Université de Montréal

Détection de changement en imagerie satellitaire multimodale

par

Redha Touati

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée à la Faculté des arts et des sciences
en vue de l'obtention du grade de Philosophiae Doctor (Ph.D.)
en informatique

Avril, 2019

RÉSUMÉ

Cette recherche a pour objet l'étude de la détection de changements temporels entre deux (ou plusieurs) images satellitaires multimodales, *i.e.*, avec deux modalités d'imagerie différentes acquises par deux capteurs hétérogènes donnant pour la même scène deux images encodées différemment suivant la nature du capteur utilisé pour chacune des prises de vues. Les deux (ou multiples) images satellitaires multimodales sont prises et co-enregistrées à deux dates différentes, avant et après un évènement. Dans le cadre de cette étude, nous proposons des nouveaux modèles de détection de changement en imagerie satellitaire multimodale semi ou non supervisés. Comme première contribution, nous présentons un nouveau scénario de contraintes exprimé sur chaque paire de pixels existant dans l'image avant et après changement. Une deuxième contribution de notre travail consiste à proposer un opérateur de gradient textural spatio-temporel exprimé avec des normes complémentaires ainsi qu'une nouvelle stratégie de dé-bruitage de la carte de différence issue de cet opérateur. Une autre contribution consiste à construire un champ d'observation à partir d'une modélisation par paires de pixels et proposer une solution au sens du maximum *a posteriori*. Une quatrième contribution est proposée et consiste à construire un espace commun de caractéristiques pour les deux images hétérogènes. Notre cinquième contribution réside dans la modélisation des zones de changement comme étant des anomalies et sur l'analyse des erreurs de reconstruction dont nous proposons d'apprendre un modèle non-supervisé à partir d'une base d'apprentissage constituée seulement de zones de non-changement afin que le modèle reconstruit les motifs de non-changement avec une faible erreur. Dans la dernière contribution, nous proposons une architecture d'apprentissage par paires de pixels basée sur un réseau CNN pseudo-

siamois qui prend en entrée une paire de données au lieu d'une seule donnée et est constituée de deux flux de réseau (descripteur) CNN parallèles et partiellement non-couplés suivis d'un réseau de décision qui comprend de couche de fusion et une couche de classification au sens du critère d'entropie. Les modèles proposés s'avèrent assez flexibles pour être utilisés efficacement dans le cas des données-images monomodales.

Mots clés: Images satellitaires multimodales ou hétérogènes, optique, radar, Fastmap, deep learning, auto-encodeur, sparse, détection de changement, paires de pixels, réseau de neurones convolutionnel, opérateur invariant, détection d'anomalies.

ABSTRACT

The purpose of this research is to study the detection of temporal changes between two (or more) multimodal images satellites, *i.e.*, between two different imaging modalities acquired by two heterogeneous sensors, giving for the same scene two images encoded differently and depending on the nature of the sensor used for each acquisition. The two (or multiple) multimodal satellite images are acquired and co-registered at two different dates, usually before and after an event.

In this study, we propose new models belonging to different categories of multimodal change detection in remote sensing imagery. As a first contribution, we present a new constraint scenario expressed on every pair of pixels existing in the before and after image change. A second contribution of our work is to propose a spatio-temporal textural gradient operator expressed with complementary norms and also a new filtering strategy of the difference map resulting from this operator. Another contribution consists in constructing an observation field from a pair of pixels and to infer a solution maximum a posteriori sense. A fourth contribution is proposed which consists to build a common feature space for the two heterogeneous images. Our fifth contribution lies in the modeling of patterns of change by anomalies and on the analysis of reconstruction errors which we propose to learn a non-supervised model from a training base consisting only of patterns of no-change in order that the built model reconstruct the normal patterns (non-changes) with a small reconstruction error. In the sixth contribution, we propose a pairwise learning architecture based on a pseudo-siamese CNN network that takes as input a pair of data instead of a single data and constitutes two partly uncoupled CNN parallel network streams (descriptors) followed by a decision network that includes fusion layers and a loss layer in the sense of the

entropy criterion.
The proposed models are enough flexible to be used effectively in the monomodal change detection case.

Index terms: Multimodal satellite images, heterogeneous images, optical, radar, Fastmap, deep learning, autoencoder, sparse, change detection, pairwise pixels, convolutional neural networks, invariant operator, anomaly detection.

TABLE DES MATIÈRES

LISTE DES TABLES

LISTE DES FIGURES

LISTE DES ALGORITHMES

REMERCIEMENTS

Je tiens d'abord à remercier très chaleureusement mon directeur de recherche, le professeur Mignotte Max, pour tous les efforts qu'il a consacrés tout au long de cette thèse. Je tiens à le remercier pour sa rigueur et pour les excellentes connaissances dans le domaine du traitement d'images qu'il a su me transmettre. J'aimerais également le remercier pour ses multiples conseils et son respect sans failles des délais serrés.

Je tiens aussi à remercier le Centre de Recherche Informatique de Montréal (CRIM) d'avoir soutenu ce projet de recherche, particulièrement, mon superviseur Mohamed Dahmane, chercheur dans l'équipe d'imagerie et vision du (CRIM), pour sa bienveillance au sein de l'équipe de vision, pour tout le temps consacré et son investissement dans ce projet de recherche, pour ces diverses discussions, remarques et appréciations et pour sa sympathie. Je tiens à remercier les autres membres de l'équipe d'imagerie et vision pour leur accueil et leur ambiance de travail.

Je remercie les membres de jury d'accepter d'évaluer cette thèse.

Je tiens à remercier aussi tous les enseignants et les responsables de notre département.

Chapitre 1

INTRODUCTION GÉNÉRALE

1.1 Introduction

En télédétection spatiale, la détection de changement est un traitement numérique dont le but est d'analyser deux ou plusieurs images (acquises par des capteurs embarqués sur des plates-formes satellites) de la même zone géographique mais à différentes dates, afin de localiser et quantifier (automatiquement) les changements existant entre ces images [65] [88].

Ce traitement est très utile et a été particulièrement étudié pour différentes applications en télédétection ou en science géographique, notamment ces dernières années pour des problèmes liés au réchauffement climatique, à la déforestation, à l'évaluation des catastrophes naturelles (inondation, tsunamis, tremblement de terre, feux de forêts, etc.), pour la gestion des ressources naturelles, l'analyse du développement urbain, le suivi agricole ou pour des applications militaires, pour ne citer que quelques unes des applications potentielles.

En fait, la télédétection (spatiale) est l'ensemble des techniques qui permettent, par l'acquisition d'images, d'obtenir de l'information sur la surface de la Terre (y compris l'atmosphère et les océans), sans contact direct avec celle-ci. Actuellement, ces images de télédétection peuvent être produites par deux types de capteurs ou de satellite différents [87]. Le premier type de capteur est le capteur passif qui permet d'acquérir des images optiques ou infrarouge thermique. Leur principe général est de déceler la lumière du jour (soleil) réfléchie par les objets de la scène dans le cas des images optiques ou l'énergie dégagée naturellement, *i.e.*, l'infrarouge thermique, par

ces objets (ce type d'image peut donc être enregistrée le jour ou la nuit). Un autre type de capteur, dit actif, est celui qui incorpore un émetteur qui irradie la scène (*i.e.*, envoie lui-même des ondes électromagnétiques grâce à une antenne dans le cas d'un instrument actif radar qui permet de générer des images SAR) puis capture les ondes réfléchies sur la surface au sol pour générer des images [87]. Les capteurs actifs sont utilisés pour examiner les longueurs d'onde qui ne sont pas produites par le soleil, par exemple les hyperfréquences. Le fluoromètre laser et radar à synthèse d'ouverture (SAR) sont des exemples de capteurs actifs. Les images satellitaires sont souvent dégradées par deux différents types de bruits; bruit additif Gaussien ou multiplicatif pour les images optiques et un bruit multiplicatif de speckle pour les images SAR ou les images données par d'autres capteurs actifs.

Dans le domaine de la détection de changement, deux images de la même zone géographique prises à deux dates différentes peuvent être capturées dans un premier cas par le même satellite sous les mêmes conditions, on parle alors d'images mono-modales et de détection de changement mono-modal [48] [32].

Dans l'autre cas, les images sont dites multi-modales et peuvent être classées généralement en trois catégories [65] [88] [15] [14]: **les images multi-sources** produites par la combinaison d'une image SAR avec une image optique ou plus généralement une image acquise par un capteur passif et l'autre actif (ou *vice-versa*). **Les images multi-senseurs** résultant de la combinaison de deux images optiques ou deux images SAR acquises par deux capteurs différents ou avec le même capteur mais avec différentes spécifications, et enfin **les images SAR multi-looking** produites par la combinaison de deux images SAR dont l'image avant est une image SAR brute et l'image après est une image SAR pré-traitée ou filtrée et exhibant ainsi différents niveaux de bruit.

On peut noter que les algorithmes de détection de changement que l'on applique le plus souvent en télédétection, en science géographique ou en géomatique sont aussi

presque directement applicables en imagerie médicale pour détecter automatiquement (ou semi-automatiquement) les changements intervenant entre deux radiographies successives (par exemple) d'un même patient et pour l'éventuelle détection et quantification (puis le suivi et le traitement) d'une anomalie entre ces deux images médicales multidates. De même, on peut penser que les algorithmes de détection de changement utilisés en télédétection seront aussi directement utilisables pour les caméras de nouvelle génération fusionnant l'infrarouge et l'optique traditionnelle.

De nombreux travaux ont été proposés ces dernières années pour essayer de résoudre ce problème de détection de changement mais presque uniquement dans le cas *mono-modal* où les images, avant et après changement, résultaient en fait d'une même modalité d'imagerie (par exemple deux images SAR représentant la scène avant et après le changement dû, par exemple, par un tremblement de terre) [48] [32] [31] [119]. Très peu de travaux ont été proposés dans le cas le plus difficile, où les deux capteurs provenaient d'une modalité d'imagerie différente comme la combinaison en télédétection d'une image, avant changement, optique et d'une image, après changement, SAR (ou encore la combinaison d'une image radio-graphique et d'une image écho-graphique en imagerie médicale) [65] [88].

La plupart des méthodes proposées pour résoudre ce problème de détection de changement mono-modal (même modalité d'imagerie) se basent sur les trois étapes classiques suivantes [48] [32]:

- Le recalage géométrique et en intensité (incluant toutes les corrections possibles) de l'image acquise avant changement et l'image obtenue après changement.

- La comparaison (pixel par pixel) entre ces deux images recalées et corrigées.

- Enfin la classification ou segmentation de cette image de différence en deux classes pour chaque pixel; "aucun changement" ou "changement".

L'étape de comparaison entre les deux images recalées dans ces méthodes repose sur un opérateur de différence pour les méthodes utilisant des images optiques et sur un opérateur de log-ratio (afin d'avoir une carte de différence moins bruitée, i.e moins sensible au bruit multiplicatif) dans le cas de détection de changement entre des images SAR, et ne s'intéresse qu'a une seule modalité d'images produite par un capteur bien précis.

Ces approches basées sur une analyse et une classification de l'image de différence, obtenue pixel par pixel, entre ces deux images d'entrée sont simples et inefficaces pour des données-images issues de modalités d'imagerie différentes [88]. En effet, les images satellitaires multi-modales présentent fondamentalement des caractéristiques et statistiques différentes et concrètement, peuvent présenter des luminances ou des textures (et donc des statistiques) très différentes pour deux identiques régions terrestres représentées par deux modalités d'imagerie différentes.

Détecter les changements entre des images satellitaires hétérogènes a récemment généré un intérêt croissant pour la communauté de recherche en télédétection [51,126, 127]. Cet engouement est principalement dû au fait que cette stratégie ne nous impose aucune condition et restriction sur l'origine et les caractéristiques des données satellitaires acquises (avant et après un évènement). Elle nous permet donc d'exploiter, sans restriction, l'énorme quantité de données hétérogènes que nous pouvons maintenant obtenir à partir des différentes archives existantes incluant les nombreux types de satellites d'observation gravitant autour de la terre ainsi que les systèmes équipés des dernières et nouvelles technologies de détection qui seront inventés demain. Finalement, ajoutons que ce domaine de recherche qui peut être aussi considéré comme la généralisation du problème classique de détection de changement mono-modal [88], et qui fusionne ou combine différents types de modalité d'imagerie, possiblement complémentaire, pourrait être bénéfique pour détecter, analyser et quantifier plus précisément le changement des surfaces terrestres ayant des propriétés complexes

soumises à des conditions extrêmes (par exemple, humidité, température, feu, glace, etc.).

La détection de changement dans les images satellitaires multi-modales est un problème de traitement d'images non trivial car cette technique a pour but de combiner des images possédant des statistiques très différentes, issues de capteurs physiques possiblement très différents. Jusqu'à présent, relativement peu de travaux ont été proposés dans la littérature. Néanmoins, bien que peu nombreux, ces travaux peuvent être regroupés en cinq grandes catégories: 1- les méthodes paramétriques [14, 65, 88], 2- les méthodes non paramétriques [29], 3- les techniques utilisant des mesures de similarité invariantes par modalité d'imagerie [2], 4- les techniques de projection ou simulation permettant de transformer la paire d'images originale dans un nouvel espace commun ou une nouvelle représentation (ou modalité) commune [117] et enfin 5- celles utilisant l'apprentissage machine [51].

Dans les techniques paramétriques, un ensemble (ou un mélange) de distributions multivariées ou métagaussiennes est généralement utilisé pour modéliser les statistiques communes ou les dépendances entre les deux modalités d'imagerie et/ou les différents types de données multi-capteurs. Dans cette catégorie, on peut citer l'approche à base de copulas proposée dans [65]. Dans cette méthode, la dépendance entre les deux images satellitaires dans des zones inchangées est modélisée par une régression quantile. La régression est appliquée selon la théorie des copules et les comparaisons fondées sur les mesures statistiques du type Kullback-Leibler pour générer une carte de similarité qui est ensuite analysée par seuillage pour détecter les zones de changement et de non-changement. Une approche statistique multivariées intéressante en deux étapes a été également proposée dans [88] [90] [89] dont la première étape vise à estimer un modèle physique, basé sur un mélange de distributions multidimensionnelles (prenant en compte le modèle de bruit, les relations entre les réponses des capteurs aux objets de la scène et leurs propriétés physiques),

dont les paramètres sont estimés par l'algorithme d'espérance-maximisation (EM) [18]. Un test statistique basé sur ce modèle permet alors d'estimer les changements. Dans le même esprit, les auteurs de [14] proposent également d'une estimation d'un mélange de distribution multidimensionnelle basée sur une nouvelle famille de distributions multivariées avec différents paramètres de forme et particulièrement bien adaptée à la détection des modifications ou changements des images SAR acquises par différents capteurs ayant différents niveaux de bruit. Le problème de ces techniques paramétriques réside dans le fait qu'elles ont été généralement spécialement conçues (par des types de distribution spécifiques) pour un type de paire de capteurs multi-modal (optique/SAR dans [65] [88] [90] ou SAR avec différents niveaux de bruit [14]). Par conséquent, ces méthodes ne sont pas toujours facilement généralisables pour une autre paire de capteurs différents. En outre, les méthodes qui ont été proposées sont quelquefois semi-supervisées car elles nécessitent généralement la disponibilité de deux images associées à une zone inchangée [65] [88] [90]. Enfin, ces méthodes nécessitent également une étape d'estimation des paramètres des lois de distribution considérées aux sens du maximum de vraisemblance (MV) laquelle peut être complexe et coûteuse en calcul.

Dans la catégorie des méthodes non paramétriques, on peut citer [29] qui combine les résultats de la segmentation pré et post-événement de l'image satellitaire avec une extension du Fuzzy C-means (FCM) exprimée dans le cadre théorique formel des incertitudes et des fonctions de croyance. Une stratégie similaire est également proposée dans [15]. Dans le même ordre d'idées, dans [120], les auteurs décrivent une stratégie de pré-segmentation basée sur l'indice spectral de différence normalisée. Les méthodes non paramétriques ont la capacité de s'adapter à une grande variété de modalités d'imagerie différentes mais sont aussi généralement moins précises qu'un modèle paramétrique traitant un type spécifique de multi-modalité et représenté par une distribution particulière dont la forme est spécifiquement adaptée à ce type de

multimodalité.

Dans la troisième famille de méthodes, utilisant des mesures de similarité invariantes par modalité d'imagerie, Alberga *et al.* [2] propose d'utiliser une technique proche de celle utilisée pour le recalage multimodal entre différentes images. Leur méthode est basée sur l'utilisation d'une combinaison de différentes mesures de similarité invariantes (telles que le rapport de corrélation, l'information mutuelle, etc.) afin d'estimer dans un premier temps, la correspondance entre les mêmes points existant dans les deux images et ensuite identifier et détecter, dans un deuxième temps les zones de changements existants entre les deux images hétérogènes.

L'intérêt principal de cette famille de méthodes repose sur le fait qu'elle n'est pas étroitement liée à un cadre mathématique particulier (analyse Bayésienne ou multivariée ou réduction de dimensionnalité pour la première ou deuxième catégorie ou analyse de régression pour la quatrième ou la cinquième catégorie). Elle est plus flexible, mais sa simplicité et l'absence de cadre mathématique formel rend difficile l'étude des propriétés de cette famille de méthodes et ses possibles améliorations.

La quatrième catégorie regroupe les techniques de projection ou simulation permettant de transformer la paire d'images originales dans une nouvelle représentation commune ou encore de projeter l'une des deux images dans la modalité d'imagerie associée à l'autre image, [117] propose de transformer la paire d'images originales dans un nouvel espace commun ou une nouvelle représentation, conçue particulièrement pour être invariant à la modalité d'imagerie et visant à mettre en évidence les changements. Dans le même esprit, Volpi *et al.* [113] tentent de trouver des projections jointes des images d'entrée en maximisant la corrélation entre les données projetées. Dans [8], les auteurs proposent une méthode de détection de changement multimodal optique/SAR pour quantifier les dommages causés, par un tremblement de terre, à chaque maison ou bâtiment individuel. À cette fin, les paramètres de chaque bâtiment sont tout d'abord estimés à partir de l'image optique et combinés avec les paramètres

d'acquisition de l'image SAR post-événement pour prédire (à l'aide de simulation) la représentation SAR du bâtiment intact. Cette image SAR simulée est ensuite comparée, en utilisant une mesure de similarité, à l'image SAR réelle afin de quantifier les dommages causés à chaque bâtiment.

Finalement, dans la catégorie des méthodes d'apprentissage automatique (machine learning) les auteurs dans [66] emploient un algorithme d'apprentissage non supervisé, appelé réseau adverse génératif (generative adversarial network) constitué de deux réseaux dont le premier génère une carte binaire et le deuxième essaie de discriminer entre le résultat du générateur et le résultat d'un algorithme de binarisation. Dans [51], les auteurs proposent d'entraîner un couple de réseaux de neurones à convolution afin de transformer l'image avant et après changement dans un espace de caractéristiques permettant de calculer une carte de différence, ensuite d'appliquer un algorithme de seuillage sur cette carte pour générer la carte de détection finale. Dans la même optique que [51], les auteurs dans [127], proposent de construire un réseau de neurones symétrique constitué d'une machine de Boltzmann restreinte [127], dont les paramètres sont ensuite mis à jour en se basant sur le résultat de clustering. Une autre méthode basée sur un réseau autoencodeur débruiteur utilise des caractéristiques sélectionnées de l'image de différence pour entrainer le réseau [126].

L'objectif de cette thèse est de proposer des algorithmes automatiques de détection de changement qui pourront aussi être possiblement semi-supervisés permettant de localiser et quantifier les changements existants entre deux images de la même zone géographique mais acquises par des capteurs hétérogènes possiblement très différents. De plus, on s'intéressera aux modèles de détection de changement temporels assez flexibles pour être aussi utilisés efficacement dans le cas général (et plus simple) des données-images mono-modales. Finalement, on s'intéressera aussi aux modèles de détection multimodale de changements temporels qui pourront possiblement et efficacement utiliser plus que deux images multimodales, multidates (avant et après

changement) et qui seront consistants, au sens statistique du terme (*i.e.*, qui devront générer des cartes de détection de changements temporels autant plus précises et fiables que le nombre de données-images multidates augmentera).

Plus précisément, nous proposons, dans le cadre de cette thèse, différents modèles non supervisés ou semi-supervisés pour la détection de changement dans une série d'images temporelles issues de différentes modalités qui s'inscrivent parmi les cinq catégories précédentes.

- Un modèle non paramétrique de détection de changement multimodale reposant sur une toute nouvelle modélisation utilisant toutes les paires de pixels de l'image est présenté dans le chapitre 2. Il permet d'exprimer pour chacune de ces paires de pixels une contrainte adaptée permettant dans un premier temps d'estimer une image de différence qui sera robuste aux différents type de modalité d'imagerie des deux images satellitaires avant et après changement. La complexité quadratique de l'estimation de cette image de différence sera réduite à une complexité linéaire grâce à une méthode d'estimation au sens des moindres carrés (de la prise en compte de toutes ces contraintes) par une technique d'optimisation qui sera adaptée de celle proposée par la technique du FastMap de Faloutsos [26]. À cette fin, une technique originale de vote majoritaire permettant de fusionner (ou combiner) plusieurs cartes de segmentations binaires résultantes de différentes stratégies de binarisation automatique afin d'obtenir une carte de segmentation fiable et sans supervision constitue la deuxième originalité de ce travail. Le modèle proposé est à la fois simple et efficace aussi dans le cas de détection de changement entre des images mono-modales.

- Un nouveau modèle reposant sur un opérateur de gradient textural spatio-temporel, invariant aux modalités d'imagerie, exprimé par des normes duales (complémentaires) et détectant à différentes échelles, les différences (en termes de hautes fréquences) de chaque région structurelle existant dans les deux im-

ages satellites hétérogènes est proposé dans le chapitre 3. Cette détection de différence donne une carte de similarité qui est ensuite dé-bruitée par un filtre spatial adaptatif utilisant les régions homogènes communes préalablement segmentées des deux images satellitaires d'entrée. Finalement, la segmentation non supervisée en deux classes de cette carte de similarité filtrée nous permet d'obtenir une carte de segmentation fiable contenant les zones de changement. Le modèle proposé s'avère aussi assez flexible pour être utilisé efficacement dans le cas de modalités d'imagerie très différentes acquises sous différentes conditions, *i.e.*, optique/radar, optique/optique, et radar/radar.

- Le chapitre 4 présente un modèle paramétrique et plus précisément une approche statistique Bayésienne Markovienne au problème de la détection de changement multimodal. La principale nouveauté de ce modèle Markovien réside dans l'utilisation d'un champ d'observations constituées d'une modélisation spatiale par paires de pixels. Une telle modélisation nous permet d'estimer comme donnée d'observation, un indice visuel robuste et quasi invariant par rapport à la (multi-) modalité d'imagerie. Pour utiliser cette donnée d'observation dans le cadre d'un modèle stochastique de vraisemblance, nous avons utilisé un algorithme d'estimation itératif qui tient compte de la variété des lois dans le mélange de lois de vraisemblance et estime les paramètres de ce mélange de distributions au sens du maximum de vraisemblance. Une fois cette étape d'estimation terminée, la solution au sens du maximum a posteriori (MAP) de la carte de détection des changements, basée sur les paramètres estimés précédemment, est ensuite calculée par un processus d'optimisation stochastique.

