]	0.844
Correlation [9]	0.670
Mutual Inf. [9]	0.580

Optical/SAR dataset (2)	Accuracy
Proposed method	**0.943**
Prendes *et al.* [8], [47]	0.918
Prendes *et al.* [7]	0.854
Copulas [7], [10]	0.760
Correlation [7], [10]	0.688
Mutual Inf. [7], [10]	0.768
Pixel Dif. [7], [45]	0.782
Pixel Ratio [7], [45]	0.813

Optical/Optical dataset (3)	Accuracy
Proposed method	**0.877**
Prendes *et al.* [9], [47]	0.844
Correlation [9], [47]	0.679
Mutual Inf. [9], [47]	0.759
Pixel Dif. [45], [47]	0.708
Pixel Ratio [45], [47]	0.661

SAR 1-look / SAR 5-looks dataset (4)	Accuracy
Proposed method	**0.821**
Chatelain *et al.* [11]	0.732
Correlation [11]	0.521
Ratio edge [11]	0.382

VHR Optical/SAR dataset (5)	Accuracy
Proposed method	**0.743**
Gregoire *et al.* [44]	0.70

SAR 3-looks/SAR 5-looks dataset (6)	Accuracy
Proposed method	**0.840**
Chatelain *et al.* [11]	0.749
Correlation [11]	0.713
Ratio edge [11]	0.737

Optical(NIR band)/Optical dataset (7)	Accuracy
Proposed method	**0.847**
Zhang *et al.* [22]	0.975
PCC [22]	0.882

SAR/Optical dataset (8)	Accuracy
Proposed method	**0.884**
Liu *et al.* [14]	0.976
PCC [14]	0.821

TABLE III
CONFUSION MATRIX FOR THE EIGHT MULTIMODAL DATASETS, i.e., [TSX/PLEIADES] (4404×2604 PIXELS), [QB02/TSX] (2325×4135 PIXELS), [PLEIADES/WORLDVIEW 2] (2000×2000 PIXELS), [SAR ONE-LOOK/SAR FIVE-LOOKS] (762×292 PIXELS), [SPOT VHR/ERS] (1318×2359 PIXELS), [SAR THREE-LOOKS/SAR FIVE-LOOKS] (400×800 PIXELS), [OPTICAL (NIR BAND)/OPTICAL] (412×300 PIXELS), AND [SAR/OPTICAL] (921×593 PIXELS)

Multimodal pair	TP	TN	FP	FN
TSX/Pleiades	661075	9448661	1106363	251917
QB02/TSX	521245	8549723	447337	95570
Pleiades/WorldView 2	342991	3166707	226958	263344
SAR 1-look/SAR 5-looks	25082	157607	25953	13862
VHR Spot/ERS	404390	1905919	520681	278172
SAR 3-looks/SAR 5-looks	38934	230128	27525	23413
Optical(NIR band)/Optical	7024	97744	18147	685
SAR/Optical	18550	464568	59353	3682

spatial resolutions (see Tables I and II). Nevertheless, we assert with high confidence that better accuracy results are obtained on satellite image pairs with high spatial resolution (datasets 1–3 versus datasets 4–6). In fact, this peculiar feature can be easily explained if we remember that our change detector is based on a temporal gradient operator applied to a local spatial gradient (see Section II-A), which tries to detect the presence or not of a common and specific high-frequency pattern (e.g., edges, contours, microtexture, etc.) between two local regions, located at the same place, but (at different times) on different satellite images. In fact, the detection of a common and specific high-frequency pattern between the two multitemporal satellite images is necessarily more robust as the image is in high resolution.

The proposed CD model is evaluated using different imaging modalities with different noise types and levels and under different spatial resolutions along with a wide variety of change events. The evaluation shows that our CD model is flexible, but also less performing, for some cases, than some other multimodal CD models proposed in the literature dealing with a specific type of noise, imaging modalities, or type of change events (see Figs. 3–5 illustrating the applicability and the efficiency of our detector for a wide variety of cases). Nevertheless, our classification accuracy rate is comparable or outperforms some state-of-the-art approaches. We think that the flexibility of our CD model is also the result of the fact that our method does not depend, as for all machine-learning-based methods, on the content of a training base that could be biased in favor of an imaging modality type, resolution, degree of noise, or type of occurring change event and also does not depend on a specific *a priori* (generally too rigid) distribution mixtures, on which parametric statistical methods heavily relies.