- Le chapitre 5 présente une nouvelle méthode basée sur la projection des deux images satellitaires dans un espace de caractéristiques commun, dans lequel les deux images hétérogènes partageront les mêmes propriétés statistiques et sur

lesquelles les méthodes classiques de détection de changement mono-modal peuvent être appliquées. Cette transformation des images avant et après est principalement basée sur une représentation de positionnement multidimensionnel (MDS) des données.

- Le chapitre 6 nous décrit une méthode de détection de changement multimodal modélisés comme étant des anomalies entre les deux images satellitaires. Le modèle proposé est basé sur l'apprentissage non supervisé des motifs hétérogènes de la classe normale, (*i.e,*, non-changée) dans un espace résiduel. Pour classer les pixels d'une nouvelle paire d'images, le modèle élaboré encode la représentation de l'entrée dans l'espace latent, *i.e.*, l'espace compact normal, ensuite reconstruit la représentation encodée à partir d'une représentation latente. L'analyse des erreurs de reconstruction permet d'identifier les zones qui ont une grande erreur de reconstruction comme des anomalies, *i.e.*, comme des zones de changement.

- Finalement, le chapitre 7 repose sur une approche d'apprentissage par paires d'entrées et un réseau pseudo siamois dont l'architecture est basée sur deux flux de réseau parallèles et partiellement non couplés. Chaque flux de réseau est lui-même un réseau de neurones convolutionnel (CNN) qui extrait un descripteur de chaque patch d'entrée. Le modèle de détection de changement comprend une étape de fusion qui concatène ses deux descripteurs de sortie dans une seule représentation multimodale, qui est ensuite réduite dans une faible dimension en utilisant conjointement des couches entièrement connectées et une fonction de perte binaire basée sur l'entropie croisée utilisée dans la couche finale. Le modèle est capable de capturer les dépendances spatiales et temporelles entre les paires d'images en entrées grâce à cette architecture d'apprentissage par paires qui prend en entrée une paire de patchs au lieu d'un seul patch. Le modèle proposé n'exige aucune sélection préalable de mélange de distributions

spécifique pour telle, ou telle modalité.

Le plan de la thèse est structuré comme suit: après notre introduction (chapitre 1), les chapitres 2 à 7 proposent six modèles de détection de changement multimodal dont le premier (chapitre 2) appartient à la famille des méthodes non paramétriques, le deuxième (chapitre 3) appartient à la classe des méthodes utilisant des mesures de similarité invariantes par modalité d'imagerie, le troisième (chapitre 4) s'inscrit dans les modèles paramétriques, le quatrième (chapitre 5) fait partie des techniques de projection ou simulation permettant de transformer la paire d'images originales dans un nouvel espace commun ou une nouvelle représentation (ou modalité) commune, et enfin les deux derniers (chapitre 6 et 7) proposent des méthodes utilisant l'apprentissage machine.

Publications et soumissions:

- R. Touati and M. Mignotte, "An energy-based model encoding non-local pairwise pixel interactions for multi-sensor change detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 1, February 2018

- R. Touati, M. Mignotte, M. Dahmane. A Reliable Mixed-Norm based Multiresolution Change Detector in Heterogeneous Remote Sensing Images. IEEE Journal of Selected topics in Applied earth Observations and Remote sensing (j-stars), January 2019, accepté.

- R. Touati, M. Mignotte, M. Dahmane. Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise-Based Markov Random Field Model. IEE Transactions on Image Processing, Mai 2019, accepté

- R. Touati, M. Mignotte, and M. Dahmane, "Change detection in heterogeneous remote sensing images based on an imaging modality-invariant mds represen-

tation, in 25th IEEE International Conference on Image Processing (ICIP'18), Athens, Greece, October 2018.

- R. Touati, M. Mignotte, M. Dahmane. Anomaly Feature Learning For Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model, IEEE Journal of Selected topics in Applied earth Observations and Remote sensing (j-stars), june 2019, révisé

- R. Touati, M. Mignotte, M. Dahmane. Pairwise Descriptors Learning For Multimodal Change Detection using Pseudo-Siamese CNN Network Model, IEEE Geoscience and Remote Sensing Letters (GRSL), March 2019, soumis

Chapitre 2

AN ENERGY-BASED MODEL ENCODING NON-LOCAL PAIRWISE PIXEL INTERACTIONS FOR MULTI-SENSOR CHANGE DETECTION

Dans ce chapitre, nous présentons notre article publié dans la revue *I*EEE Transactions on Geoscience and Remote Sensing, intitulé: **An Energy-Based Model Encoding Non-Local Pairwise Pixel Interactions For Multi-Sensor Change Detection** . Nous exposons ce dernier dans sa langue originale de publication.

Abstract

Image change detection is a challenging problem, particularly when images come from different sensors. In this paper, we present a novel and reliable change detection model which is first based on the estimation of a robust similarity-feature map generated from a pair of bi-temporal heterogeneous remote sensing images. This similarity-feature map, which is supposed to represent the difference between the multi-temporal multi-sensor images, is herein defined, by specifying a set of linear equality constraints, expressed for each pair of pixels existing in the before-and-after satellite images acquired through different modalities. An estimation of this over-constrained problem, also formulated as a non-local pairwise energy-based model, is then carried out, in the least square sense, by a fast linear-complexity algorithm based on a multidimensional scaling (MDS) mapping technique. Finally, the fusion of different binary segmentation results, obtained from this similarity-feature map by different automatic thresholding algorithms, allows us to precisely and automatically classify the changed and unchanged regions. The proposed method is tested

on satellite data sets acquired by real heterogeneous sensor and the results obtained demonstrate the robustness of the proposed model compared to the best existing state-of-the-art multi-modal change detection methods recently proposed in the literature.

2.1 Introduction

Nowadays, change detection (CD) is a major application and also an active research topic in remote sensing image processing since it plays an important role in various application domains, including environmental monitoring, deforestation, urban planning, land or natural disaster/damage monitoring and management to name a few. Until now, many change detection approaches have been proposed for addressing the classical *mono-modal* change detection issue [9, 10, 32, 33, 50] which occurs when the pairs of images are obtained from the same sensor or, more generally, the same imaging modality. In this *mono-modal* case, the two images, recorded at two different times but under similar imaging conditions, are generally first coregistred and corrected (preprocessing) and then, most often, used to generate a difference image by differencing or (log-)rationing. Finally the resulting difference image is then segmented into two classes to distinguish changes of interest of the land cover/land use.

A less explored and more challenging problem is the so-called *multi-modal* change detection problem which is based on pairs of images obtained from different imaging modalities. In this case, the two input (before and after change) images present

[1] By changes of interest of the land cover, it must be understood that we do not seek to detect, in this work, changes such as atmospheric effects, including haze, persistent cloud cover, phenological changes, thin snow or ice cover, soil moisture, shadow, etc. In this work, we are just referring to land cover changes such as major construction or excavations, flooding, earthquake, deforestation, etc.

radically different image statistics (along with possibly different spatial & spectral resolutions) which cannot be compared with traditional methods borrowed from *mono-modal* change detection approaches relying on a simple pixelwise difference model. Multi-modal change detection is especially appealing for several reasons. In facts, in furthermore to generalize the *mono-modal* case, this processing treatment has obviously less restrictive considerations about the formation of the input data pair since it must adapt itself to the characteristics of data with different natures. As a consequence, it should be more robust to natural variations in environmental variables such as soil moisture or phenological state (such as flowering, maturing, drying, senescence, harvesting, etc.) that cannot be avoided and well taken into account and corrected in the preprocessing step of a classical *mono-modal* change detection approach. Another interest is its inherent practicality that it could bring in several emergency situations. For example, it is useful in the case when an optical image of a given area is provided by an available remote sensing image archive data, and only a new Synthetic Aperture Radar (SAR) image can be acquired (for technical reasons, lack of time, availability or atmospheric conditions) in an emergency situation for the same area. In addition to providing a wide variety of information and properties about the study area, let us stress out that the additional information provided by two different sensors, could also be used to our advantage, to improve the accuracy of the final change detection map. This can be efficiently achieved if one succeeds in modeling the complementary and supplementary information provided by the two different imaging modalities, with modelling techniques borrowed, for example, from data fusion-based classification theory. Finally, let us also mention that this *multi-modal* approach may be useful and sometimes indispensable in some specific cases, such as forest monitoring in tropical or boreal areas for which SAR, thanks to its ability to penetrate heavy clouds and fog, is often used as a complement to optical data. Another example where SAR and optical sensors are complementary, is the case of frequently snow-covered regions of high altitudes since SAR is also able to

penetrate a thin snow layer.

Up to now, relatively few research works have been developed in change detection using heterogeneous remote sensing images. Among the few existing models proposed in the literature, we can however, mention the theoretical approach proposed in [65]. In this study, the model is based on the assumption that some dependence indeed exists between the two images in unchanged areas and more precisely rely on the estimation of the local statistics of the first image through the point of view (in a statistical sense) of the second one (and vice versa). This dependence is modeled by quantile regression applied according to the copula theory and Kullkack-Leibler-based comparisons of these above-mentioned local statistics are applied to define a change measure, which is then finally analyzed by thresholding, in order to detect between change and no change areas. However, this method remains supervised since it requires to learn the cumulative distribution function of the pixel intensity in the after image, conditioned to "no-change" hypothesis (*i.e.* the so-called copula) by using a manually selected (carefully chosen) training set of samples in the after image. Let us note that the model is also not easily generalizable in the case when more than one image, before and after a given event, is available and also not well suited (in terms of modeling, speed and efficiency) to be used for images acquired with homogeneous sensors. Another recent study was proposed by Prendes *et al.* [88] to overcome multi-sensor variability problems in change detection. The authors propose an interesting multivariate statistical approach aiming to estimate a physical model, based on a mixture of multi-dimensional distributions which both takes into account the relationships between the sensor responses to the objects contained in the observed scene, the physical properties of these objects and the statistical properties of the noise corrupting the images. The parameters of this multidimensional mixture model are estimated by the expectation-maximization (EM) algorithm [18] which are then subsequently used to infer the relationships between the sensor physical

properties involved through manifold learning. A statistical test based on this model allows to estimate the changes. An extension of this model, taking advantage of the correlations between adjacent pixels *via* a Markov Random Field model, has also been proposed by the same authors [90]. However, this method also assumes a training set and more precisely that two training images associated with an unchanged area are available. Also the method has been designed for heterogeneous multi-sensor in the case of optical/SAR data and is not easily generalizable for another pair of different sensors. Besides, it requires a lot of EM estimations (nearly one for each pixel), each one relying on a good unsupervised estimation of the optimal number of existing components. Another CD approach for heterogeneous multi-sensor SAR data based on multi-dimensional distribution mixture estimation has also been proposed in [14]. In particular, the authors have studied a new family of multivariate distributions whose margins are univariate gamma distributions with different shape parameters referred to as multi-sensor multivariate gamma distributions (MuMGDs) which are well suited for detecting changes in SAR images acquired by different sensors having different numbers of looks. The parameters of this multidimensional mixture model are estimated by the Maximum Likelihood (ML) or Inference Function for Margins (IFM) algorithm. Also the method, and more precisely the family of MuMGDs has been especially designed for heterogeneous SAR sensors and can not be easily generalized for other or different sensors. Finally [2] proposes to use a methodology borrowed from co-registration (used in the field of medical imagery), based on the use of similarity measures (such as correlation ratio, mutual information, etc.) and to use the correspondence between the same points in the two images to detect eventual changes existing between the two data acquisitions. A comparison between the results of the performance of tested similarity measures is reported which indicates that the mutual information and the Cluster Reward Algorithm (CRA) seems the best indicator for *multi-modal* (optical/SAR) change detection. The CRA measure is built from the joint and the marginal probabilities, as the mutual information, and

has a large value when the joint histogram has little dispersion (thus indicating a good correlation score). Nevertheless, these two measures are sensitive to the dimension of the estimation windows used for the pixel statistics and the similarity measure calculation.

Contrary to change detection techniques based on a classical pixelwise modeling approach, we propose, as first and main contribution of this work, a new change detector relying on the set of all pairs of (possibly non-local) pixels existing in the before and after remote sensing images. This allows us to build a robust similarity feature map, especially well suited to estimate the difference between heterogeneous sensors exhibiting radically different image statistics. In our model, a set of linear equality constraints is expressed for each pair of pixels (in terms of grey levels or local statistics difference) and this over-constrained problem is then embedded or formulated into a final energy-based model encoding all the local pairwise pixel interactions. The quadratic complexity in the number of pixels of this resulting energy-based model, is reduced to a linear complexity procedure, thanks to the FastMap-based optimization procedure proposed by Falutsos and Lin [26]. This technique acts as an efficient and fast global minimizer of the cost function, integrating all the pairwise constraints, of our model by performing geometric linear projections (using the cosine law) in a n-dimensional space over an axis defined by a pair of pixels from the image (or in our application, from a pair of images) called pivots. Conceptually, the FastMap treats each distance or constraints between a pair of pixels (in terms of grey level difference) as a *spring* between the pixels, and tries to rearrange the grey level values of each pixel to minimize the stress of the springs (*also called the stress function*) or equivalently to satisfy all the constraints in the least square (LSQ) sense. Moreover, as second contribution, changed and unchanged areas are then finally identified, from this latter similarity feature map, by fusing the results of different automatic thresholding algorithms. In this way, we efficiently combine the intrinsic properties and

criteria related to the different automatic thresholding algorithms in order to further increase the robustness and reliability of our *multi-modal* change detection strategy.

Let us note that, within the FastMap-based optimization and energy-based model framework encoding the non-local pairwise pixel interactions, we can mention the recent gait analysis model proposed in [75], which allows us to convert a video sequence of depth images of a human gait (on a treadmill) into an informative color map providing a quick overview of asymmetry existing in a given gait cycle for a rapid clinical diagnosis. In this model, using a gait video datacube, the pairwise interactions are defined to encode the degree of similarity existing between two gait movements (represented by two temporal depth signals) taken on two different locations on a human's body surface walking on a treadmill and such that the (pairwise) distance is defined as zero if the two motions are either pointwise similar or in perfect phase opposition (*i.e.*, with a phase difference of half a gait cycle as it is normally the case for legs and arms during the gait cycle of a healthy subject). The set of distances between each pair of pixels is then used by the FastMap algorithm to generate a final mapping in which these distances should then code (as constraint) the L2-norm of color difference existing between this pair of pixels. By this means, two pixels (or two points located on the human's body surface) that shares the same color on this mapping have to be considered as symmetric (and conversely, all the more anti-symmetric as their color difference is high).

The remainder of this paper is organized as follows: Section 7.2 and 2.3 describe respectively the proposed change detection technique and the optimization procedure related to this model which allows us to estimate the similarity-feature map, from which changed and unchanged areas are then identified in Section 2.4 by combining the results of different automatic thresholding algorithms. Section 7.3 presents a set of experimental results and comparisons with existing *multi-modal* and *mono-modal* change detection algorithms. Finally, Section 7.4 concludes the paper.

2.2 Proposed Change Detection Model

Let us consider two (co-registered) bi-temporal remote sensing (N pixel size) images, y^{t_1} and y^{t_2} acquired at two times (before and after a given event), in the same geographical area, from different sensors or from the same sensor but without the correction step, in terms of radiometric, atmospheric and distortion consistencies and characteristics.

In order to estimate y^D, the similarity feature map, which is supposed to represent the difference between the multi-temporal (multi-sensor) images, we rely on an improved version of the model introduced in [108] for the *mono-modal* change detection problem. In this model, first, we have to specify an over-determined set of constraints to be satisfied (for y^D) and expressed for each pair of pixels $<s,t>$ existing in each of the two multi-temporal images y^{t_1} and y^{t_2}. The similarity map y^D is then seen as a solution to this set of constraints *via* the following non-local pairwise cost function to be optimized:

$$\hat{y}^D = \arg\min_{y^D} \sum_{<s,t>_{s\neq t}} \left(\beta_{s,t} - \left\|y_s^D - y_t^D\right\|_2\right)^2 \tag{2.1}$$

where the summation is done over all the pairs of pixels existing in the similarity feature image y^D to be estimated and $\|.\|_2$ is the Euclidean distance. In Eq. (2.1), the set of $\beta_{s,t}$, represents the set of $N(N-1)/2$ equality constraints expressed for each pair of pixels $<s,t>$, in terms of difference of grey levels (or local statistics), in order to obtain a reliable similarity feature image y^D in which unchanged pixels will be associated to small gray-level values whereas changed pixels will present rather

large values [2] . These constraints are the following:

First, let us assume that two distinct pixels at locations s and t belong to the class *urban* at time t_1 and still belong to the same class (*urban*), at time t_2. In this case, these two pixels should both belong to the (same) class label *unchanged area* in the binary segmentation of y^D. Let us consider another scenario: let us assume that two distinct pixels at locations s and t belong to the class *urban* at time t_1 and both belong to the class *river*, at time t_2 (due to a flooding event). In this case, these two pixels should both belong to the (same) class label *changed area* in the binary segmentation of y^D. This two scenario can be summarized, as first constraint, as follows:

CONSTRAINT #1: Two distinct pixels $<s,t>$ should both belong to the class label *unchanged pixels* or both belong to the class label *changed pixels*, in the binary segmentation of y^D, if $y_s^{t_1}$ and $y_t^{t_1}$ have similar grey level (or similar local statistics), **and** if $y_s^{t_2}$ and $y_t^{t_2}$ have also similar grey level (or similar local statistics). To satisfied this constraint, y_s^D and y_t^D should both be assigned to a small grey-level value in y^D or should both be assigned to a high grey-level value in y^D (since $s \,\&\, t$ should finally share the same label in the binary segmentation of y^D) or equivalently, this constraint requires that the grey level difference between y_s^D and y_t^D is small.

Conversely, if the two pixels at locations s and t belong to a same class at time t_1 (for example *urban*) and a different class at t_2 (for example *urban* for pixel s and *river* for pixel t) or conversely. In this case, these two pixels should belong to a different

[2] Let us note that our model can handle separately the individual channels or bands of a multi- or hyper spectral sensor system since, in our energy-based model, the difference between each pair of pixels can be formulated as a Euclidean distance between two d-dimensional spectral vectors with d being the number of spectral bands. By handling the bands separately, the similarity-feature map $\hat{y}^D$ is estimated according to a similar (but opposite) criterion (*i.e.*, difference of "preservation of spectral distance") as the one often used as a criterion in the compression of hyperspectral images [69].

class label in the binary segmentation of y^D; *i.e., unchanged pixels* for one of the two pixels and *changed pixels* for the other. This belonging to different class labels, in the binary segmentation of y^D, requires that two different (grey-level) values to be assigned to these two pixels in y^D (so that the binary segmentation of y^D correctly assigns two different classes to these two pixels). This leads us to the CONSTRAINT #2.

The third and last case which leads us to the CONSTRAINT #3, involves a situation in which two pixels $<s,t>$ belong to a pair of different classes at time t_1 (for example *urban* for s and *vegetation* for t) and also belong to a pair of different classes, different from the first pair, at time t_2 (for example *urban* for s and *river* for t) or conversely. In this case, $<s,t>$ should also belong to a different class label in the binary segmentation of y^D and this requires that the grey level difference between y_s^D and y_t^D is high (see Fig. 2.1).

In summary, the three above-specified constraints, in terms of pairwise grey level difference in y^D, for each pair of locations $<s,t>$, can be quite well satisfied by using (in Eq. (2.1)) the following pairwise distance between pixels $<s,t>$ at time t_1 and t_2 (which was empirically found and inspired from the max-Symmetric χ^2 distance combined with the city block distance [12]):

$$\begin{aligned}\beta_{s,t} \;=\; \Bigg| & \max\left(\frac{\left|y_s^{t_1}-y_t^{t_1}\right|}{y_s^{t_1}}, \frac{\left|y_s^{t_1}-y_t^{t_1}\right|}{y_t^{t_1}}\right) \\ & - \max\left(\frac{\left|y_s^{t_2}-y_t^{t_2}\right|}{y_s^{t_2}}, \frac{\left|y_s^{t_2}-y_t^{t_2}\right|}{y_t^{t_2}}\right)\Bigg| \end{aligned} \tag{2.2}$$

where we recall that $y_s^{t_1}$ and $y_s^{t_2}$ is respectively the grey level (or a local statistics vector) at pixel s in, respectively, the *before* and *after* image (*i.e.,* at time t_1 & t_2). In our model, Eq. (2.1) thus become a composite cost function encoding our $N(N-1)/2$ constraints given by the *observed data* composed of all the pairwise pixels existing in y^{t_1} and y^{t_2}. Optimization of Eq. (2.1) will ensure a robust similarity feature

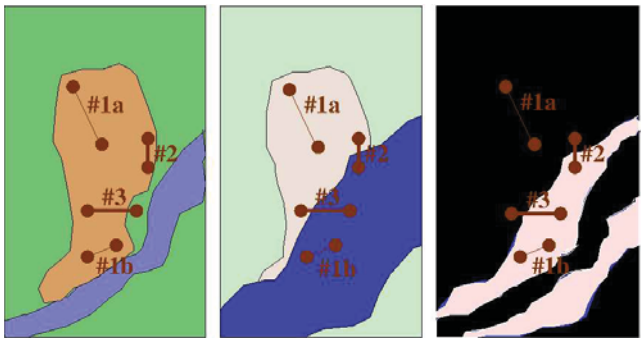

image t_1 image t_2 similarity map

Figure 2.1. Illustration of the four constraints ($\#1a$, $\#1b$, $\#2$, $\#3$) corresponding to the scenario described in Section 7.2. From left to right; image at time t_1 before a flooding event, (with the $urban$ region at the center, the $vegetation$ region all around the image and the $river$ region represented by a narrow, elongated region at bottom right of the image), image at time t_2 after a flooding event, and (ideal binarized) similarity map y^D (with the white region corresponding to the $changed\ area$) with the link (between each pair of pixels considered) drawn in such a way that its thickness is proportional to the associated distance defined by Eq. (2.2) between the grey levels (or local statistics vector) of each considered pair of pixels.

map $\hat{y}^D$ with land cover changes presenting significantly different values from those associated to the pixels belonging to unchanged areas. Nevertheless, it is important to note that the estimation of $\hat{y}^D$, according to Eq. (2.1), does not necessarily ensure that low grey-level value is assigned for the pixel belonging to the *unchanged area* and conversely that high grey-level values are assigned for the *changed area*. It could be the opposite. Nevertheless, let us mention that this latter case can be easily and automatically detected with a correlation metric or more simply by assuming that the land cover change is often much smaller than the unchanged area and once detected, we can easily remedy it by simply inverting the grey level values of the estimated image $\hat{y}^D$ before its (binary) segmentation (see Section 2.4).

Let us note that the major advantage of the proposed model lies in its flexibility given by its LSQR criterion. Indeed, contrary to Maximum a posteriori (MAP) and ML approaches [14, 88] the proposed model does not require an explicit knowledge of the data distribution and also a (ML) parameter estimation step of these distribution laws, which can be complex and/or of very different natures, in the *multi-modal* CD case, since the images (before and after the change) are from different modalities. Besides, contrary to machine learning-based approaches, the proposed scheme does not also require a large and representative and supervised training set. In addition, contrary to recent methods that seek to transform the original pair of temporal images into a new feature space or representation that significantly highlights the changes and which may be somewhat regarded as a CD method which could be invariant to imaging modality [117], the proposed method has also the advantage of not requiring the same number of spectral bands for the two (temporal) satellite images (as it is most often the case in practice for multi-modal CD since the two imaging modalities are assumed to be different).

Table 2.1. Accuracy rate of change detection on the four heterogeneous datasets obtained by the proposed method and the state-of-the-art *multi-modal* change detectors (first upper part of each table) and *mono-modal* change detectors (second lower part of each table).

TSX / Pleiades	Accuracy
Proposed method	**0.867**
Prendes *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

QB02 / TSX	Accuracy
Proposed method	**0.949**
Prendes *et al.* [87, 90]	0.918
Prendes *et al.* [88]	0.854
Copulas [65, 88]	0.760
Correlation [65, 88]	0.688
Mutual Inf. [65, 88]	0.768
Pixel Dif. [88, 104]	0.782
Pixel Ratio [88, 104]	0.813

Pleiades / WorldView2	Accuracy
Proposed method	**0.853**
Prendes *et al.* [87, 89]	0.844
Correlation [87, 89]	0.679
Mutual Inf. [87, 89]	0.759
Pixel Dif. [87, 104]	0.708
Pixel Ratio [87, 104]	0.661

SAR 1-look / SAR 5-looks	Accuracy
Proposed method	**0.781**
Chatelain *et al.* [14]	0.732
Correlation [14]	0.521
Ratio edge [14]	0.382

2.3 FastMap-Based Model Optimization

Let us note that the function to be minimized (Eq. (2.1)) is also the so-called *stress* function used as a criterion in the mapping based on multidimensional scaling (MDS) technique [106] [16]. MDS has already been successfully used in a number of practical applications, such as color image segmentation [68, 71], hyperspectral compression [69], asymmetry detection [75], human action recognition [107], and database

browsing and visualization [42] to name a few.

In our case, MDS is able to estimate a mapping, *i.e.* a grey-level similarity image y^D, such that the distances between each pair of grey-level values associated to pixels s and t are close of $\beta_{s,t}$ as faithfully as possible (in the least square sense). Nevertheless, the originally proposed MDS algorithm (called *metric MDS* [106] [16]) is not appropriate in our application (and more generally for all large scale applications) because it requires an entire $N \times N$ distance matrix to be stored in memory with a $O(N^2)$ complexity (N being the number of pixels). Instead, we have herein used a fast alternative, called FastMap [26] whose main advantage is its linear complexity (thanks to a Nyström [78] approximation of the estimation of the eigenvectors and eigenvalues of the distance matrix) compared to the other MDS procedures.

In the proposed application, the FastMap allows us to find a mapping y^D with a linear complexity such that the distances between each pair of grey-level values associated to pixels s and t are close of $\beta_{s,t}$ as much as possible. To this end, we recall that the first step, and an essential element of the FastMap algorithm, is to select two objects (pixels in our case) to form the projection line. These two pixels, also called pair of anchor nodes or pivots (or pivot line) are selected such that the distance ($\beta_{s,t}$ in our application) is maximal. To accomplish such a task Falutsos and Lin proposed a linear heuristic algorithm, based on a deterministic procedure called "choose-distant-objects". The second step is to project any other object (pixels) onto this orthogonal axis (pivot line) by employing the cosine rule.

However, the price paid for the low linear complexity of the FastMap is its sensitivity to outliers and non-linearities. In our case, this characteristic may give a poor or noisy estimation of the similarity image y^D. In order to get a more reliable estimation, an interesting solution is obtained by averaging the estimations from different pivot lines. To this end the linear heuristic and deterministic procedure proposed by Falutsos and Lin can be easily modified in order to propose more than one pivot line.

2.4 Fusion-based Segmentation Step

Finally, in order to achieve more robustness, changes are then identified, from the (previously estimated) similarity image y^D, by combining the results of $T = 5$ different automatic thresholding algorithms[3] (namely [41, 91, 97, 121, 124]). In this way, this strategy (already been used in [62]) allows us to synergistically integrate multiple different criteria, for which these binary segmentation algorithms have been designed to be optimal in order to further increase the robustness and reliability of our proposed segmentation scheme. In our application, this binary fusion process is simply achieved by using a majority vote filter using a three dimensional window $W \times W \times T$ whose the first two dimensions are spatial and the third dimension indexes the different binary thresholded maps to be fused. In our application, this majority vote is achieved with a 3D window which is spatially centered on the pixel to be classified, and that collects the binary class labels of the different binary thresholded maps and finally by assigning to that central pixel, the class label that has the majority vote. This strategy ensures both the spatial regularization of the final fused (detection) map result and also a reliable decision fusion between results obtained by different thresholding strategies.

2.5 Experimental Results

2.5.1 Results on Multi-Modal Datasets

To evaluate the efficiency of our proposed model, we validate our approach on three real pairs of heterogeneous remote sensing images (see Fig. 2.2), provided by the

[3] Let us note that the concept of combining classifiers for the improvement of the performance of individual classifiers is known, in machine learning field, as a committee machine, ensemble classifiers, ensemble methods or mixture of experts [21, 99]. In this context, Dietterich [21] have provided an accessible and informal reasoning, from statistical, computational and representational viewpoints, of why ensembles can improve results.

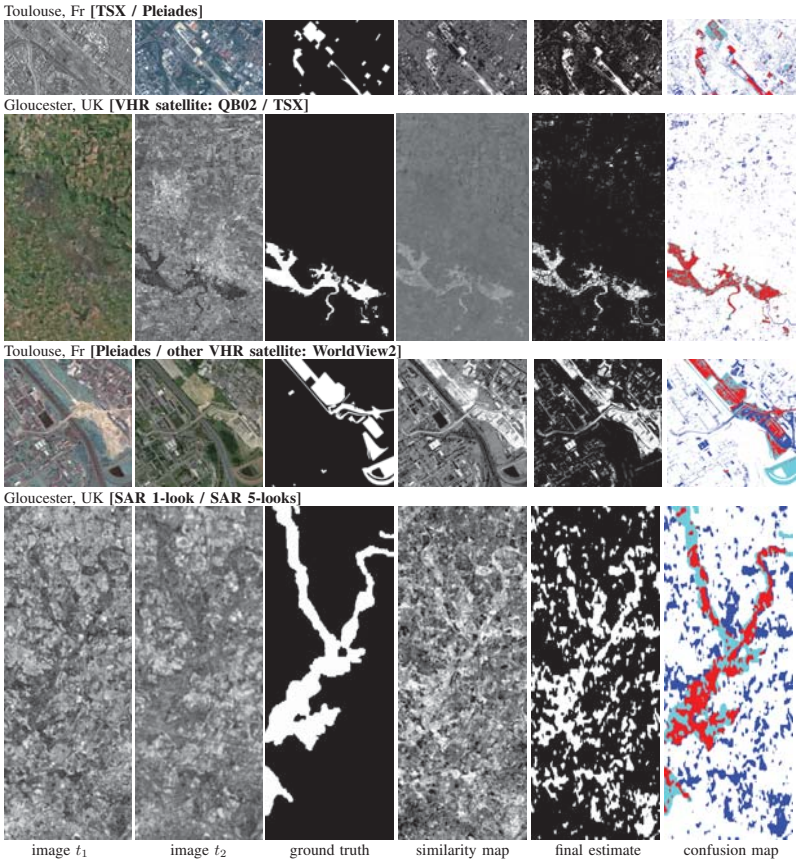

Figure 2.2. From left to right; image t_1 (before event), image t_2 (after event), ground truth, estimated similarity feature map $\hat{y}^D$, final binary map result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN). From top to bottom; *multi-modal* image pair: SAR/Optical (image from TerraSAR-X / Pleiades satellite of Toulouse, France), optical/SAR (image from QB02 / TerraSAR-X satellite of Gloucester, UK), heterogeneous optical/optical (image from Pleiades / WorldView2 of Toulouse, France), heterogeneous SAR/SAR (image from SAR 1-look / SAR 5-looks of Gloucester, UK).

CNES center (French National Centre for Space Studies), and already used in [88,89] and for which the different change mask constructions were provided by a photo interpreter. Besides, we have at our disposal one pair of heterogeneous SAR images given by [14]. This allows us to compare the performance of our model with the four existing state-of-the-art *multi-modal* change detection algorithms in this field,

namely the one introduced in [88,89] (and its improved version proposed in [90]), the multidimensional EM-based model proposed in [65] and the method proposed in [14] for heterogeneous multi-sensor SAR data. Besides, we have also compared our result with change detector traditionally used in *mono-modal* approaches provided by the ORFEO Toolbox [104].

• The first *multi-modal* dataset is a pair of SAR/optical satellite images (Toulouse, France), with size 4404×2604 pixels, before and after a construction. The SAR image was taken by the TerraSAR-X satellite (Feb. 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of CNES) satellite (July 2013). The TSX image was co-registered and re-sampled by [87] with a pixel resolution of 2 meter to match the optical image.

• The second one is a pair of optical/SAR satellite images (Gloucestershire region, in southwest England, near Gloucester), with size 2325×4135 pixels, before and after a flooding (on a mixture of urban and rural area). The optical image is a screenshot from Google Earth and comes from the Quick Bird 02 (QB02) VHR satellite (15 July 2006) and the SAR image was acquired by the TerraSAR-X satellite (July 2007). The TSX image presents a resolution of 7.3 meters and the QB02 image (with resolution of 0.65 meter and 0% cloud cover) was co-registered and re-sampled by [87] to match this resolution.

• The third dataset is a pair of different optical images, with different sensor specifications (*i.e.* spectral bands), with size 2000×2000 pixels (with the same resolution of 0.52 meter and 0% cloud cover), before and after the construction of a building (in the urban area of Toulouse, France). The first optical image is captured by the Pleiades sensor (May 2012) and the second optical image is a screenshot from Google Earth and is acquired by WorldView2 satellite from three (Red, Green and Blue) spectral bands (11 July 2013). The WorldView2 VHR-image was co-registered by [87] to match the Pleiades image.

• The fourth *multi-modal* dataset [14] is a pair of SAR/SAR satellite images

Table 2.2. Confusion matrix for the four *multi-modal* **datasets i.e., [TSX/Pleiades] (**4404×2604 **pixels), [QB02/TSX] (**2325×4135 **pixels), [Pleiades/WorldView2] (**2000×2000 **pixels), [SAR 1-look / SAR 5-looks] (**762×292 **pixels).**

Multi-modal pair	TP	TN	FP	FN
TSX / Pleiades	4.83%	81.96%	10.08%	3.13%
QB02 / TSX	4.36%	90.58%	3.00%	2.06%
Pleiades / WorldView2	9.08%	76.30%	8.54%	6.08%
SAR 1-look / SAR 5-looks	9.88%	68.22%	14.24%	7.66%

(Gloucester, U.K.) before and during a flood, with size 762×292 pixels, acquired by RADARSAT satellite. The numbers of looks for the before SAR image is 1-look image (Sept. 2000) and the numbers of looks for the after image is 5-looks (Oct. 2000). These two SAR images have a resolution of about 40 meters.

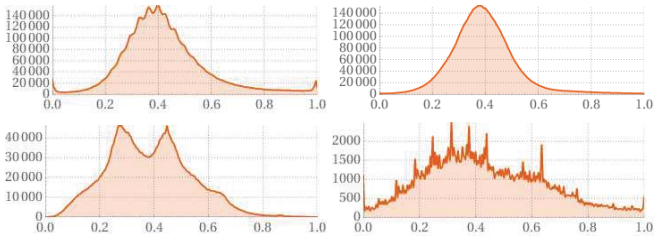

Figure 2.3. Histogram of the four similarity-feature maps of the four *multi-modal* **image pairs generated by the FastMap (see Fig. 2.2).**

We have considered the pairwise distance formula given by Eq. (2.2) where $y_s^{t_1}$ corresponds to the simple grey level of the image (and not a local statistics vector around a neighbourhood of s). In the case of an optical image, this also requires the conversion of the possible color image to a grayscale image. We have finally considered the final majority vote with a squared window spatial size set to $W = 3 \times 3$.

We have summarized respectively in Tables 2.1 and 2.2 the accuracy rates and the

confusion matrix obtained by our approach, compared with the four existing *multi-modal* change detection methods (see also Fig. 2.2) and some classical change detection methods borrowed from *mono-modal* techniques. We can notice that the proposed model outperforms quantitatively the four existing state-of-the-art approaches recently published in this field.

Fig. 2.4 shows the binary maps obtained by the Prewitt [91], Kapur [41], Zack [124], Yen [121] and Shanbhag [97] binarizers on the feature similarity map generated by the FastMap in the case of the second and fourth *multi-modal* dataset and the fusion results obtained by the proposed fusion strategy based on a three dimensional ($3\times3\times5$) majority vote filter. We can notice that the different binarizers estimate a different optimal threshold leading to a different binary map since different criteria are used. Nevertheless, the proposed fusion strategy ensures both an efficient spatial and consensus regularization, even if the statistical distribution of the feature similarity map is not clearly bi-modal (see Fig. 2.3).

We can notice that some histograms of the similarity map are not bi-modal. In our case, this is not a problem since four of the five binarizers, used in our procedure, does not necessarily assume that the histogram is bi-modal. For example, the so-called triangle method presented in [124] proposes to construct a line between the histogram peak and the farthest end of the histogram and the threshold is the point of maximum distance between the line and the histogram. Another binarization method which is applicable, even if the histogram is not bi-modal, is the binarizer proposed in [41] which uses the entropy concept. In this case, the threshold is estimated such that the entropies of distributions above and below are maximized. In the same spirit, [121] uses the maximum correlation criterion as a more computationally efficient alternative to entropy measures. Finally [97] proposes an extension of the method proposed in [41]. Only the binarizer proposed in [91] seeks two modes in the histogram and thus relies on the presence of a bi-modal shape of the histogram. The method

consists in iteratively smoothing the histogram (using a running average of size 3) until two peaks remain; the threshold is then the minimum or midpoint between the two peaks. Nevertheless, algorithmically, if a bi-modality in the histogram is not detected after a maximum number of iterations, the threshold is generally the grey value corresponding to the highest peak. All these different binarizers generally ensure the diversity which is then needed for a reliable subsequent fusion process.

Let us stress out that the proposed model can also be easily generalized in the case where more than one image, before and after a given event, is available. Indeed, this can be easily done by considering the following averaged pairwise distance:

$$\overline{\beta}_{s,t} = \frac{1}{N_p} \sum_{<b,a>} \beta_{s,t}(y^b, y^a) \tag{2.3}$$

with $\beta_{s,t}(y^a, y^b)$ the distance expressed in Eq. (2.2) for an image respectively belonging to the before (and after) event set y^b (y^a) and where the averaging is done over all possible pairs of images available before and after a given event ($< b, a >$) (and N_p is the number of averaging pairs). Let us note, however, that this technique can be applied only if the date of change event is known beforehand. This averaging procedure could even improve the estimation of $\beta_{s,t}$ since the averaging procedure is a reliable strategy to reduce the noise of any estimation procedure. In addition it would be interesting to study, in this multiple before and after image case, the effect of a median, harmonic or geometric mean operator instead of this arithmetic mean-based operator.

Fig. 2.5 shows a comparison of the similarity-feature maps obtained by Prendes *et al.*'s method [87] and the proposed method on the three first *multi-modal* dataset. By comparison with Prendes's method, the proposed CD method seems to visually produce more distinctly two clustering structures (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 to a more spatially and properly regularized (or less noisy) similarity-feature maps. It is interesting to note

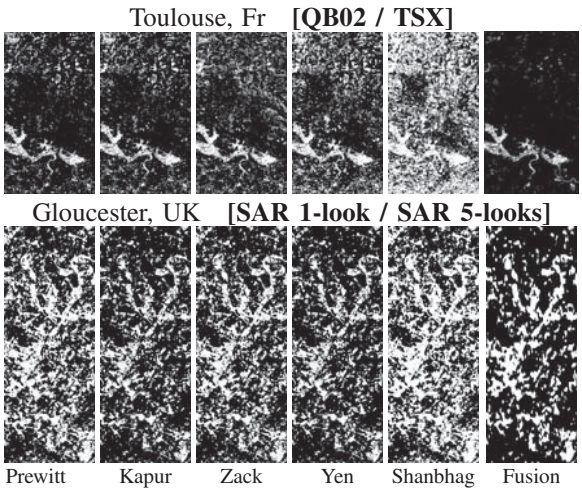

Figure 2.4. Individual binary CD maps given by respectively, the Prewitt [91], Kapur [41], Zack [124], Yen [121], Shanbhag [97] binarizers on the similarity-feature map generated by the FastMap (see Fig. 2.2) and fusion results using a majority vote filter using a three dimensional ($3\times\ 3\times\ 5$**) window.**

that our *multi-modal* CD strategy is able to detect very thin structure in the *changed area* class, such as the thin S-shaped region in the middle bottom of the middle image (contrary to the Prendes *et al.*'s method). We can also notice that some false positives are detected in the same locations in the two methods (see the rectangular shape at the upper-right of the bottom-left quadrant of the third image). Let us note that, in our case, the similarity-feature maps closely depends on the pairwise distance used (see Eq. (2.2)). A clever and more discriminative pairwise distance metric would allow us to obtain a better similarity-feature map. In addition, it is worth mentioning that the proposed method still remains perfectible if a better binarization strategy is found.

2.5.2 Results on Mono-Modal Datasets

In order to demonstrate that our approach is flexible enough to also be efficiently used in *mono-modal* change detection (*i.e.* with homogeneous sensor), we present a

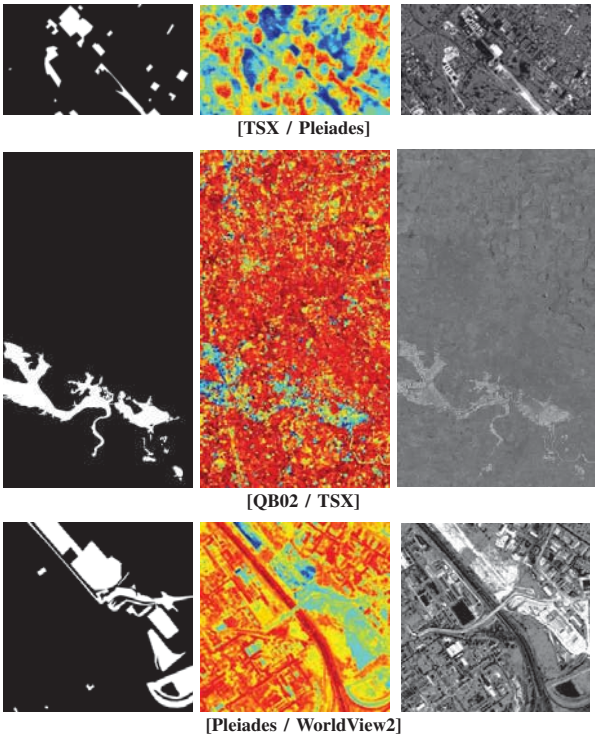

Figure 2.5. Comparison of the similarity-feature maps obtained by Prendes *et al.*'s method [87] and the proposed method on the three first *multi-modal* dataset. From lexicographic order, ground truths, similarity-feature map obtained by Prendes *et al.*'s method in false colors (red areas represent high similarity between the two images, while blue areas correspond to low similarity) and similarity-feature maps obtained by the FastMap-based proposed method for, from top to bottom, the TSX/Pleiades, QB02/TSX and Pleiades/WorldView2 datasets.

set of experimental results obtained on different real, publicly available, mono-modal optical, multi-temporal, multispectral, airborne SAR or radar data sets with available ground truth. In this case, we use the following and simple pairwise distance $\beta_{s,t}$:

$$\beta_{s,t} = \left| |y_s^{t_1} - y_t^{t_1}| - |y_s^{t_2} - y_t^{t_2}| \right| \tag{2.4}$$

which turned out a bit more efficient that the distance used in *multi-modal* case. In addition, for the *mono-modal* case, we have considered the final majority vote filter

Table 2.3. Accuracy rate of change detection obtained by different state of the art methods, on Bern (ERS-2), Ottawa (RADARSAT), and Beijing (Airborne SAR) datasets.

Method / Dataset	Proposed method	[32]	[56]	[48]	[59]	[31]	[119]
Bern (ERS-2)	.993	.996	.997	.996	.996	-	-
Ottawa (RADARSAT)	.943	.972	.965	.974	-	.988	-
Beijing (SAR Airborne)	.986	-	-	-	-	-	.997
Nb. of images tested	17	3	2	3	2	5	2

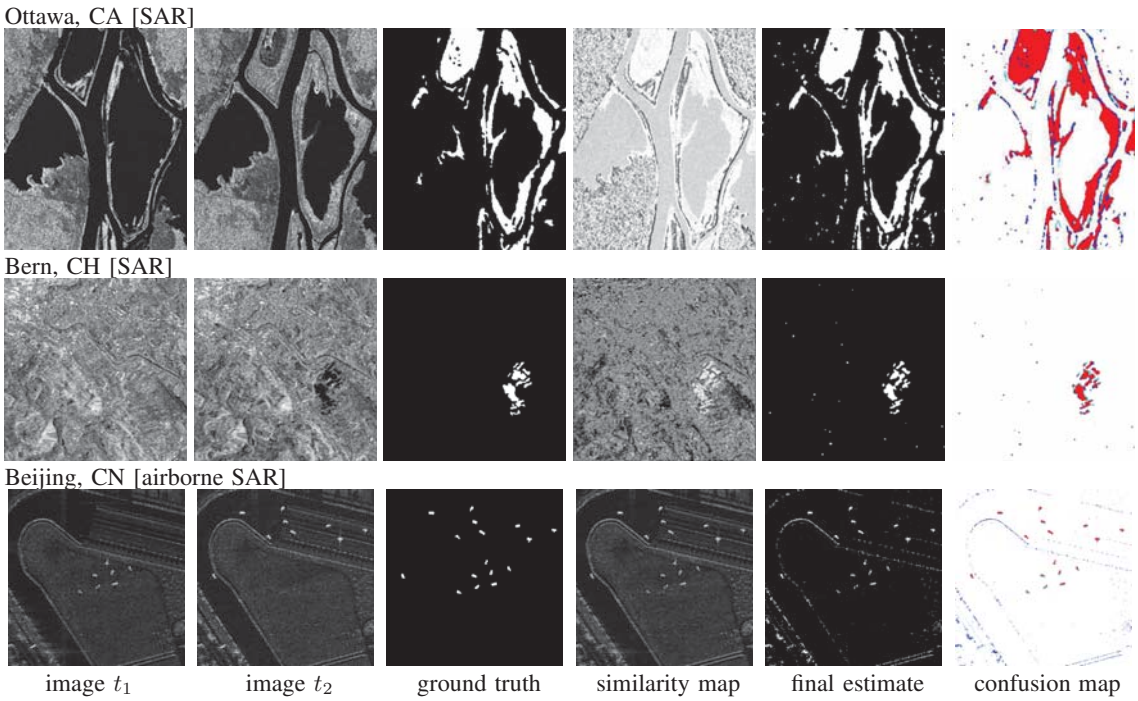

Figure 2.6. Experimental results on *mono-modal* SAR (second, third) and airborne SAR (fourth) dataset: Ottawa, Bern, Beijing. From left to right; image acquired at time t_1 and t_2, ground truth, similarity feature map, final (changed/unchanged) binary segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by our approach.

with a squared window size set to $W = 3 \times 3$.