Technically speaking, the first ($z_s^{D_1}$) CD operator favors sparse segmentation in terms of candidate CD regions and used alone would increase the false negative rate [see Fig. 6(a)] contrary to the second ($z_s^{D_2}$) CD operator, which, used alone, encourages diversity for detecting changes while reducing the false negative rate but increasing the false positive rate [see Fig. 6(b)]. The mixture of these two complementary CD operators has the merit to get well-balanced class accuracies instead of the use of only one of the two CD operators that would favor one of the two classes. The evolution of the average classification accuracy according to the number of (pyramid) levels shows that three levels are, in fact, a good compromise between the integration of relevant information at different resolution levels or frequencies and the loss of information due to irrelevant information or noise detected at higher scales and the loss of information due to the FastMap-based dimensionality reduction technique [see Fig. 7(b)]. Finally, the number of superpixels slightly affects the average classification accuracy because the fact that small segmentations errors can be accumulated from the SLIC segmentation algorithm applied on the before and the after images [see Fig. 7(a)].

By knowing this, a further improvement of our method would be to include a reliable high-frequency noise reduction step of

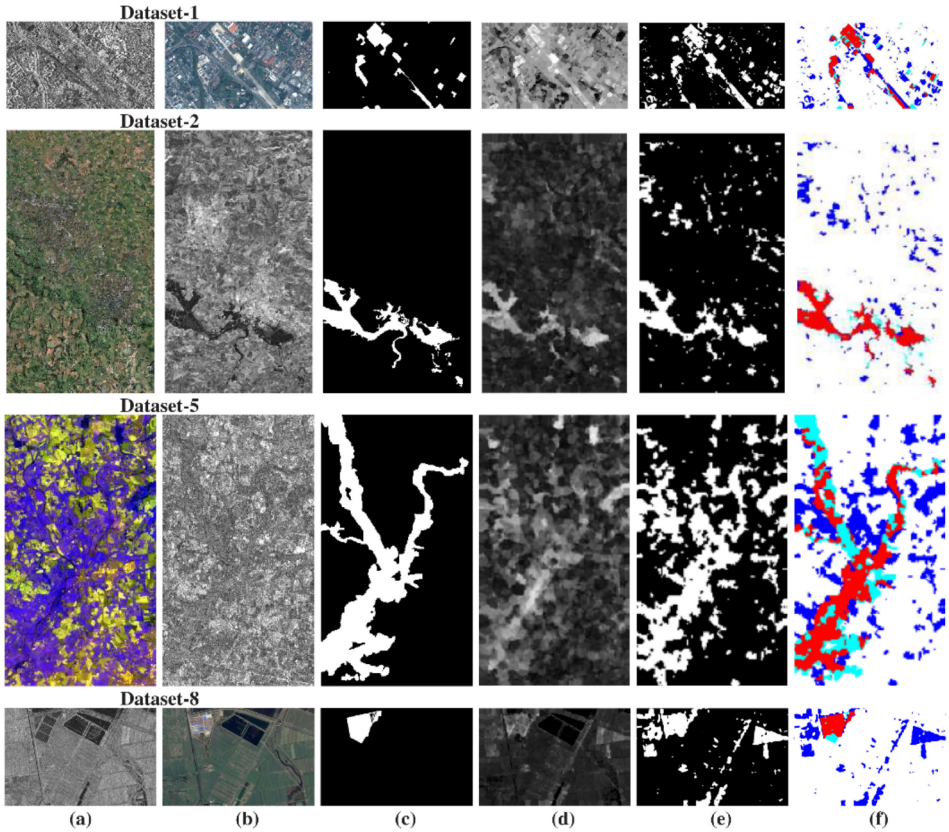

Fig. 3. Heterogeneous (multisource) optical/SAR and SAR/optical datasets: (a)–(c) image t_1, t_2, and ground truth and (d)–(f) filtered similarity map, final (changed–unchanged) segmentation result, and confusion map (white: TN, red: TP, blue: FP, and cyan: FN) obtained by the proposed approach.

the two input images, as very first preprocessing. However, let us note that finding a reliable (multimodal) denoising method in our case is not trivial, since the statistics of the noise are radically different in the case of passive optical sensors (additive and Gaussian noise) and active SAR sensors (multiplicative and speckle noise). Thus, in the case of multisource SAR+optical images, this denoising technique should be different and adaptive. It is also the case for multilooking images, in which the spatial averaging and different filtering (generally used to reduce the speckle noise) transforms the noise degradation into a mixture of independent additive and multiplicative correlated noise process, which becomes very difficult to reduce.

Fig. 8 presents a visual comparison between the CD similarity map obtained by our method and the one obtained by the SoA methods. By comparison with SoA methods [9]–[11], [44], the proposed CD method seems to visually produce more distinctive binary cluster-like structure (modeling the unchanged and changed areas) a bit more separated and more compacted (with lower internal variance within a cluster) and with less overlap. Besides, our method yields more spatially and properly regularized (or less noisy) similarity-feature maps.