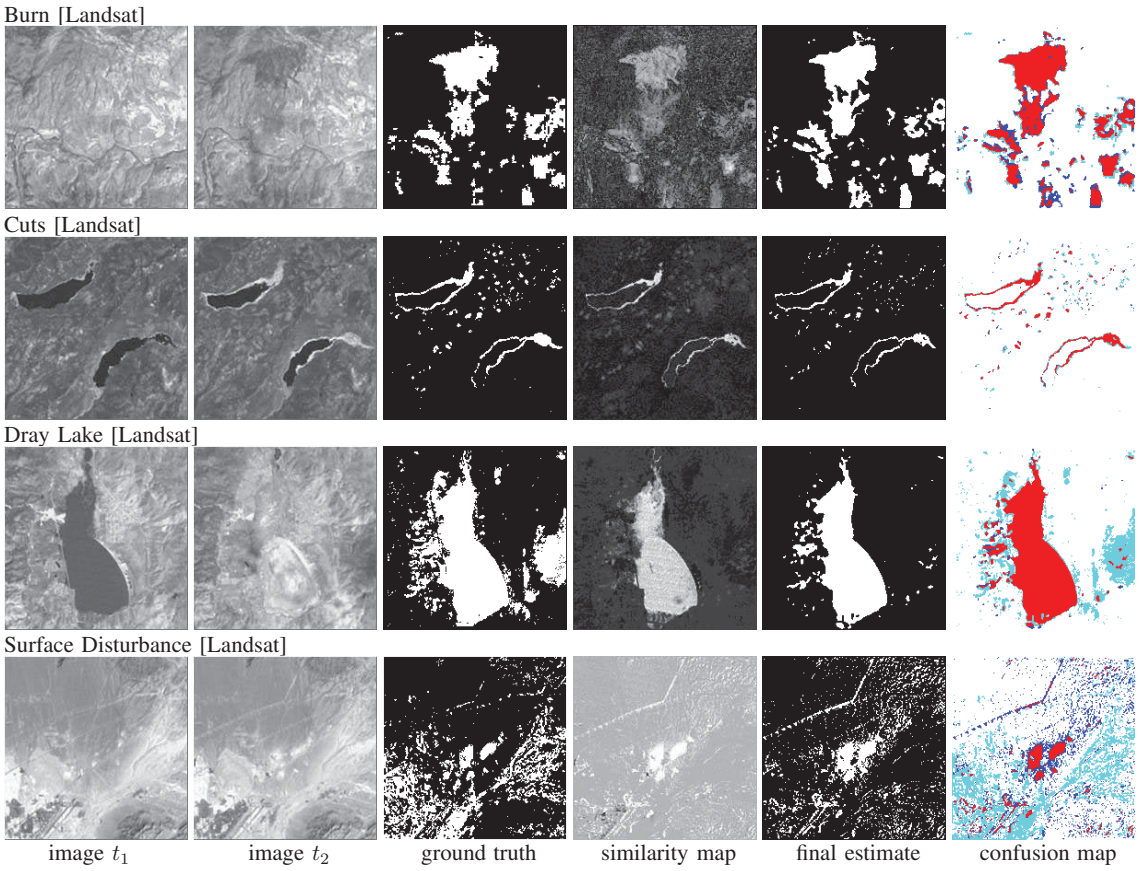

Figure 2.7. Experimental results on *mono-modal* optical Eros center (first) dataset: BURN, CUTS, DRAY LAKE, SURFACE DISTURBANCE; From left to right; image acquired at time t_1 and t_2, ground truth, similarity feature map, final (changed/unchanged) segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by our approach.

• The first dataset[4] (see Fig. 2.7) is a pair of optical satellite images produced by the EROS data center in southwest U.S., corresponding to a part of Reno-Lake Tahoe

[4] • The first mono-modal dataset: BURN, CUTS AND DRAY LAKE images and their ground truths have been downloaded from:
http://geochange.er.usgs.gov/sw/changes/natural/reno-tahoe/
SURFACE DISTURBANCE and its ground truth has been downloaded from: https://geochange.er.usgs.gov/sw/changes/anthropogenic/vegas/const.html

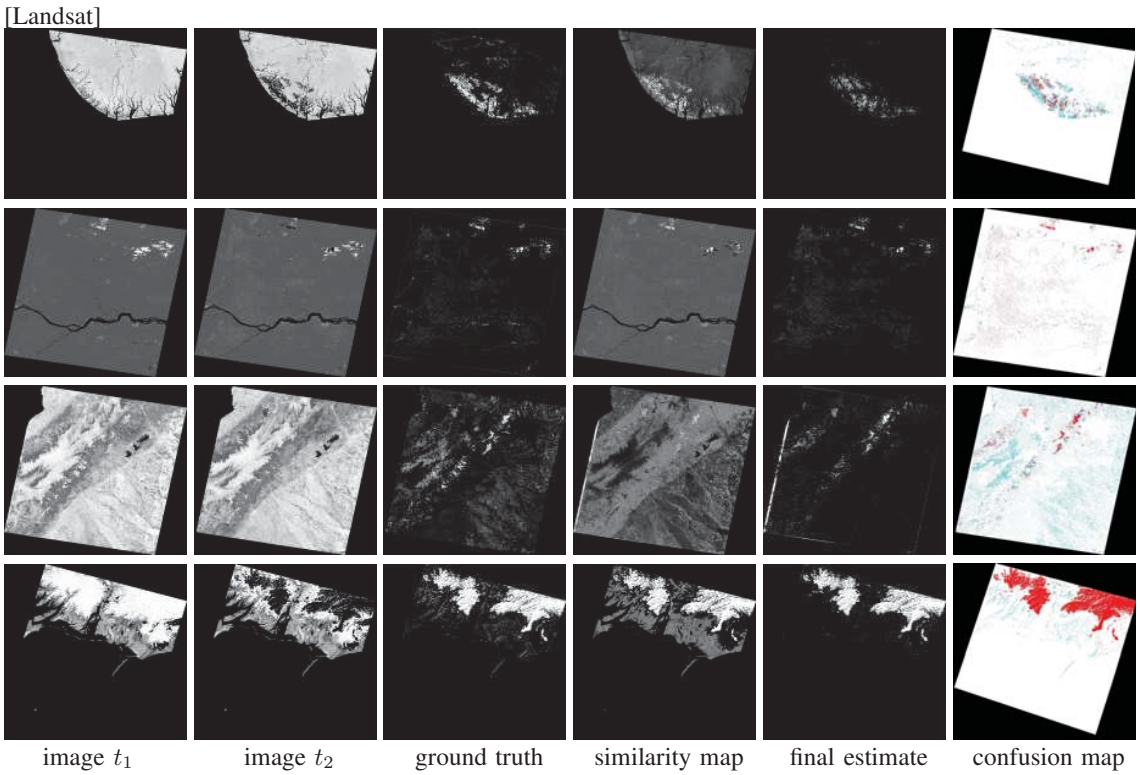

Figure 2.8. Experimental results on *mono-modal* **UMD-NASA (fifth) dataset; From left to right; image acquired at time t_1 and t_2, ground truth, similarity feature map, final (changed/unchanged) binary segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by our approach.**

area of Nevada (acquired on Aug. 5, 1986, and Aug. 5, 1992), with size 200×200 pixels, captured by the Landsat Multi-spectral Scanner. The Burn images show a change that results from forest fire phenomena. The Cuts images show a change

• Images of the second, third and fourth mono-modal dataset and their ground truths have been provided by Dr. Y. Li [49] and Dr. B. Xiong [119].

• Images of the fifth dataset have been provided by UMD-NASA and downloaded from:
http://glcf.umd.edu/data/landsatTreecover/
and their ground truths from: http://www.landcover.org/

described by a decrease in the surface area of the lake that results from drought effects. The Dray Lake images show a change that corresponds to the beginning and culmination of drought conditions in the western U.S. The Surface disturbance images show increased surface disturbance due to construction or excavations for construction including road resurfacing or paving.

• The second dataset [32] [56] [48] [31] (see Fig. 2.6) is provided by the Defence Research and Development Canada (DRDC), Ottawa (Canada), and are two multi-temporal SAR images relating to Ottawa, with size 290×350 pixels, acquired by the RADARSAT SAR sensor respectively, in July 1997 during the summer flooding, and Aug. 1997 after the summer flooding.

• The third dataset [32] [56] [48] [59] (see Fig. 2.6) is a pair of two multi-temporal SAR images with a size 301×301 pixels (pixel resolution is 12.5 meters), acquired by ERS-2 satellite (the European Remote Sensing satellite). It presents a natural phenomenon, generally occurring during the rainy season in the Switzerland area, near the city of Bern, in April 1999 before the flooding and in May 1999 after the flooding.

• The fourth dataset [119] (see Fig. 2.6) shows a pair of X-band airborne SAR (intensity) images with size 900×900 pixels (pixel resolution is 0.5 meter), acquired over a field in Beijing, China, on Apr. 4 and 6, 2004. It shows the number and positions of the vehicle on the field which were different during the two data acquisition dates.

• The fifth dataset (see Fig. 2.8) is a collection of images with size 7660×7402 pixels (pixel resolution is 30 meters), provided by the NASA/USGS Global Land Survey (GLS) [96], captured by the multispectral scanner Landsat-5 (TM) and Landsat-7 Enhanced Thematic Mapper Plus (ETM+) and showing various change phenomena in landscape, in different areas, between 2000 and 2005. For each pair of images of the same area, this dataset proposes a ground truth image containing the different evolutions undergone by the area for five years (thrust drills, loss of trees,

etc.).

Table 2.3 summarizes the different change detection accuracy rates obtained by our approach with a comparison with other *mono-modal* "state of the art" approaches [32] [56] [48] [59] [31] [119] for different datasets with different imaging modalities (with the total number of images tested in each case). We can see that the different changed-unchanged detection binary map results match fairly the different regions present in the ground truth, and that the most changed regions for the different imagery modalities are well recognized by our strategy (see Figs. 2.6, 2.7 and 2.8).

2.5.3 Shadow Effects

In this work (see footnote[1]), our goal is to detect changes of interest in the land cover or land use. Until now, we have respectively considered, in our *multi-modal* experiments, a major or a localized minor construction and two types of flooding (Fig. 2.2) and in the mono-modal case; a deforestation (due to a forest fire), two examples of decrease of a given lake's surface area (resulting from drought effects), a surface disturbance (*i.e.*, an excavations/construction for road paving) (Fig. 2.7), two different floodings, the detection of vehicles in an agricultural field (Fig. 2.6) and various change phenomena in the landscape such as thrust drills, loss of trees, changes in tree cover over time, etc. (Fig. 2.8).

As an additional experiment, it would be also interesting to see how the proposed CD model behaves when one of the two images has glow and shadow effects. To this end, for the homogeneous CD detection case, we have considered a stereo panchromatic data set provided by [105], with size 900×900 pixels (pixel resolution is 5 meter) and captured by Cartosat-1 satellite sensor. This pair of panchromatic images is acquired over Arges region (Romania near Piatra Craiului national park), on Oct. 2008 and Nov. 2009 and shows a forest changes caused by storms, and containing many shadow areas caused by steep terrain due to the mountainous forest area [105]. We have applied our CD model with and without any preprocessing step on the image

pair. As pre-processing, we use a simple (double) histogram matching method [98]. More precisely, the *before* image is histogram matched to the *after* image to give the pre-processed *before* image and the *after* image is then histogram matched to the latter (pre-processed *before*) image. We show in Fig. 4.5 the obtained results with a comparison in Table 2.4 with other state-of-the-art *mono-modal* change detectors studied in [105]. The result shows that our method is also robust in this *mono-modal* case. Nevertheless, it would also have been interesting to evaluate how our model behaves in the *multi-modal* case involving shadow effects, especially between SAR and optical images since the shadow is a quite different phenomenon between these two imagery modalities that cannot be corrected with a simple preprocessing scheme as a simple histogram matching method. This special case still remains to be studied.

2.5.4 Discussion

We can also notice that the rate accuracy of our method remains comparable, although slightly lower than the other *mono-modal* "state of the art" approaches but above all that the strength of the proposed model is its ability to process a wide variety of satellite imaging modalities (*i.e.*, multi-temporal, multispectral, airborne SAR or radar data) potentially degraded by different noise types and different noise levels (see, for example, the Fig. 2.6 where the SAR images are corrupted by different speckle noise levels). This peculiarity certainly comes from the fact that our model is, before all, designed to be used for the *multi-modal* change detection case. The average accuracy rate obtained by our change detection approach over 17 image pairs stemming from this five different *mono-modal* datasets with the distance expressed by Eq. (2.4) is $\rho = 0.94$ (94%). With the distance expressed by Eq. (2.2), especially well suited for the *multi-modal* change detection case, the average accuracy rate obtained on this five different *mono-modal* datasets is $\rho = 0.92$ (92%).

Consequently, we can say that the proposed method has also the defect of its main quality. Its ability to process a wide variety of imaging modalities (with different noise

types and levels) explains why it will be also less accurate than a specific mono-modal CD model only dealing with a specific type of noise and for which the similarity map, obtained by some local operations follows a particular mixture of distributions whose each distribution's shape may be theoretically estimated and for which the parameters of the finite distribution mixture can then be efficiently estimated with an EM like algorithm to finally obtain a reliable binary CD map.

We can also notice that the proposed model has, comparatively, more difficulties to separate the changed and unchanged areas when the SAR imaging modality is involved (see Fig. 2.2). This behavior can be probably explained by the inherent multiplicative speckle noise degrading the quality of any SAR images and creating, for each land cover class, a kind of macro-texture with grainy patterns (and referring to variations in radar brightness that are larger than many resolution cells). More precisely, this can be explained by the fact that the pairwise distances, used in our energy-based model (see Eqs. (2.1) and (2.2)), with y_v^t corresponding to the simple grey level at site v, can not fully model a coarse texture. In a multi-modality case involving SAR imaging, a more appropriate model would have been to consider local statistics around the pixel and therefore a distance computed between two feature vectors instead of two scalars. Nevertheless, experimentally, it would seem that a complex distance (*i.e.*, a more complex, realistic model) also leads to a harder optimization problem and finally a more approximated solution given by the Fastmap optimization procedure. In our case, a good solution of a simpler, approximate model seems preferable than an approximate solution of a complex (and maybe more realistic) model.

2.6 Conclusion

In this paper, we have proposed a new model for change detection in heterogeneous remote sensing images. Our method is mainly based on the estimation of a robust

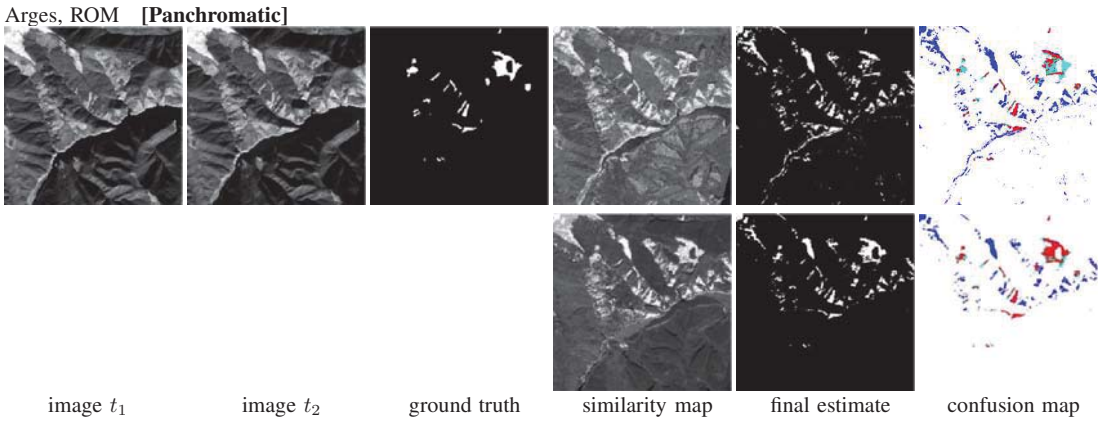

Figure 2.9. Panchromatic data set: image t_1, t_2, ground truth; similarity feature map; final (changed/unchanged) segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach. First row presents results obtained without any preprocessing step. Second row presents results obtained with a double histogram matching method based-preprocessing step.

Table 2.4. Kappa statistic of change detection on the Panchromatic shadow dataset obtained by the proposed method and other unsupervised (first upper part of the table) and supervised (second part of the table) state-of-the-art mono-modal change detectors [105].

Method	Kappa
Proposed method (with preprocessing)	**0.513**
Proposed method (without preprocessing)	**0.281**
kMNF OPTI‡ [105]	0.487 - **0.509** - 0.506 - 0.501 - 0.487 - 0.475
Height Difference‡ [105]	0.127 - 0.316 - 0.469 - **0.526** - 0.0 - 0.0
CVA‡ [105]	0.07 - 0.242 - 0.403 - **0.457** - 0.0 - 0.0
k-Means [105]	0.472
ICDA [105]	0.495
OSVM [105]	0.478
Random Forests [105]	0.432

‡based on different threshold levels given in increasing order

similarity feature map, containing the difference caused by the event between the bi-temporal multi-sensor images involved, and which is formulated as the solution of a set of constraints expressed for each pixel pair *via* a global cost function. A Fastmap

based optimization and then a simple fusion step, used to combine a set of binary segmentation maps generated by several automatic thresholding algorithms on this similarity feature map, then allows us to identify between the changed and unchanged areas. The proposed method is unsupervised and does not require a training data set or the estimation of an important parameter and can be used for any pairs of heterogeneous sensors. Besides, the proposed method is flexible since it can also be efficiently used in *mono-modal* change detection (*i.e.* with homogeneous sensor). It can be easily generalized in the case where more than one image, before and after a given event, is available or be used to handle separately the individual bands of a multi- or hyperspectral image (with d spectral bands), by simply formulating the constraint or difference between each pair of pixels, as the distance existing between two d-dimensional spectral vectors. Finally the model is perfectible by identifying a better pairwise distance or a better binarization strategy and its time complexity is linear with the total pixel number.

Acknowledgement

The authors are very grateful to acknowledge Dr. J. Prendes and CNES (French National Centre for Space) for sharing the *multi-modal* dataset in order to validate our model. The authors would also like to thank Dr. Y. Li [49], Dr. B. Xiong [119] and UMD-NASA for having provided us, respectively, the second and third, the fourth, and the fifth mono-modal dataset [96]. The authors would also like to express their gratitude to Dr. Jiaojiao Tian who put at our disposal the change detection Panchromatic shadow data set [105] and also for the time spent in providing the comparison results. We also thank all the other researchers who kindly made their databases available for our study and the comparisons made in this paper. The authors are finally grateful to the four anonymous reviewers for their numerous comments and suggestions that helped improve both the scientific content and the

presentation of this paper.

Chapitre 3

A RELIABLE MIXED-NORM BASED MULTIRESOLUTION CHANGE DETECTOR IN HETEROGENEOUS REMOTE SENSING IMAGES

Dans ce chapitre, nous présentons notre article accepté dans la revue *I*EEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, intitulé: **A Reliable Mixed-Norm based Multiresolution Change Detector in Heterogeneous Remote Sensing Images** . Nous exposons ce dernier dans sa langue originale de soumission.

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 paper, we present a novel and reliable unsupervised change detection 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 change detection estimation particularly robust in terms of sensitivity/specificity trade-off. Second, the detection process, embedded in a multi-resolution 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.

3.1 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) [55] 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 non-trivial 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 [23, 36, 77, 122] [33], namely; environmental monitoring, deforestation, urban planning, 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 [55, 88–90]. In fact, the practical and technical advantages of such multimodal analysis procedure are obvious both technically and practically [65, 88]. First, let's emphasize that a heterogeneous CD approach may be useful and sometimes

indispensable in some emergency cases. Since SAR sensors can operate regardless of weather conditions, even at night, *i.e.* with less restrictive conditions compared to optical imaging [88–90]. We can give the representative case of an optical image of a given area which is provided by an available remote sensing image archive data and only a new SAR image can be acquired for technical reasons, lack of time, availability or atmospheric conditions in an emergency situation for the same area [88–90]. A similar example can be given in 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 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 [14, 88–90, 109]. Let us also stress that, since multimodal CD must be adaptable to heterogeneous data with different statistics, this procedure may turn out also 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 mono-modal 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 [45]) for further improving the change detection 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 [65] [88] [14] [51] but generally we can divide them into three main categories: parametric, non-parametric or invariant similarity measure or operator-based models.

In parametric techniques, a mixture or a set of parametric multivariate distri-

butions are 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 [65] 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 Kullkack-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 [88–90] whose the first step aims at estimating a physical model, based on a mixture of multi-dimensional distributions (both taking into the noise model, the relationships between the sensor responses to the objects and their physical properties), with the expectation-maximization (EM) algorithm [18]. A statistical test based on this model then allows to estimate the changes. In the same spirit, the authors in [14] 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 [65, 88, 90] or SAR with different numbers of looks [14]) and consequently, they are not easily generalizable for another pair of different sensors. Besides, these method are in fact semi-supervised since they generally require (as training set) that two training images (sometimes manually selected and carefully chosen) associated with an unchanged area are available [65, 88, 90]. Let us finally add that these methods also require a Maximum Likelihood (ML) parameter estimation step of the distribution laws considered, which can be complex and computationally expensive.

Among nonparametric methods, an energy minimization model has been specifi-

cally designed in [109] 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 square 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 [117], 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.* in [113] find joint projections of the paired input images by maximizing the correlation between the projected data with a canonical correlation analysis. Another representation which 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.* in [29] propose a general multidimensional evidential reasoning (MDER) approach using the segmentation results of the pre and post-event satellite images with an extension of the Fuzzy C-means (FCM) clustering under belief function framework and whose result is directly used as basic belief assignment in their MDER approach. A similar strategy is also proposed in [15]. In the same vein, a pre-segmentation strategy based on the Normalized Difference Spectral Index is described in [120]. Let us note that machine learning-based methods are also non parametric (in the sense that they do not assume a specific parametric distribution for the data) and deep learning methods through conditional adversarial networks [66], convolutional coupling networks [51], or method based on deep feature representation [126], binary support vector classifier [11], multi-classifier systems [24] or based on simple K-nearest neighbors technique [53] have also been recently proposed and turn out to be valuable for the multimodal CD problem. In fact, these non-parametric 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 depend on the availability of an adequate massive amounts of representative training data.

Finally, in the third family of method, Alberga *et al.* [2] propose to use a technique closed to the co-registration 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 [8] is presented a CD method to quantify the damages caused by an earthquake to each individual building, using pre-event optical image and post-event SAR images. To this end, the parameters of each building, estimated from the optical scene and combined with the acquisition parameters of the actual post-event SAR scene are both 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 two first above-mentioned categories of models (parametric and non-parametric) 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 work, we propose a new imaging modality invariant change detector which belongs to the third family of above cited methods. Compared to our preliminary model [110], 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 term 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 are combined, thanks to the Multidimensional Scaling (MDS) mapping method, to generate a similarity feature map, which turns out 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 non-parametric 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 change detection problems in multimodal case; Namely, multisensor image pairs with i) heterogeneous optical images, multisource image pairs with ii) SAR+optical or optical+SAR images and finally, iii) multilooking SAR images.

The remainder of this paper is organized as follows: Section 7.2 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 7.3 presents a set of experimental results and comparisons with existing multimodal change detection algorithms. Finally, Section 7.4 concludes the paper.

3.2 Proposed Change Detection Model

The proposed model takes as input two bi-temporal 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 a four-step procedure:

• We first estimate a set of multiscale features aiming at detecting the structural difference in terms of high-frequency components of each local region (2D 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 that, consequently, our estimation model could estimate a degenerate overfit solution to this problem by classifying all pixels to be *unchanged* (see Sect. 3.2.1 and 3.2.2).

• In order to both reduce the noise and to 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 Multidimensional Scaling (MDS) mapping technique (see Sect. 3.2.3).

• To further reduce the noise of this similarity map, we then apply a spatially adaptive filters based on the super-pixel representation of the before and after satellite images (3.2.4).

• 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 Sect. 3.2.5). More precisely, the different steps of our approach are:

3.2.1 Imaging Modality Invariant Change Feature

Let us consider two (previously co-registered) bi-temporal 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 coregistred temporal image 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) [23, 36, 77, 122] [33]. 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 [110] particularly when the input images are acquired by different sensor types. Indeed, the gray or color value of each pixel is not a 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 that 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 an 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 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 [17] 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 NDVI (normalized difference vegetation index) [102] in multispectral imagery or the polarization ratio of SAR data [60] 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 can not be straightforward used and exploited in a multimodal change detection system except in the specific multispectral/optical

[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 band-pass capabilities, the human visual system (HVS) can recognize, even in a complex SAR image with strong correlated speckle, the specific high-frequency spatial textural pattern created by an urban area.

case introduced in [53] in which a NDVI image, combined with an optical (SPOT) image, are both projected in a common feature space for the convenience of change detection.

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:

$$z_s^{D_1} = \sum_{<s,s'>\in W_n} \left| |\mathbf{y}_s^{t_1} - \mathbf{y}_{s'}^{t_1}|_1 - |\mathbf{y}_s^{t_2} - \mathbf{y}_{s'}^{t_2}|_1 \right| \tag{3.1}$$

where the summation is done over all pairs of pixels at location $< s, s' >$ contained in a $N_n \times N_n$ squared window W_n including the central pixel located at site s. This summation allows us both 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 location $< s, s' > \in W_n, s \neq s'$) obtained by gathering together all the gray (or color) values contained in a $N_{s'} \times N_{s'}$ 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 differences between these two (feature) vectors are computed in the L_1 norm sense ($|.|_1$).

A simple way to improve our CD result accuracy consists in considering and estimating the dual and complementary version of the above expressed (in Eq. 3.1) similarity map by considering the same local spatio-temporal gradient operator but

expressed in terms of the infinity norm (which is the dual norm of the L_1 norm [93] [123]). In this regard, a second similarity map z^{D_2} is estimated, at every pixel s of the image, by the following operator:

$$z_s^{D_2}=\sum_{<s,s'>\in W_n} \max_{1\leq i\leq N_{s'}\times N_{s'}} \left| |y_i^{t_1}(s)-y_i^{t_1}(s')| - |y_i^{t_2}(s)-y_i^{t_2}(s')| \right| \tag{3.2}$$

with $y_i^{t_1}(s)$ is the i-th component of the pixel vector $\mathbf{y}_s^{t_1}$ or the i-th pixel value taken within the $N_s \nsim N_{s'}$ 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 been already investigated and observed recently in machine learning theory for improving feature selection techniques or for finding a support vector machine based classification rule with minimal generalization error [37] but also 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 [28], denoising [6], image deconvolution [82], or in Fluorescence diffuse optical tomographic (FDOT) reconstruction [5], 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 encourages sparsity contrary to $L_{p>2}$ (and especially L_∞ norm) that favours diversity [44]. This is what we have observed in our multimodal change detection or two-class segmentation problem; the spatio-temporal gradient operator based on the L_1 norm favours sparse segmentation result contrary to the one based on the L_∞ norm which rather encourages diversity.

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 the

take 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 favour of one class.

3.2.2 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 two in the horizontal and vertical directions), has the intrinsic capability to represent and re-organize 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 both makes our change detector robust against to noise and somewhat scale-invariant.

To this end, we construct two 3-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 Eqs (3.1)-(3.2)) 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 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 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)

3.2.3 Similarity Feature Map Estimation

Finally in order to further reduce both the noise of the estimation and also 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 3.2.5), 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 (1D) with the linear-complexity version of the Multidimensional Scaling (MDS) mapping method, called the FastMap technique[2] [26]. 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

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

Algorithme 1: FastMap

[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 [26]. 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

3.2.4 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 3.2.5) 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 estimate 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) [1] 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 superpixels by grouping pixels with a local k-means clustering method and second, exploit a connected components algorithm to remove the generated small isolated regions by merging them into the nearest large superpixels.

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

Algorithme 2: SLIC segmentation

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 over-segmented map y^S (with thus smaller superpixels) in which the set of pixels inside each new superpixel has the appealing property to both 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 classified 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 work that consists in averaging each pixel value of y^D, inside each superpixel of y^S, between them. Conceptually, this later 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. 3.1 and Algorithm 3 which simply averages out each y^D values of each segment).

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**

    $val \leftarrow val + y^D_s$

    $nb \leftarrow nb + 1$

  end

  foreach *pixel* p_s *(at location* s*)* $\in b_i$ **do**

    $\overline{y}^D_s \leftarrow (val/nb)$

  end

end

Algorithme 3: Superpixel-based filter

[3] if x^{t1} denotes the segmentation or the subdivision of the image y^{t1} into a set of superpixels or regions: $x^{t1} = \{R^{t1}_1, R^{t1}_2, \ldots, R^{t1}_{N_1}\}$ and x^{t2} is the subdivision of y^{t2}, *i.e.*, $x^{t2} = \{R^{t2}_1, R^{t2}_2, \ldots, R^{t2}_{N_2}\}$. Every pixel of the image pair (y^{t1}, y^{t2}) is thus associated to an unique region in the set x^{t1} and an 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}.

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 $\left(\mathbf{x_1}, ..., \mathbf{x_N}\right)$

Algorithme 4: Two-class clustering

3.2.5 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 which 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, as well as 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 [54] [4]. 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. 3.2 and Algorithm 4).

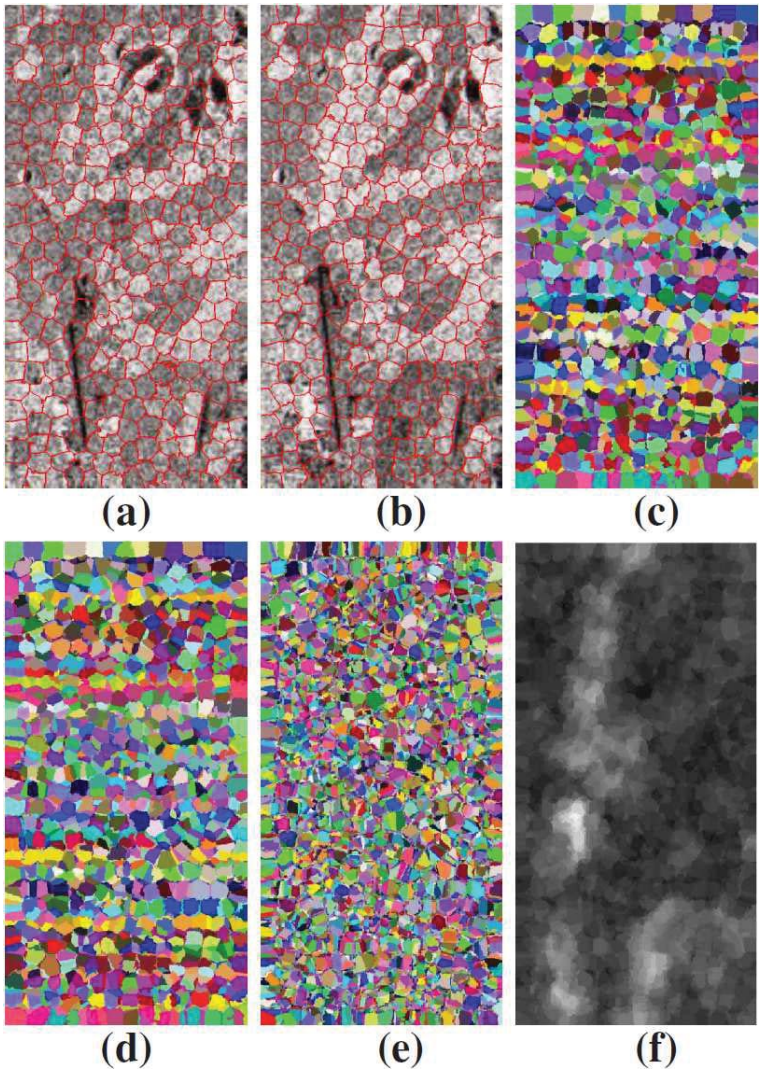

Figure 3.1. Superpixel based filtering step on SAR 3-looks/SAR 5-looks dataset (sixth dataset, cf. 7.3.1). (a-b) superpixel contour superimposed on y^{t_1} (before) and y^{t_2} (after) satellite image, (c-d) segmentation into superpixel regions (on images y^{t_1} and y^{t_2}), (e) segmentation intersection 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)).

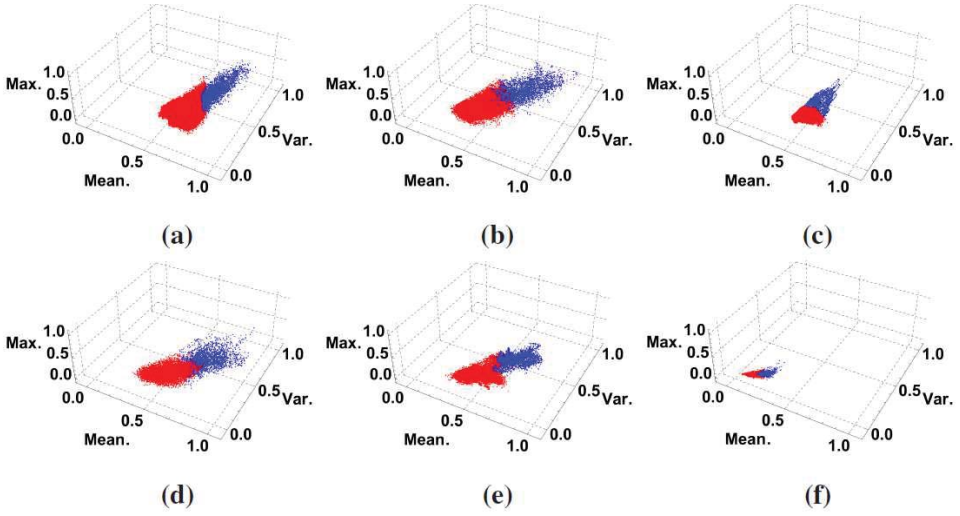

Figure 3.2. 3D feature space for the local textural features (mean, variance, maximum gray level) of the filtered feature similarity map y^D related to different heterogeneous datasets (a-f); red and blue colors represent, respectively, the unchanged and changed clusters or classes found by the K-means algorithm.

3.3 Experimental Results

To validate our approach, we present in this section a series of tests conducted on different real heterogeneous datasets, chosen to reflect the three possible change detection conditions in multimodal case; Namely, two heterogeneous optical images, heterogeneous SAR images, one optical and one SAR images. This allows us to compare the performance of the proposed method with different state-of-the-art multimodal change detection algorithms recently proposed in this field [65] [88] [90] [14] [89] [64] in different multimodal CD conditions. In this benchmark, all the ground-truth images (or change detection mask) was provided by an expert photo interpreter. We also compare the obtained results with other change detector traditionally proposed in mono or multimodal cases and provided by the ORFEO Toolbox [104] [63]. In our implementation, we have used the FastMap and SLIC C++ codes kindly provided by their authors and freely available on the web.

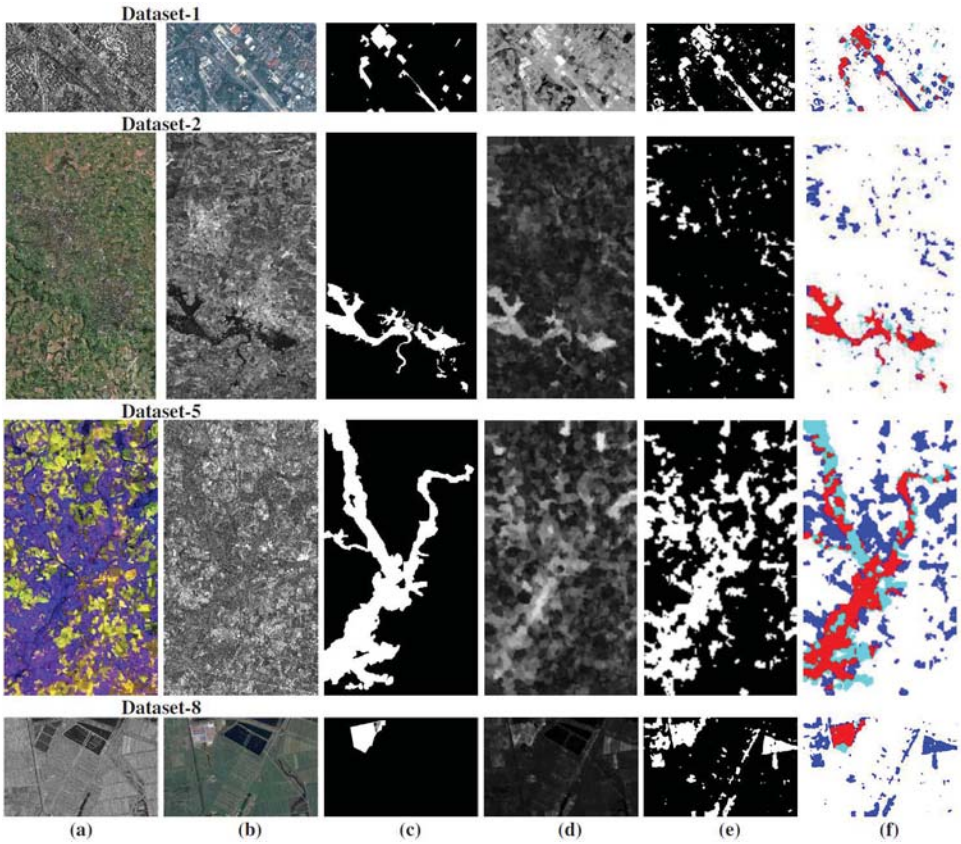

Figure 3.3. Heterogeneous (multisource) Optical/SAR and SAR/Optical datasets: (a-c) image t_1, t_2, ground truth; (d-f) filtered similarity map; final (changed-unchanged) segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

3.3.1 Heterogeneous Dataset Description

• The first heterogeneous dataset is a pair of SAR/optical satellite images (Toulouse, France), with size 4404 × 2604 pixels, before and after a construction. The SAR image was taken by the TerraSAR-X satellite (Feb. 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of CNES, Centre National d'Etudes Spatiales) satellite (July 2013). The TSX image was co-registered and re-sampled by [87] with a pixel resolution of 2 meter to match the optical image.

• The second one is a pair of optical/SAR satellite images (Gloucestershire region, in southwest England, near Gloucester), with size 2325×4135 pixels, before and after a flooding taking place in an urban and rural area. The optical image comes from the Quick Bird 02 (QB02) VHR satellite (15 July 2006) and the SAR image was acquired by the TerraSAR-X satellite (July 2007). The TSX image presents a resolution of 7.3 meters and the QB02 image (with resolution of 0.65 meter and 0% cloud cover) was co-registered and re-sampled by [87] to match this resolution.

• The third dataset shows two Heterogeneous optical images acquired in Toulouse (Fr) area by different sensor specifications (size 2000×2000 pixels with a resolution of 0.5 meter). 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 (11 July 2013) after the construction of a building. The WorldView2 VHR-image was co-registered by [87] to match the Pleiades image.

• The fourth dataset [14] is a pair of SAR/SAR satellite images (Gloucester, U.K.) before and during a flood event caused by intense and prolonged rainfall, overwhelming the drainage capacity, on a urban and agricultural/rural areas, with size 762×292 pixels, acquired by RADARSAT satellite with different number of looks. The numbers of looks for the before SAR image is 1-look image (Sept. 2000) and the numbers of looks for the after image is 5-looks (Oct. 2000). These two SAR images have a resolution of about 40 meters.

• The fifth dataset [63, 64] 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 on Sept. 1999 before a flooding event. The SAR image is captured by the European Remote Sensing (ERS) satellite (around Nov. 2000) during the flooding event. The resolution of these two images are about 10 meters [63].

• The sixth dataset [14] shows a pair of heterogeneous satellite images (size $400 \times$

800 pixels and resolution of 10 meters) 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 respectively 3-looks and 5-looks.

• The seventh dataset is composed of two heterogeneous optical images. 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 (Sept. 1995) with spatial resolution of 30 meters. The second optical image comes from Google Earth (Jul. 1996), and is an RGB image with spatial resolution of 4 meters. After co-registration, these two images are re-sampled at the same pixel-resolution 412×300 pixels.

• The eighth data set 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 (Jun. 2008) with spatial resolution of 8 meters. The optical image comes from Google Earth image (Sept. 2012) and its a combination of aerial photography imaging with a satellite imaging (produced respectively by QuickBird and Landsat-7) with a spatial resolution of 4 meters. After co-registration, these two images are re-sampled at the same pixel-resolution 921×593 pixels.

Table 3.1 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.

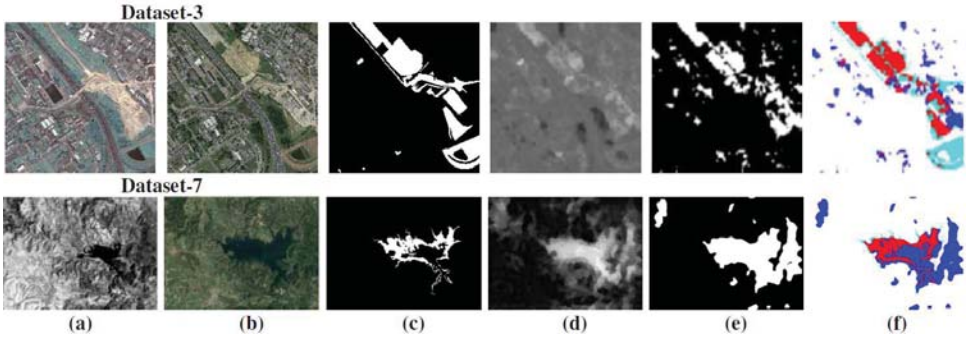

Figure 3.4. Heterogeneous (multisensor) Optical/Optical dataset: (a-c) image t_1, t_2, ground truth; (d-f) filtered similarity map; final (changed/unchanged) segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

Table 3.1. 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

3.3.2 Results & Evaluation

In all the experimental results, we have considered the simple gray level of the image (and not a local statistics vector around a neighbourhood of s) (see Eqs (3.1) and (3.2)). 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) is re-scaled for all sites s of the image between $[0-255]$. We have considered $N_p = 3$ levels of the multiresolution pyramidal structure and $N_n = 7$, $N_{s'} = 3$ for each operator applied at each scale of this pyramid (see Section 3.2.2). Finally, for the superpixel-

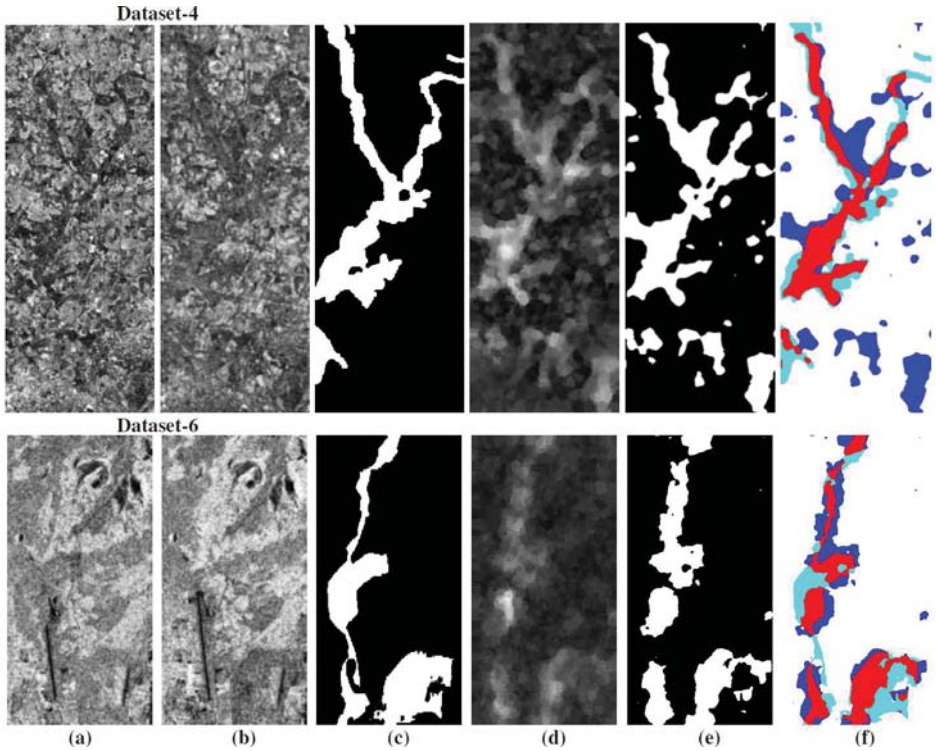

Figure 3.5. Heterogeneous (multilooking) SAR/SAR datasets: (a-c) image t_1, t_2, ground truth; (d-f) filtered similarity map; final (changed/unchanged) segmentation result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

based filtering step (see Section 3.2.4), 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 measure the percentage of the correct changed and unchanged pixels:

$$\text{PCC} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \tag{3.3}$$

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 belong-

Table 3.2. Accuracy rate of change detection 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 mono-modal change detectors (second lower part of each Table).