The average accuracy rate obtained on the eight multimodal dataset based on the dual CD operators is 85.38%. With the CD operators expressed in formulas (1) and (2), the average accuracy rate obtained on this eight multimodal datasets is, respectively, 73.81% and 64.35%. Fig. 6 presents a visual comparison between two binary maps resulting from the application of (only) the first ($z_s^{D_1}$) and second ($z_s^{D_2}$) CD operators and visually showing how the two different binary maps complement each other (see the second row in Fig. 3)

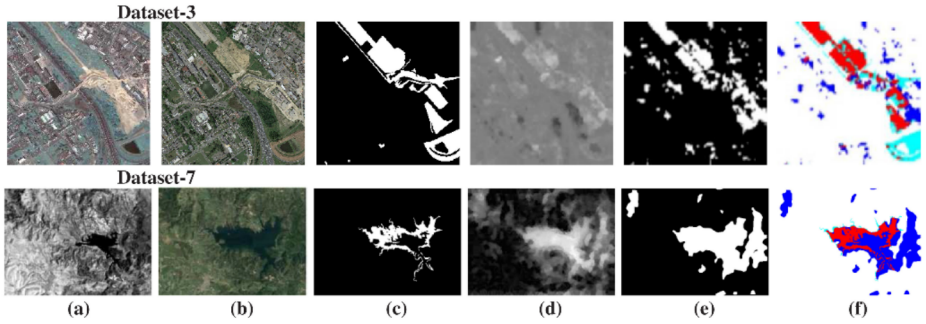

Fig. 4. Heterogeneous (multisensor) optical/optical dataset: (a)–(c) image t_1, t_2, and ground truth and (d)–(f) filtered similarity map, final (changed/unchanged) segmentation result, and confusion map (white: TN, red: TP, blue: FP, and cyan: FN) obtained by the proposed approach.

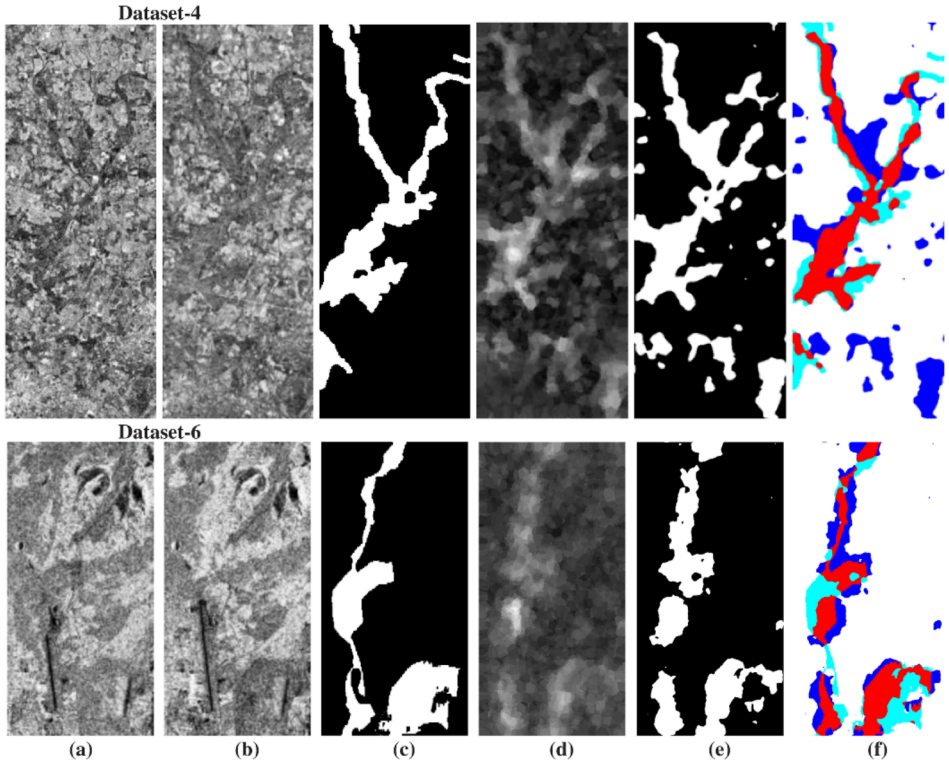

Fig. 5. Heterogeneous (multilooking) SAR/SAR datasets: (a)–(c) image t_1, t_2, and ground truth and (d)–(f) filtered similarity map, final (changed/unchanged) segmentation result, and confusion map (white: TN, red: TP, blue: FP, and cyan: FN) obtained by the proposed approach.