SAR/Optical dataset (1)	**Accuracy**
Proposed method	**0.881**
Prendes *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

Optical/SAR dataset (2)	**Accuracy**
Proposed method	**0.943**
Prendes *et al.* [87, 90]	0.918
Prendes *et al.* [88]	0.854
Copulas [65, 88]	0.760
Correlation [65, 88]	0.688
Mutual Inf. [65, 88]	0.768
Pixel Dif. [88, 104]	0.782
Pixel Ratio [88, 104]	0.813

Optical/Optical dataset (3)	**Accuracy**
Proposed method	**0.877**
Prendes *et al.* [87, 89]	0.844
Correlation [87, 89]	0.679
Mutual Inf. [87, 89]	0.759
Pixel Dif. [87, 104]	0.708
Pixel Ratio [87, 104]	0.661

SAR 1-look / SAR 5-looks dataset (4)	**Accuracy**
Proposed method	**0.821**
Chatelain *et al.* [14]	0.732
Correlation [14]	0.521
Ratio edge [14]	0.382

VHR Optical/SAR dataset (5)	**Accuracy**
Proposed method	**0.743**
Gregoire *et al.* [64]	0.70

SAR 3-looks/SAR 5-looks dataset (6)	**Accuracy**
Proposed method	**0.840**
Chatelain *et al.* [14]	0.749
Correlation [14]	0.713
Ratio edge [14]	0.737

Optical(NIR band)/Optical dataset (7)	Accuracy
Proposed method	**0.847**
Zhang *et al.* [126]	0.975
PCC [126]	0.882

SAR/Optical dataset (8)	Accuracy
Proposed method	**0.884**
Liu *et al.* [51]	0.976
PCC [51]	0.821

ing 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 corresponding to the unchanged pixels wrongly classified as changed.

A comparison with different state of the art approaches [65] [88] [90] [14] [89] [64] is summarized in Table 7.2. We have also summarized in Table 7.3 the confusion matrix obtained by the proposed change detector. From Table 7.2, we can see that the rate

Table 3.3. 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 1-look / SAR 5-looks] (762×292 pixels), [Spot VHR / ERS] (1318×2359 pixels), [SAR 3-looks / SAR 5-looks] (400×800 pixels), [Optical (NIR band) / Optical] (412×300 pixels), [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

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 Figures 6.4, 6.6, and 3.5) but also multitemporal image pairs with different spatial resolutions (see Tables 7.2 and 3.1). Nevertheless, we assert with high confidence that better accuracy results are obtained on satellite image pairs with high spatial resolution (datasets 1-2 & 3 versus datasets 4-5-6). 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 3.2.1), that tries to detect the presence or not of a common and specific high frequency pattern (e.g., edges, contours, micro-texture, 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 all the 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 figures 6.4, 6.6, and 3.5 illustrating the applicability and the efficiency of our detector for a wide variety of cases). Nevertheless, our average 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 learning machine based methods, on the content of a training base that could be biased in favour 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 favours sparse segmentation in term of candidate CD regions and used alone would increase the false negative rate (see Fig.3.6 (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.3.6 (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 operator that would favour one of the two classes. The evolution of the average classification accuracy according to the number of (pyramid) levels shows that 3 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.3.7 (b)). Finally, the number of superpixels affects slightly 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.3.7 (a)).

By knowing this, a further improvement of our method would be to include a reliable high frequency noise reduction step of the two input images, as very first pre-processing. 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.3.8 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 [65] [14] [89] [64], 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 to a 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 the formula 1 and 2 the average accuracy rate obtained on this eight multimodal datasets is respectively 73.81% and 64.35%. Fig. 3.6 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 also Fig.6.4 at second row.)

3.3.3 Parameter Sensitivity

In this section we study the impact of the four internal parameters of our CD model (see Section 6.3.2) on the final detection result. These parameters include the following: 1-] the two sizes N_n, $N_{s'}$ of the squared window used in our detection features (see Section 3.2.1) 2-] N_p the number of levels used in the pyramid representation (see Section 3.2.2) and finally, 3-] N_s the number of superpixels used in the SLIC segmentation algorithm (see Section 3.2.4). To that end, we have shown, in Fig. 3.7, the evolution of the average accuracy obtained (on the eight considered heterogeneous image pairs) when the N_p and N_s parameters evolve around their set value (the other parameters being constant and set to their set value, *i.e.,* $N_{s'}=3, N_n=7, N_p=3, N_s=300$). Concerning, the parameters $N_{s'}$ and N_n, since N_n is necessarily greater than $N_{s'}$ (the local $N_{s'} \times N_{s'}$ window describing the texture features around the pixel is necessarily included into a larger $N_n \times N_n$ *search window*), we obtain the following accuracy results;

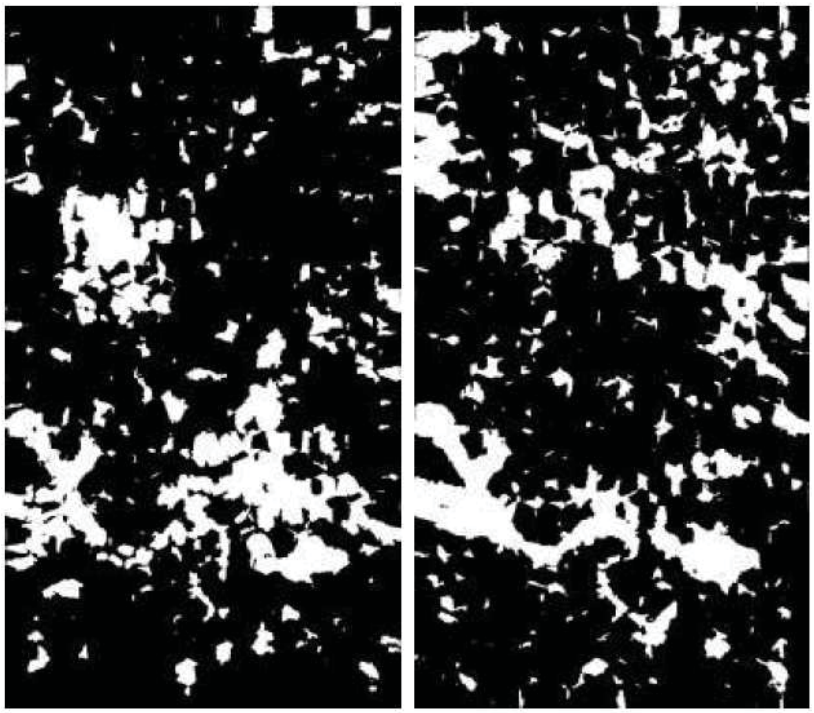

Figure 3.6. The two complementary binary maps resulting from the application of only the first (a) ($z_s^{D_1}$) and second (b) ($z_s^{D_2}$) CD operators on the second (optical/SAR) pair of satellite images.

0.8538, 0.8463, 0.8452, 0.8335 for respectively $(N_{s'}, N_n) = \{(3,7)(5,9)(7,11)(9,13)\}$. All these experiments show that the proposed model is not too much sensitive or dependent of one or its four internal parameters.

3.4 Conclusion

In this paper, we have proposed a novel and simple change detection method in heterogeneous remote sensing images. The proposed method is based on an imaging modality-invariant operator that detects at different resolution levels, the common specific high-frequency pattern of each structural region existing in the two hetero-

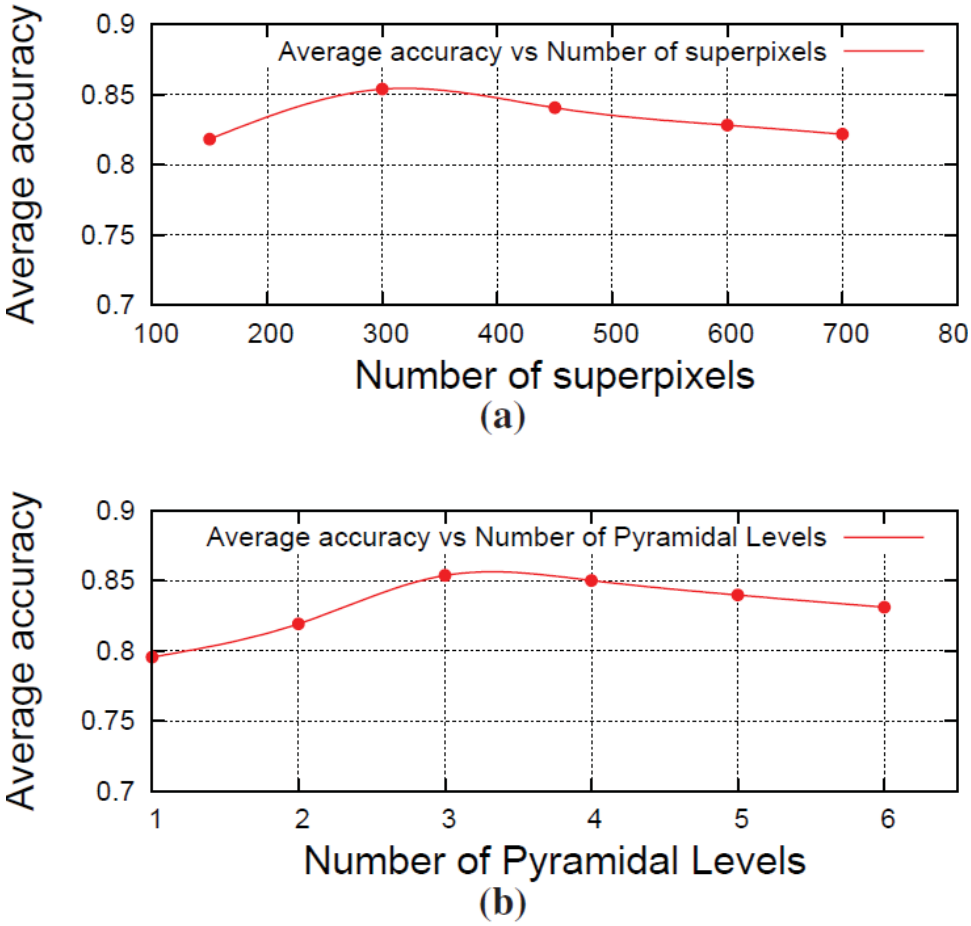

Figure 3.7. Evolution of the average accuracy using different parameters (the other parameters being constant and set to their set value): (a) number of superpixels, (b) number of pyramid levels.

geneous satellite images. The dual norm formulation of this new detector was found to give a reliable CD estimation whose distribution (given by the confusion matrix) is well balanced with no bias in favour of a particular class. Qualitative and quantitative results show that the proposed method is effective and performs particularly well, without requiring any preprocessing, on different types of input satellite images (multisensor optical images, multisource SAR+optical or multilooking SAR images), degraded with possibly different types of noise or different level of noise, and showing different kind of changes due to a major urban construction and/or changes due to different types of natural phenomenon.

Acknowledgement

We would like to acknowledge the Computer Research Institute of Montreal **(CRIM)** and the Ministry of Economic Science and Innovation **(MESI)** of the Government of Québec to have supported this work. We would like also to acknowledge all other researchers that made at our disposal the change detection dataset in order to validate the proposed change detector.

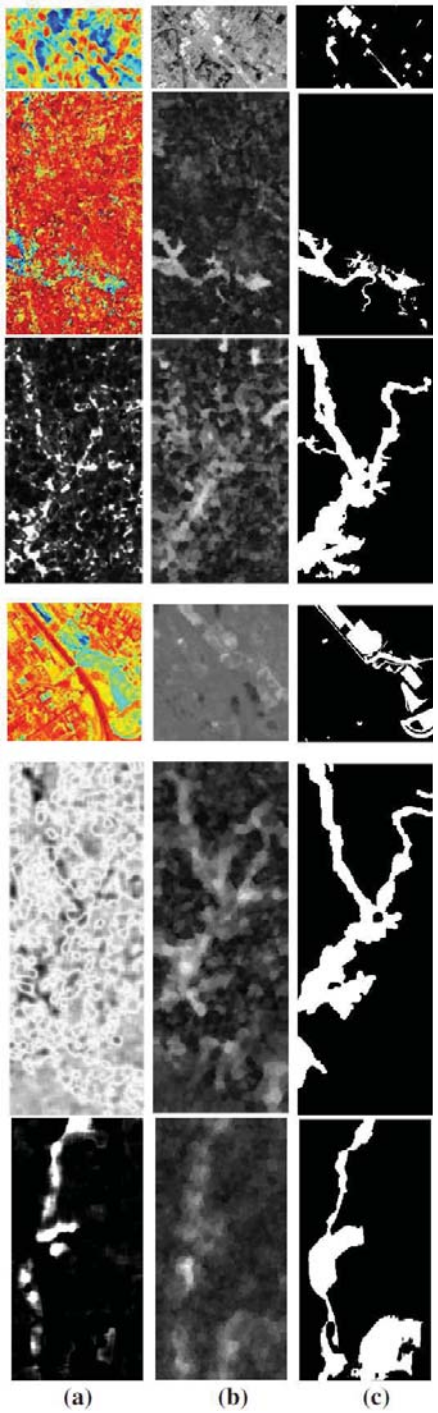

Figure 3.8. Comparison of the similarity map obtained by: (a) the SoA method presented by Prendes *et al.* (row 1, 2 & 4) [88] by Gregoire *et al.* [64] (row 3) and Chatelain *et al.* [14] (row 5 & 6) and (b) the proposed method in the case of, from top to bottom: multisource Optical/SAR and SAR/Optical images (datasets #1, #2 & #5), multisensor Optical/Optical images (dataset #3) and SAR/SAR (datasets #4 & #6) (c) ground truth.

Chapitre 4

MULTIMODAL CHANGE DETECTION IN REMOTE SENSING IMAGES USING AN UNSUPERVISED PIXEL PAIRWISE BASED MARKOV RANDOM FIELD MODEL

Dans ce chapitre, nous présentons notre article accepté dans la revue *I*EEE Transactions on Image Processing, intitulé: **Multimodal Change Detection in Remote Sensing Images Using an Unsupervised Pixel Pairwise Based Markov Random Field Model** . Nous exposons ce dernier dans sa langue originale de soumission.

Abstract

This work presents a Bayesian statistical approach to the multimodal change detection (CD) problem in remote sensing imagery. More precisely, we formulate the multimodal CD problem in the unsupervised Markovian framework. The main novelty of the proposed Markovian model lies in the use of an observation field built up from a pixel pairwise modeling and on the bitemporal heterogeneous satellite image pair. Such modeling allows us to rely instead on a robust visual cue, with the appealing property of being quasi-invariant to the imaging (multi-) modality. To use this observation cue as part of a stochastic likelihood model, we first rely on a preliminary iterative estimation technique that takes into account the variety of the laws in the distribution mixture and estimates the parameters of the Markovian mixture model. Once this estimation step is completed, the Maximum a posteriori (MAP) solution of the change detection map, based on the previously estimated parameters, is then computed with a stochastic optimization process. Experimental results and

comparisons involving a mixture of different types of imaging modalities confirm the robustness of the proposed approach.

4.1 Introduction

Multimodal Change Detection (CD) [55] is a procedure used to identify any land cover changes that occurred between two satellite images acquired at different times, in the same geographical area but by different kinds of sensors. Multimodal CD is a growing interest task which can be considered as a generalization of the basic and classic monomodal CD problem as it requires less stringent requirements about the characteristics and origin of the acquired data. It is also a challenging task since, such a procedure must be powerful and flexible enough to model any existing heterogeneous data types (thus sharing different statistics) in remote sensing imagery and to handle the same problems that have been already solved by monomodal CD techniques [25, 95, 128] [33] such as anomaly and target detection (eventually in the presence of diurnal and seasonal variations), natural, land or environmental monitoring, damage monitoring (earthquake, flooding, landslides, etc.) or urban planning, to name a few.

Multimodal CD has recently aroused a growing interest, in the remote sensing community since this technique allows to relax the assumption of homogeneous and co-calibrated measurements and consequently to exploit the huge amount of heterogeneous data, we can now get from various archives or from different types of existing Earth observing satellites. In addition, the practical and technical advantages of such multimodal analysis procedure are obvious and are widely described in the literature, for instance [109]. Finally, let us add that the different imaging modalities may be complementary and this sensor fusion technique could potentially be exploited (not only in Geoscience imaging [45]) for further improving the change detection and analysis of land surfaces with complex properties subject to extreme conditions (e.g. temperature, fire, ice, etc.).

Despite its undeniable potential, there are relatively few research works that have been devoted to heterogeneous or multimodal CD using machine learning or image processing. Nevertheless, we can identify four main categories. First non-parametric based techniques such as learning machine algorithms (since these techniques do not assume explicitly a specific parametric distribution for the data) [11,24,51,53,66,126] or unsupervised non-parametric based procedures, that do not require supervised training step, such as the energy based model, in the least-squares, sense proposed in [109] and satisfying an overdetermined set of constraints, expressed for each pair of pixels existing in the before-and-after images. Secondly, algorithms relying on similarity measures with invariance properties according to the imaging modality [2,8,110]. Thirdly procedures mainly based on a transformation or projection of the two multimodal images to a common feature space, in which the two heterogeneous images share the same statistical properties and on which classical monomodal CD methods can then be applied [15,29,58,111,113,117,120]. Finally parametric models that we now describe in more details since the proposed model fits into this category. In parametric techniques, a set (or mixture) of multivariate distributions are generally used to model the joint statistics or the dependencies between the two imaging modalities. More precisely, local models of dependence between unchanged areas are modeled according to the copula theory in [65] and based on these models, Kullkack-Leibler-based comparisons on local statistical measures are then used to generate a similarity map which is subsequently binarized. An appealing two-step multivariate statistical approach has also been proposed in [87–89] where the first step aims to estimate a physical model, based on a mixture of multi-dimensional distributions (both taking into the noise model, the relationships between the sensor responses to the objects and their physical properties). A statistical test based on this model then allows to estimate the changes. In the same spirit, the authors in [14] 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 with different numbers of looks.

Herein, we propose a different statistical approach, relying on an observation field built up from a pixel pairwise modeling on the bitemporal heterogeneous satellite image pair. This allows us to indirectly model the joint statistics or the dependencies between the two imaging modalities and to finally base our CD (or binary segmentation) model on a relevant imaging modality-invariant visual cue whose likelihood model parameters can be fully estimated within the standard ICE (Iterative conditional estimation) framework [86, 94] with ML (Maximum Likelihood) estimator in the Least Square (LS) sense. Once the estimation step is completed, the MAP (Maximum a posteriori) solution of the change detection map, based on the previously estimated parameters, is then computed with a stochastic optimization strategy.

The remainder of this paper is organized as follows: Section 4.2 describes proposed unsupervised Markovian CD model by first defining the ingredients of the proposed MRF model (likelihoods and priors), and the proposed strategy based on a two-step procedure; namely a parameter estimation step and a segmentation step. Section 7.3 presents a set of experimental results and comparisons with existing multimodal change detection algorithms. In this section, we describe the robustness assessment for our proposed technique. Finally, Section 7.4 concludes the paper.

4.2 Unsupervised Markovian CD Model

Herein, we formulate the multimodal CD problem in the unsupervised Bayesian framework. To this end, a possible and interesting approach is a two-step process. First, a parameter estimation step is conducted to infer the likelihood model parameters (in the ML sense). Then a second step is devoted to the binary segmentation or change detection itself based on the value of estimated parameters. [73].

Let y^{t_1} and y^{t_2}, a pair (co-registered) bi-temporal remote sensing (N pixel size)

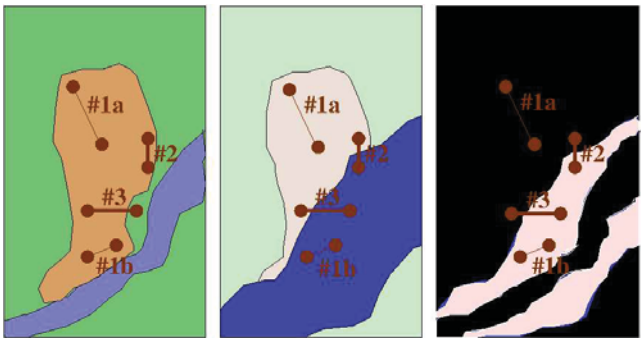

image t_1 image t_2 CD map

Figure 4.1. In lexicographic order; (synthetic) image before a flooding event, with an *urban* **region at the center, a** *vegetation* **region all around the image and a** *river* **crossing the image from right to left (bottom); image of the same area (and obtained by another imaging modality, thus with different colored textures) after a flooding event, and ground truth CD map (with the white region corresponding to the** *changed area*). **Illustration of the four pixel pair locations** $<s,t>$ **leading to the four possible cases (**$\#1a$ & $\#1b$ **: low value for** $y_{<s,t>}$ **implying that** $< s,t >$ **must share the same class label in the CD map** x**,** $\#2$ & $\#3$ **: high value for** $y_{<s,t>}$ **implying that** $<s,t>$ **must share a different class label between** s **and** t **in the final CD map** x **to be estimated. The link (between each pair of pixels considered) is drawn from such way that its thickness is proportional to the value that Eq. (4.1) could give.**

images acquired at two different times (before and after a given event), in the same geographical area, and from different sensors. We first consider $\mathbf{X} = \{X_s, s \in S\}$ the random label field located on the same rectangular lattice S of N sites s associated to the two input images, with each X_s taking its value in the discrete set $\Lambda_{label} = \{e_0 = \textit{no-change},\ e_1 = \textit{change}\}$.

4.2.1 Observation Field

In the classic monomodal (or homogeneous) CD case, the two coregistred images y^{t_1} and y^{t_2} are first compared pixel by pixel in order to generate a *difference image* by differencing or (log-)rationing (*i.e.,* by using a temporal gradient or a log temporal

gradient operator). This latter *difference image* is such that the pixels associated with land cover changes present gray-level values significantly larger, compared to those associated with unchanged areas and this visual cue based on the norm of the temporal luminance gradient $|y^{t_1} - y^{t_2}|$ is a robust cue on which the *observation field* and the likelihood distributions of a MRF model can be built up. In the multimodal (or heterogeneous) case, this temporal gradient is not a robust and reliable cue. Indeed, the color or grey value of each pixel is not a useful information since the gray levels of the same region, in y^{t_1} and y^{t_2} may be radically different according to the characteristics of the two different imaging modalities. Conversely, $y_s^{t_1}$ and $y_s^{t_2}$ may be locally coded with the same (grey or color) value in the two imaging modalities but representing two completely different textures or regions.

In our application, in order to rely on a robust visual cue with the specific property to be (nearly) invariant to the imaging modality, we have considered a pixel pairwise modeling, estimated from (y^{t_1}, y^{t_2}) and for each pixel pairs $<s,t>$ existing in S, with the following symmetric relation:

$$y_{<s,t>} = \left| |\mathbf{y}_s^{t_1} - \mathbf{y}_t^{t_1}|_1 - |\mathbf{y}_s^{t_2} - \mathbf{y}_t^{t_2}|_1 \right| \tag{4.1}$$

where $|.|_1$ is the L_1 norm and $\mathbf{y}_s^{t_1}$ and $\mathbf{y}_s^{t_2}$ represents a local statistics vector at pixel s (that will be made explicit in the following) in the *before* and *after* image.

This visual cue $y_{<s,t>}$ already proposed, in a simplified version without texture in [109][1], is defined as a function of the pixel pair $< s,t >$ and (y^{t_1}, y^{t_2}). This is discriminant in our application since, whatever the imaging modality, $y_{<s,t>}$ will give a high value for two pixels at sites s and t that must belong to two different class labels (*no-change*/*change* in our case) in the CD binary map (to be estimated) and conversely, will give a low value, for two pixels at sites s and t that must share the

[1] in which the authors define a set of constraints which will be satisfied, in the least squares (LS) sense, by a multidimensional scaling-based constraint model aiming to generate a soft CD map that is then binarized.

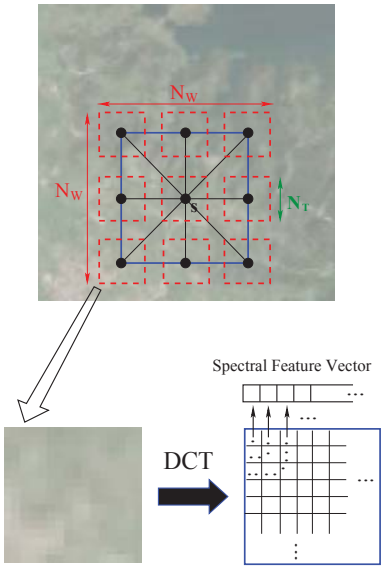

Figure 4.2. We consider, for each pixel s, a sub-sample $\mathcal{G}_s$ of 8 pairs of pixels $<s,t>$ in which the pixel t is regularly distributed around a squared window of size $N_w \times N_w$ (with $N_w = 41$ in our application). Besides $\mathbf{y}_s$ and $\mathbf{y}_t$ (see Eq. (4.1)) is in fact a radially-integrated (DCT) spectral feature vector encoding the textural and structural information existing around each local squared region of size $N_T \times N_T$ ($N_T = 16$) centered at the considered pixel.

same class label (see Fig. 4.1 and its caption).

To use this cue in our Bayesian framework, we first consider that the set of $y_{<s,t>}$ values are a realization of a random variable vector $\mathbf{Y}_{<s,t>} = \{Y_{<s,t>}, Y_{<s,u>}, ..., Y_{<u,v>}, ...\}$ gathering the $N(N-1)$ random variables associated to each site pair, that we herein call the random (pixel pairwise) observation field and secondly that $\mathbf{X}_{<s,t>}$ is its corresponding random (pairwise) label field taking its value in $\Lambda_{label_{<s,t>}} = \{id, di\}$. The pixel-pairwise label *id* means that the pixel at location s an t must share the same (**id***entical*) class label in the final CD map $\hat{x}$ to be estimated (leading to the configuration $<x_s = change, x_t = change>$ or $<x_s = no - change, x_t = no - change>$). Conversely, $x_{<s,t>} = di$ means that we have a **di***fferent* configuration, *i.e.*, either the configuration $<x_s = change, x_t = no - change>$ or $<x_s = no - change, x_t = change>$.

In our application, in order to decrease the computational load of our algorithm and to keep a quasi-linear complexity with respect to the number of image pixels, we consider for each pixel, a sub-sample $\mathcal{G}_s$ of 8 pairs of pixels regularly distributed around a squared window of size $N_w \times N_w$ centered around the pixel s (see Fig. 4.2). Besides, we consider at site s or t (image before t_1 or after t_2) a feature vector $\mathbf{y}$ (see Eq. (4.1)) encoding the textural and structural information existing around each local squared region of size $N_T = 16 \times N_T = 16$ centered at the considered pixel (see Fig. 4.2). To this end, in our application, we first estimate the Discrete Cosine Transform (DCT) of each local squared window, compute its module (*i.e.,* its absolute value since DCT is real) and then apply a half circular or Radial Integration Transform (RIT) (using a bi-linear interpolation) to estimate a spectral descriptor vector of size $N_T/2$. Since this texture descriptor is obtained from the compressed domain, this has the ability to be both, robust to noise (several denoisers are built from a filtering in this DCT domain [67,74]), be strongly reduced in size, while combining the properties to encode a texture with rotation and translation invariance. In addition, compared to a Discrete Fourier Transform (DFT), the DCT has a higher compression efficiency and above all, its spectrum is less biased than the DFT spectrum (especially when this one is computed on small images) due to the even-symmetric extension properties of DCT that avoids the generation of artifacts or spurious spectral components created by edge effects caused by the inherent periodic nature of the DFT. Also, DCT uses real computations, unlike the complex computations used in DFT. This makes the computation of DCT extremely fast[2].

[2] For the implementation of this step, we have used the very fast 16×16 (FFT2D) DCT package implemented in C code by Takuya Euro (functions DDCT16X16S tested in program SHRTDCT.C) and available online at http address given in [81].

4.2.2 Likelihood Distributions

To use the observation measure $y_{<s,t>}$ (see Eq. (4.1)) in a Bayesian settings, we must, before all, estimate the (marginal / conditional) likelihood distributions of $Y_{<s,t>}$ in the two possible cases; identical pixel-pairwise label $x_{<s,t>}{=}id$ or not $x_{<s,t>}{=}di$.

Identical Pixel-Pairwise Label Distribution

In our experiments, we have noticed that, if $x_{<s,t>} = id$, $P_{Y_{<s,t>}|X_{<s,t>}}$ is well approximated, for a given s, by an exponential distribution $p_{id} = \mathcal{E}(.;\lambda)$ with shape (or inverse rate) parameter λ, *i.e.*:

$$\begin{aligned} p_{id}(y_{<s,t>}) &= P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}=id}) \\ &= \frac{\exp\left(-y_{<s,t>}/\lambda\right)}{\lambda} \cdot H(y_{<s,t>}) \end{aligned} \tag{4.2}$$

with the right-continuous Heaviside step function, $H(x)$ where $H(0) = 1$ and $\lambda > 0$ (which makes the distribution supported on the interval $[0\ \ \infty]$).

This approximation can be justified and understood if we notice that, for a pixel pair $<s,t>$ located in a spatially and temporally homogeneous region (*e.g.*, cases $\#1a\,\&\,\#1b$ illustrated in Fig. 4.1), *i.e.* for $x_{<s,t>} = id$, $y_{<s,t>}$ is in fact related to the norm of a first order temporal gradient over a n-order (n is the distance in pixel between s and t) spatial gradient and the gradient norm of the intensity image is known to be well approximated by a simple exponential distribution [85] or its numerous variant (such as its truncated [19] [20], generalized [115] or long-tail version with a shape and scale factor [70, 72]).

Different Pixel-Pairwise Label Distribution

In the case of $x_{<s,t>} = di$ (different pixel-pairwise labels), we have empirically noticed that the Gaussian law $p_{di} = \mathcal{N}(.; \mu, \sigma^2)$ is well adapted to describe the measure $y_{<s,t>}$:

$$\begin{aligned} P_{di}(y_{<s,t>}) &= P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}=di}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_{<s,t>} - \mu)^2}{2\sigma^2}\right) \end{aligned} \tag{4.3}$$

Let us note that, in the case of a heterogeneous pair of images and two heterogeneous temporal regions ($x_{<s,t>} = di$), this distribution is consistent with the central limit theorem and the fact that this results from the addition of lots of different phenomena (*i.e.,* lot of numerical differences achieved between many possible different textural feature vectors, coded by different imaging modality with possibly different scales, etc.).

Data Likelihood and Posterior Distribution

Now, if we assume that the pairwise data $Y_{<s,t>}$ are independent conditionally on the pairwise labeling process $X_{<s,t>}$, and take into consideration the sub-sample $\mathcal{G}_s$ of pairs of pixels defined in Section 4.2.1 (and shown in Fig. 4.2), one gets:

$$P_{\mathbf{Y}_{<s,t>}|\mathbf{X}_{<s,t>}}(.) = \prod_{s \in S} \prod_{\substack{<s,t> \\ t \in \mathcal{G}_s}} P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}}) \tag{4.4}$$

In addition, if we consider that the distribution of $\mathbf{X}$ is stationary and Markovian and choose a standard prior for the distribution of the labeling process $\mathbf{X}$ and that the CD map x defines $x_{<s,t>}$ without ambiguity, one gets for the posterior distribution:

$$P_{\mathbf{X}|\mathbf{Y}_{<s,t>}}(.) \propto \prod_{s \in S} \prod_{\substack{<s,t> \\ t \in \mathcal{G}_s}} P_{Y_{<s,t>}|X_{<s,t>}}(.) \cdot P_X(x) \tag{4.5}$$

If we consider a standard isotropic Pott-type prior model relative to the second-order neighborhood system η_s, with identical potential value β for the different (horizontal

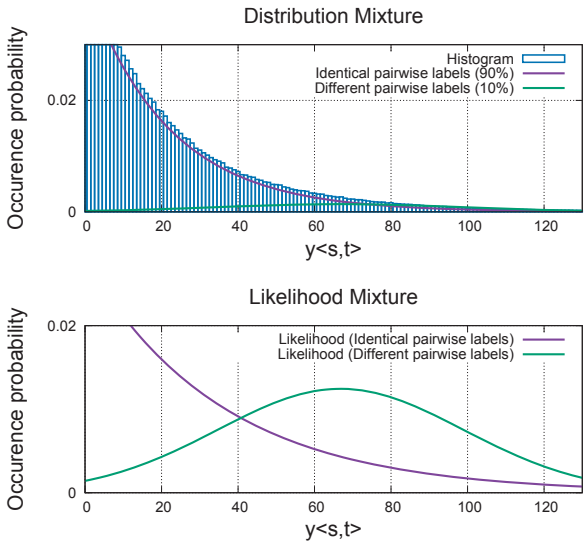

Figure 4.3. From top to bottom; Distribution mixture: Histogram of $y_{<s,t>}$ associated to the heterogeneous image pair Dataset-3 and the two weighted (90% of identical pairwise labels and 10% of different pairwise label) mixture components that are estimated by the ICE procedure (see Section 4.2.3). Likelihood mixture: the two preceding likelihood distributions (without proportion priors) that are estimated by the ICE procedure.

vertical, right diagonal or left diagonal) *cliques* $\langle s,t\rangle$ of η_s, thus a model favouring for $\hat{x}$, homogeneous regions of the same class *no-change* or *change*; *i.e.*, $P_X(x) \propto -\beta \exp\{\sum_{\langle s,t\rangle\in\eta_s}[1-\delta(x_s,x_t)]\}$ [7], (where δ is the delta Kronecker function) $\hat{x}$, the CD map to be estimated becomes the global maxima of the following corresponding posterior probability:

$$\begin{aligned}
\hat{x} &\propto \arg\max_x \prod_{s\in S} P_{X_s|Y_{<s,t>}}(.) \\
&\propto \arg\max_x \prod_{s\in S} \Bigg\{ \prod_{\substack{<s,t> \\ t\in\mathcal{G}_s}} P_{Y_{<s,t>}|X_{<s,t>}}(.) \\
&\cdot \underbrace{\exp -\Big\{\beta \sum_{\langle s,t\rangle\in\eta_s}[1-\delta(x_s,x_t)]\Big\}}_{P_{X_s}(x_s)} \Bigg\}
\end{aligned} \tag{4.6}$$

In this context, the corresponding posterior energy to be minimized is:

$$\begin{aligned}
U(x,y) = &\sum_{s\in S} \sum_{\substack{<s,t> \\ t\in\mathcal{G}_s}} -\ln P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}}) \\
&+ \sum_{\langle s,t\rangle\in\eta_s} \beta[1-\delta(x_s,x_t)]
\end{aligned} \tag{4.7}$$

and $\hat{x}_{\text{MAP}} = \arg\min_x\{U(x,y)\}$.

4.2.3 Iterative Conditional Estimation

Principle

In our unsupervised Markovian segmentation case, we have to estimate in a first step (*estimation step*), the parameter vector $\mathbf{\Phi}_{y_{<s,t>}}$ which defines respectively the likelihood distributions $p_{id}(y_{<s,t>})$ and $P_{di}(y_{<s,t>})$ (or $P_{Y_{<s,t>}|X_{<s,t>}}(y_{<s,t>|x_{<s,t>}})$ for each two classes $x_{<s,t>}$ of $y_{<s,t>}$. (see Equations (4.2)-(4.3)), *i.e.*, the parameter vector $\mathbf{\Phi}_{y_{<s,t>}}(\lambda,\mu,\sigma)$ gathering the scale parameter of the exponential law $p_{id}(y_{<s,t>})$ and the mean μ and σ parameters of the Gaussian distribution $p_{di}(y_{<s,t>})$.

In our case, this estimation step is particularly challenging for three reasons; first, one has to deal with a mixture of different distributions (exponential and Gaussian) which are also strongly mixed (see Fig. 4.3) and which also exhibits different mixing proportions (generally the class *di* is under weighted ($<15\%$) because this class is related to the fewer pixel-pairwise labels, or transitions, existing between the class *change* and the class *no-change* (see Fig. 4.1).

To this end, we resort to the ICE [86,94] iterative procedure which is able to cope with different distributions and which experimentally turned out to be more efficient than the classical Expectation Maximization (EM) [18] algorithm or its stochastic version; the Stochastic EM (SEM) [61]. This efficiency can be explained by the fact that the ICE [86, 94] procedure can also be viewed as the stochastic and Markovian version of the EM procedure and thus is also constrained by the distribution of $\mathbf{X}$ defined as stationary and Markovian.

The ICE procedure first requires to find an estimator $\hat{\mathbf{\Phi}}_{y_{<s,t>}} = \mathbf{\Phi}(x_{<s,t>}, y_{<s,t>})$ providing an estimate of $\mathbf{\Phi}_{y_{<s,t>}}$ based on the complete data configuration $(x_{<s,t>}, y_{<s,t>})$. Random field $\mathbf{X}_{<s,t>}$ being un-observable, the iterative ICE procedure thus defines the parameter $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$, at step $[k+1]$, as the conditional expectations of $\hat{\mathbf{\Phi}}_{y_{<s,t>}}$ given $Y_{<s,t>} = y_{<s,t>}$ and the current parameter $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$. The fixed point of this iteration corresponds to the best approximations of $\mathbf{\Phi}_{y_{<s,t>}}$ in terms of the mean squared error [94]. By denoting E_k the conditional expectation based on $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$, this iterative procedure is defined as follows:

- One takes an initial value $\mathbf{\Phi}^{[0]}_{y_{<s,t>}}$
- $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$ is computed from $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$ and from $y_{<s,t>}$ using:

$$\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}} = E_k\big[\hat{\mathbf{\Phi}}_{y_{<s,t>}}(x,y) \,|\, \mathbf{Y}_{<s,t>}{=}y_{<s,t>}\big]$$

The computation of this expectation is impossible in practice, but we can approach

it thanks to the law of large numbers [94]:

$$\mathbf{\Phi}_{y_{<s,t>}}^{[k+1]} = \frac{1}{n}\left[\hat{\Phi}_{y_{<s,t>}}(x_{<s,t>}^{(1)}, y_{<s,t>}) + \cdots + \hat{\mathbf{\Phi}}_{y_{<s,t>}}(x_{<s,t>}^{(n)}, y)\right]$$

where $x_{<s,t>}^{(i)}, i = 1, \ldots, n$ are realizations drawn from the posterior distribution: $P_{X_{<s,t>}|Y_{<s,t>},\Phi}(x_{<s,t>}|y_{<s,t>}, \mathbf{\Phi}_{y_{<s,t>}}^{[k]})$.

In our application, since x completely defines $x_{<s,t>}$ without ambiguity[3], these realizations can be drawn from the posterior distribution $P_{X|Y_{<s,t>},\Phi}(x|y_{<s,t>}, \mathbf{\Phi}_{y_{<s,t>}}^{[k]})$ (see Section 4.2.2 and Eq. (4.5)). As it turns out, $n = 1$ is sometimes found sufficient (or even better) to get good estimates [94]. It is the case in our unsupervised Markovian CD model, and we actually chose $n = 1$ in our experiments.

ICE-Based ML estimator

For the Gaussian law, a ML estimate of (μ, σ^2), based on the complete data configuration, can be easily given by the empirical mean and empirical variance. If $N_{di} \stackrel{\triangle}{=} \#\{x_{<s,t>} = di\}$, one gets:

$$\hat{\mu}(x_{<s,t>}, y_{<s,t>}) = \hat{\mu}(x, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=di} y_{<s,t>}}{N_{di}} \tag{4.8}$$

$$\hat{\sigma}^2(x, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=di} (y_{<s,t>} - \hat{\mu})^2}{(N_{di} - 1)} \tag{4.9}$$

For the exponential law, if $N_{id} \stackrel{\triangle}{=} \#\{x_{<s,t>} = id\}$, a ML estimate of the shape parameter is:

$$\hat{\lambda}(x, y_{<s,t>}) = \frac{\sum_{x_{<s,t>}=id} y_{<s,t>}}{N_{id}} \tag{4.10}$$

In our Bayesian CD framework, we do not need to estimate the proportion of each class. Nevertheless, the mixing proportion can be easily estimated within this procedure with the empirical frequency estimator; $\pi_{id} = N_{id}/(N_{id} + N_{di})$ and $\pi_{di} = N_{di}/(N_{id} + N_{di})$.

[3] but the inverse is not true.

ICE Algorithm

$\mathbf{\Phi}_{y_{<s,t>}}(\lambda, \mu, \sigma^2)$ are thus estimated with the ICE procedure in the following way:

• *Parameter Initialization*: we start from a CD map x randomly sampled from two classes (*change / no-change*) and start from $\mathbf{\Phi}^{[0]}_{y_{<s,t>}} = (\lambda^{[0]}, \mu^{[0]}, \sigma^{2\,[0]})$.

• **ICE procedure**: $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$ is then computed from $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$ as follows:

1. *Stochastic Step*: using the Gibbs sampler, one realization x of the CD map is simulated according to the posterior distribution $P_{X/Y_{<s,t>}}(x/y_{<s,t>})$, with parameter vector $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$. More precisely, for each site s (lexicographically), we sample x_s with the local version of Eq. (4.5), *i.e.*,

$$P_{X_s|Y_{<s,t>}}(.) \propto \prod_{\substack{<s,t> \\ t \in \mathcal{G}_s}} P_{Y_{<s,t>}|X_{<s,t>}}(.) \cdot P_{X_s}(x_s) \tag{4.11}$$

   (a) with $P_{Y_{<s,t>}|X_{<s,t>}}$ an Exponential law for $x_{<s,t>}=id$ (see Section 4.2.2).

   (b) with $P_{Y_{<s,t>}|X_{<s,t>}}$ a Gaussian law for $x_{<s,t>}=di$ (see Section 4.2.2).

2. *Estimation Step*: the parameter vector $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$ is estimated with the ML estimator of the "complete data" (see Eqs (4.8), (4.9), (4.10)).

3. Repeat until convergence is achieved; *i.e.*, if $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}} \not\approx \mathbf{\Phi}^{[k]}_{y_{<s,t>}}$, we return to Stochastic Step.

In our application, one has to deal with a mixture of different distributions which are strongly mixed with unbalanced mixing proportions (see Fig. 4.1). This makes the convergence of the ICE procedure still difficult in some cases. Thus, it is necessary to add an additional hard constraint. In our application, we can capture the fact that the shape parameter λ of the exponential distribution $p_{id}(y_{<s,t>})$ is in fact not too far from its shape parameter computed from the set of $y_{<s,t>}$ regardless of its label $x_{<s,t>}$

(id or di) since there are generally fewer labels di (let $\lambda^\star$ this parameter value). In fact, since the true shape parameter λ of the exponential distribution $p_{id}(y_{<s,t>})$ is computed from $y_{<s,t>}$ given $x_{<s,t>} = id$, λ is thus computed from a subset of smaller values of $\{y_{<s,t>}\}$, or equivalently, we can surely assert that a reliable estimation for λ is necessarily a value inferior to $\lambda^\star$. We model this by imposing the hard constraint $\lambda = \lambda^\star/\alpha$ for the different iteration of the ICE procedure.

In order to further help the iterative ICE procedure, we start, at iteration $^{[0]}$ with $\mathbf{\Phi}^{[0]}_{<s,t>} = (\lambda^{[0]}, \mu^{[0]}, \sigma^{[0]})$, with $\mu^{[0]} = 2\,\lambda^{[0]}$ (with $\lambda^{[0]}{=}\lambda^\star$) and $(\sigma^2)^{[0]} = 1000$ to model the fact that the *mean* of the Gaussian is generally greater to the λ parameter and that the variance of the Gaussian is generally around 1000. We finally use the *Stochastic Step* with a Gibbs sampler with a temperature equals to 0.25 in order to allow a fast convergence and to reduce the number of explored solutions around the initialization values.

Table 4.1. Description of the four heterogeneous datasets

Dataset	Date	Location	Size (pixels)	Event (& Spatial resolution)	Sensor
1	Sept. 1995 - Jul. 1996	Sardinia, It	412×300	Lake overflow (30 m.)	Landsat-5 Thermic (NIR band) / Optical
2	July 2006 - July 2007	Gloucester, UK	2325×4135	Flooding (0.65 m.)	TerraSAR-X / QuickBird 02
3	Feb. 2009 - July 2013	Toulouse, Fr	4404×2604	Construction (2 m.)	TerraSAR-X / Pleiades
4	May 2012 - July 2013	Toulouse, Fr	2000×2000	Construction (0.52 m.)	Pleiades / WorldView 2

4.2.4 Segmentation Step

Once the estimation step is completed, the MAP (Maximum a posteriori) solution of the CD map x, based on the previously estimated parameters, is then computed. In our application, the energy function (see Eq. (4.7)) is complex and the MAP solution is difficult to estimate (essentially due to the strongly mixed likelihood mixture model which is possibly of slightly different shapes according to the type of multimodality). In order to avoid local minima we must resort to a simulated annealing (SA) procedure

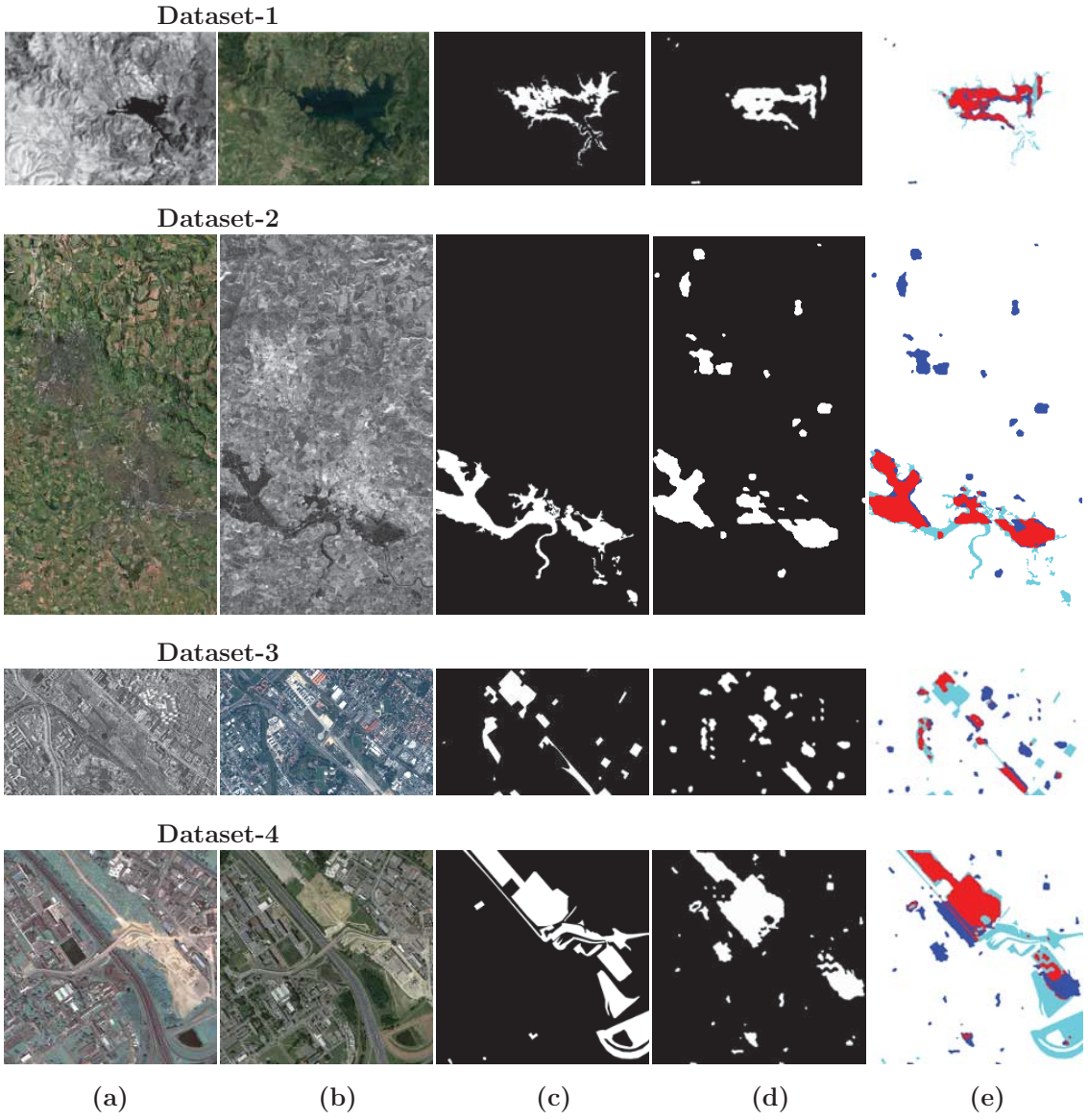

Figure 4.4. Heterogeneous datasets (see Table 4.1). (a-c) image t_1, t_2, ground truth; (d) final (changed-unchanged) segmentation result and (e) confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

[7] with a sufficient number of iterations about 200000 iterations in our application), or equivalently by varying the temperature of a Gibbs sampler (see Eq. (4.11)) from the initial temperature $T_o = 1.25$ to $T_{final} = 0.01$ with a slow geometric decreasing

Table 4.2. Confusion matrix in terms of number of pixels and percentage for the four heterogeneous datasets *i.e.*, [TM/TM], [TSX/QB02], [TerraSAR-X/Pleiades], [Pleiades/WorldView 2] (see Table 4.1).

Multimodal pair	TP	TN	FP	FN
Thermic/Optical (Landsat-5)	5189 (67.3%)	114007 (98.4%)	1884 (1.6%)	2520 (32.7%)
QB02/TerraSAR-X	5272 (69.3%)	10782 (78.3%)	2990 (21.7%)	2337 (30.7%)
TerraSAR-X/Pleiades	4025 (35.8%)	124468 (95.7%)	5611 (4.3%)	7217 (64.2%)
Pleiades/WorldView2	15904 (41.8%)	199794 (94.3%)	12164 (5.7%)	22138 (58.2%)

schedule such as $T = T_o \times (0.999975)^k$.

Once $\hat{x}_{\text{MAP}}$ is estimated, it is important to note that, due to the pixel (label) pairwise modelling, there are two global minima to the optimization problem defined in Eq. (4.7). One for the solution ("1" for *change* class and "0" for *no-change* class) and the second one corresponding to its binary inverse (*i.e.*, its binary complement, with "0" for *change* class and "1" for *no-change* class). In our case, this ambiguity can be easily resolved with a correlation metric or more simply by assuming that the land cover change is generally smaller than the unchanged area.

It takes between 30 and 70 minutes to perform a SA (depending on the image size) with so many iterations for a non-optimized C++ code running on Linux on a $i7-930$ Intel CPU, 2.8 GHz. Nevertheless, by considering a Jacobi-type version of the Gauss-Seidel based SA procedure [40], the final energy-based minimization procedure can be efficiently implemented by using the parallel abilities of a graphics processor unit (GPU) with a speed gain up to (about) 200 [40].

4.3 Experimental Results

4.3.1 Heterogeneous Dataset Description

To validate our approach, we present in this section a series of tests conducted on four real heterogeneous (multimodal) datasets, reflecting different change detection

Table 4.3. Accuracy rate of change detection on the four heterogeneous datasets obtained by the proposed method and the state-of-the-art multimodal change detectors (first upper part of each Table) and mono-modal change detectors (second lower part of each Table).

Optical(NIR band)/Optical [#1]	Accuracy
Proposed method	**0.964**
Touati *et al.* [111]	0.942
Zhang *et al.* [126]	0.975
PCC [126]	0.882

SAR/Optical [#2]	Accuracy
Proposed method	**0.955**
Touati *et al.* [109]	0.949
Touati *et al.* [110]	0.932
Prendes *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

Optical/SAR [#3]	Accuracy
Proposed method	**0.909**
Touati *et al.* [109]	0.867
Touati *et al.* [111]	0.878
Prendes *et al.* [87, 90]	0.918
Prendes *et al.* [88]	0.854
Copulas [65, 88]	0.760
Correlation [65, 88]	0.688
Mutual Inf. [65, 88]	0.768
Pixel Dif. [88, 104]	0.782
Pixel Ratio [88, 104]	0.813

Optical/Optical [#4]	Accuracy
Proposed method	**0.862**
Touati *et al.* [109]	0.853
Touati *et al.* [110]	0.870
Prendes *et al.* [87, 89]	0.844
Correlation [87, 89]	0.679
Mutual Inf. [87, 89]	0.759
Pixel Dif. [87, 104]	0.708
Pixel Ratio [87, 104]	0.661

conditions in multimodal case (see Table 4.1); Namely, (#1 and #4) two multisensor optical datasets (*i.e.*, same sensor type but with two different optical sensors or same satellite sensor but with different specifications), (#2-#3) two multisource datasets (*i.e.*, different sensor types), respectively optical/SAR and SAR /optical. This allows us to compare the performance of the proposed method with different state-of-the-art multimodal change detection algorithms recently proposed in this field [65, 88, 109–111] in different multimodal CD conditions, and also for a wide variety of changed event when the resolution varies from 0.52 to 30 meters. In this benchmark, all the ground-truth images (change detection mask) was provided by an expert photo interpreter.

4.3.2 Results & Evaluation

In all the experimental results, we have considered the simple grey level of the image (and thus converted, when necessary, the optical color image to grayscale), reduce the size of the image such that its maximal size (length or width) is around 500 pixels and use a double histogram matching.

The internal parameter of our Markovian model are for, from decreasing order of importance, the parameter α of the data likelihood (see Subsection 4.2.3), the parameter β of the prior model (see Subsection 4.2.2) and the length N_w of the graph $\mathcal{G}_s$ (see Subsection 4.2.1 and Fig. 4.2) for which the sensitivity is not important. We do not consider the parameter N_T as an important internal parameter; in fact, we have taken $N_T = 16$ in order to use the very fast (since 16 is a power of 2) DCT package implemented in C code by [81][4]. In our application, the DCT is thus applied on the grey-scale band of the image or the gray-level band resulting from the grayscale conversion of the three color bands (for a color image). For all the experimental results, we use $\alpha = 1.5, \beta = 0.1, N_w = 41$.

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: PCC = (TP+TN)/(TP+TN+FN+FP) where TP, TN, FN, FP designate classically the true positives, negatives, and false negatives and positives.

A comparison with different state of the art approaches [65, 88, 109–111] is summarized in Table 7.2. We have also summarized in Table 7.3 the confusion matrix obtained by our proposed Markovian CD model. From Table 7.2, we can see that the rate accuracy of our method performs very well and outperforms in average the other state-of-the-art approaches.

[4] We have also tested $N_T = 8$ and noticed that the classification results was slightly altered in our application.

The average accuracy rate obtained on the four multimodal dataset based on our Markovian CD approach is 92.3% with well balanced confusion matrices (see Table 7.3).

4.3.3 Results on Homogeneous Dataset with Shadow Effects

As an additional experiment, it is also interesting to see how the proposed unsupervised Markovian CD model behaves and adapts in the presence of homogeneous images (see Fig. 4.5) when one of the two images has glow and shadow effects. To this end, for this (non-trivial) homogeneous CD detection case, we have considered a stereo panchromatic data set provided by [105], with size 900×900 pixels (pixel resolution is 5 meters) and captured by the Cartosat-1 satellite sensor. This pair of panchromatic images is acquired over the Arges region (Roumania near Piatra Craiului national park), on Oct. 2008 and Nov. 2009 and shows a forest changes caused by storms, and containing many shadow areas caused by steep terrain due to the mountainous forest area [105]. From Table 4.4, we can see that the kappa coefficient of our method is correct and quite comparable to others state-of-the-art homogeneous CD approaches, though slightly less good (than the methods purely dedicated and optimized for the monomodal case). In fact, our model remains ideally and best suited for the multimodal CD case with a mixture of distributions specifically chosen to take into account a (quite large) number of pairs of rather different imaging modalities usually observed in remote sensing.

4.3.4 Discussion

Concerning the technical specifications of the proposed model, we have noticed that the L1 norm, for the pixel pairwise spatio-temporal difference (used as visual cue) in Eq. (4.1), is slightly more robust than the L2 norm for which we obtain an average accuracy rate (obtained on the four multimodal datasets) of 89.3% (versus 92.3% for

Panchromatic shadow dataset

(a) (b) (c) (d) (e)

Figure 4.5. Panchromatic data set: (a-c) image t_1, t_2, ground truth; (d) final (changed-unchanged) segmentation result and (e) confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

Table 4.4. Kappa statistic [105] $(po - pe)/(1 - pe)$ **(with po = observed accuracy = (TP + TN) / (TP + FP + FN + TN) and pe = expected accuracy = [(TP + FP)(TP + FN) + (FN + TN)(FP + TN)] / [(TP + FP + FN + TN)^2]) of change detection on the Panchromatic shadow dataset obtained by the proposed method and comparisons other unsupervised (first upper part of the table) and supervised (second part of the table) state-of-the-art monomodal change detectors [105].**

Method	Kappa
Proposed method	**0.403**
Touati *et al.* [109] (with preprocessing)	**0.513**
Touati *et al.* [109] (without preprocessing)	0.281
kMNF OPTI* [105]	0.487 - **0.509** - 0.506 - 0.501 - 0.487 - 0.475
Height Difference* [105]	0.127 - 0.316 - 0.469 - **0.526** - 0.0 - 0.0
CVA* [105]	0.07 - 0.242 - 0.403 - **0.457** - 0.0 - 0.0
k-Means [105]	0.472
ICDA [105]	0.495
OSVM [105]	0.478
Random Forests [105]	0.432

*based on different threshold levels given in increasing order

the L1 norm). Besides, it is important to mention that our choice concerning the likelihood distributions was made after a pre-study where we empirically tried different mixtures of statistical laws. More precisely, we have successively tried different

law combinations including, for identical pixel-pairwise labels (in addition to the exponential law that was finally used), an half Gaussian, Rayleigh and Gaussian laws along with for different pixel-pairwise labels, (in addition to the Gaussian law that was finally used); a Rayleigh, an exponential and finally an uniform distribution. The best combination was the mixture of Exponential/Gaussian likelihood distributions used in our model and presented in Section 4.2.2.

From the experiment, we can notice that the CD result in multisensor optical Dataset 4 is the least accurate of the four examples given. We think that this can be explained by several reasons. The first one is due to the macro texture generated by the (high-resolution) satellite view of the (Toulouse) urban area. DCT features have more difficulties to model such macro textural patterns and is in fact better suited to model micro-textural features usually present in a lower resolution satellite image (as datasets #1-3). The second reason is due to the nature of change. In this image, the change (*i.e.*, an area under construction) can be subtle and light and thus difficult to distinguish even with a trained eye. Thirdly, the different colors between the two optical images give, after grey-level conversion, different grey levels which may further complicate the CD result.

It is interesting to notice that, in a way, the proposed herein model can be viewed as the Markovian version, thus in the ML sense of the LS model, based on the Multi-dimensional scaling (MDS) mapping proposed in [109] (however, we herein consider a slightly different observation field including texture information).

Let us also note that, in the ML criterion sense, we try to maximize the posterior probability of a given (pair of observation(s) and consequently this one is thus closely related both to the choice of the observation field (in our case $y_{<s,t>}$ and also, above all, the choice of the mixture of distributions (in our case Exponential/Gaussian). We think that more flexible (or generalized) distribution laws would be perhaps more suited to the heterogeneous remote sensing imagery (*i.e.,* thus leading to a better

model) but this flexibility would be at the cost of a more complicated (already very complex and computational demanding) final optimization procedure.

The overall unsupervised Markovian CD proposed model is outlined in pseudo-code in Algorithm 5.

The C++ code running on Linux, data, and all that is necessary for reproduction of the results shown in this paper is freely accessible at http address[5] .

4.4 Conclusion

In this paper, we have addressed the problem of change detection in heterogeneous remote sensing. Although this issue has become important, due to the huge amount of heterogeneous data, we can now get from various archives or from existing (and different types of) Earth observing satellites, it has only received little attention in the literature. In addition, this issue has really been very little discussed in the statistical field and, to our knowledge, no Bayesian or Markovian-based multimodal CD method has been proposed until now. This paper fills the gap by proposing a complete unsupervised Markovian approach which has been validated on a number of real multimodal bitemporal satellite image pairs and whose the main novelty, and not only in Geoscience imaging, lies in the use of an observation field built up from a pixel pairwise modeling. In fact, in our application, in order to decrease the computational load of our algorithm, we consider for each pixel, a sub-sample of pairs of pixels. Nevertheless, the proposed MRF model turns out to iteratively propagate information *via* this sub-sample of pairs of pixels very efficiently during the estimation or segmentation step, while keeping a quasi-linear complexity with respect to the number of image pixels. We also think that the concept of pixel pairwise modeling

[5] http://www.iro.umontreal.ca/∼mignotte/ResearchMaterial

M3CD Algorithm

Input: Pair of bi-temporal satellite images: (y^{t_1}, y^{t_2})
Output: A binary CD segmentation map: x

α Hard constraint param. of the likelihood model
β Regularization term of the Markov. prior model
N_w Length of the graph $\mathcal{G}_s$
r Cooling rate of the simulated annealing (SA)
$T_{o,f}$ Initial and final Temp. of the SA

$\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$ Parameter vector $(\lambda^{[k]}, \mu^{[k]}, \sigma^{[k]})$ gathering the scale, mean and variance of the likelihood mixture at iteration k
$x^{[k]}$ CD (binary) map at iteration k

1. Initialization Step

∘ Co-Registration of the image pair (y^{t_1}, y^{t_2})
∘ Conversion of (y^{t_1}, y^{t_2}) into grayscale (if necessary)
∘ Image size reduction of (y^{t_1}, y^{t_2}) until the maximum (length or width) side is ≈ 500 pixels
∘ Double histogram matching on (y^{t_1}, y^{t_2})

2. Parameter Estimation Step

▷ Initialization:
. $x^{[0]} \leftarrow$ Rand. sampling from 2 classes (*change / no*)
. $\lambda^\star \leftarrow \sum y_{<s,t>} / (N_{id} + N_{di})$ (shape param. of $\{y_{<s,t>}\}$)
. $\mathbf{\Phi}^{[0]}_{y_{<s,t>}} \leftarrow (\lambda^{[0]}, \mu^{[0]}, \sigma^{[0]})$
with $\mu^{[0]} = 2\lambda^{[0]}$, $\lambda^{[0]} = \lambda^\star$ and $(\sigma^2)^{[0]} = 1000$
. $k \leftarrow 0$

▷ ICE Procedure:
while $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}} \not\approx \mathbf{\Phi}^{[k]}_{y_{<s,t>}}$ **do**
  for *each pixel at site s (lexicographically)* **do**
    Sample $x_s^{[k]}$ with Posterior dist. $P_{X_s|Y_{<s,t>}}(.)$ based on graph $\mathcal{G}_s$, temperature $T = 0.25$ and $\mathbf{\Phi}^{[k]}_{y_{<s,t>}}$ [see Eq. (11)]
  ∘ $\mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$ is estimated with the ML estimator of the "complete data" $(y_{<s,t>}, x^{[k]})$ [see Eqs (8-10)]
  ∘ Hard constraint: $\lambda^{[k+1]} = \lambda^\star / \alpha$
  . $k \leftarrow k + 1$
$\mathbf{\Phi}_{y_{<s,t>}} \leftarrow \mathbf{\Phi}^{[k+1]}_{y_{<s,t>}}$

3. CD Segmentation Step

▷ Initialization: $k \leftarrow 0$ and $T \leftarrow T_o$

▷ Simulated Annealing Procedure:
while $T > T_f$ **do**
  for *each pixel at site s (lexicographically)* **do**
    Sample $x_s^{[k]}$ according to the Gibbs Posterior distribution $P_{X_s|Y_{<s,t>}}(.)$ based on graph $\mathcal{G}_s$, temperature T and $\mathbf{\Phi}_{y_{<s,t>}}$ [see Eq. (11)]
  . $k \leftarrow k + 1$ and $T \leftarrow T_0 \cdot \mathrm{r}^k$

Algorithme 5: M3CD (Markov Model for Multimodal Change Detection) algorithm

can be interesting for other issues in traditional digital image processing, not only in Geoscience imaging, since the underlying framework based on pixel-pairwise affinity can really model complex statistical phenomena with possibly important invariance properties.

Acknowledgement

We would like to acknowledge the Computer Research Institute of Montreal **(CRIM)** and the Ministry of Economic Science and Innovation **(MESI)** of the Government of Québec to have supported this work. We would like also to acknowledge all other researchers that made at our disposal the change detection dataset in order to validate the proposed anomaly change detection model.

Chapitre 5

CHANGE DETECTION IN HETEROGENEOUS REMOTE SENSING IMAGES BASED ON AN IMAGING MODALITY-INVARIANT MDS REPRESENTATION

Dans ce chapitre, nous présentons notre article présenté dans la conférence *I*EEE International Conference on Image Processing, intitulé: **Change Detection in Heterogeneous Remote Sensing Images Based On an Imaging Modality-Invariant MDS Representation**. Nous exposons ce dernier dans sa langue originale de publication.

Abstract

In this paper, we propose a new multimodal change detection in remote sensing. The proposed method is based on a projection of the two multisensor satellite images to a common feature space, in which the two heterogeneous images share the same statistical properties and on which any classical monomodal change detection methods can be applied. This transformation of the before and after images is mainly based on a Multidimensional Scaling(MDS) representation which can be also viewed as a *de-texturing* approach of the two multisource images. Experimental results involving different types of imaging techniques confirm the reliability of the proposed approach.

5.1 Introduction

Multimodal Change Detection (CD) consists in identifying any land cover changes/uses that may have occurred between two satellite images acquired on the same geographi-

cal area, at different times, by two different kinds of imaging techniques. It is a recent and challenging task in the area of remote sensing, also called *multi-sensor data fusion*, that actually generalizes the classical monomodal CD issue [10, 33, 108] already used for solving the environmental monitoring, geological resources surveys and disaster detection/localization and quantification to name a few. The combination of images acquired by different sensor types (*e.g.* active and passive) or the finding of reliable imaging modality-invariant features, coming from different data sources is a difficult task. However, this difficulty is widely compensated by the numerous practical and technical advantages of such multimodal analysis procedure. Indeed, with the development of satellite and remote sensing imaging technology, a huge amount of heterogeneous data are acquired every day and stored in data archives for later use. By this fact, it can happen that, for example, an optical image of an area, provided by an archive, have to be necessarily combined with a new SAR image (of the same area) for technical reasons, lack of time, availability or atmospheric conditions in an emergency situation (SAR sensors can operate regardless of weather or thermal conditions, even at night, *i.e.* with less restrictive conditions compared to optical imaging). It is also worth mentioning that, since a multimodal CD analysis processes heterogeneous data with different statistics, this new technique may be more robust to natural variations in environmental variables such as soil moisture or phenological states or shading effects which should not be detected as major land cover changes. Until now, among the few research works that have been devoted to heterogeneous CD problem, we can identify four main categories. Namely; parametric models, non-parametric or learning machine based methods, algorithms based on operators using spatial and temporal similarity measures with invariance according to the imaging modality or finally, procedures mainly based on a transformation or projection of the two multimodal images to a common feature space, in which the two heterogeneous images share the same statistical properties and on which classical monomodal CD methods can then be applied. In parametric models, a mixture or a set of parametric

multidimensional distributions are generally used to model the joint statistics or the dependencies between the two imaging modalities [14, 65]. Sometimes, these models take also into account the noise characteristics and the relationships between the sensor responses to the objects and their physical properties [88, 90]. The main problems related with these parametric models are that they have been especially designed with specific distribution laws related to a type of multimodal sensors and are not easily generalizable for another pair of different sensors. In addition, these methods require a Maximum Likelihood (ML) parameter estimation step of the considered distribution, which can be complex and computationally expensive. Sometimes, these models are also semi-supervised and rely on a training set to fit the parametric model. Among nonparametric methods, an energy minimization model has been specifically designed and solved in the least-squares sense in [109] 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. Deep learning methods through conditional adversarial networks [66] or convolutional coupling networks [51] have also been proposed and turn out to be valuable for the multimodal CD problem. In fact, these nonparametric methods have the ability to adapt to a wide variety of different imaging modalities (with possibly different noise types and levels) but are also generally less accurate than a parametric model dealing with a specific type of multimodality represented by a particular distribution whose shape is clearly theoretically determined. In the third family of methods relying on similarity measures with invariance according to the imaging modality, Alberga *et al.* [2] propose to use a technique closed to the co-registration and based on the use of a combination of different invariant similarity measures (such as correlation ratio, mutual information, etc.). Also, authors in [8] presented a CD method to quantify the damages caused by an earthquake to each individual building from a pre-event optical and post-event SAR images. In this work, simulation is used to predict the expected SAR signature of each building from the optical image which is then compared to the actual SAR scene to quantify the

damages caused to each building. In [110], an imaging modality-invariant operator that detects the common specific high-frequency pattern of each structural region existing in the two heterogeneous satellite images is proposed. Finally, in the last category in which the bitemporal image data is projected to a common feature space for comparison convenience, [117] proposes also a representation, especially designed to highlight the changes. Another representation which turn out to be invariant to imaging modality is given by a classical segmentation. In this way, Liu *et al.* in [29] propose a general multidimensional evidential reasoning approach for estimating the segmentation map of the two satellite images which are then easily and subsequently compared. In this work, we propose a new multimodal CD method belonging to the last category and based on a common feature space thus making possible the direct comparison between the two input images. This transformation of the before and after images can be viewed as a *de-texturing* approach [68] of each satellite image.

5.2 Proposed Change Detection Model

Imaging Modality Invariant Projection: This step aims at finding a common feature space in which the pixels of the two satellite image should ideally have the same statistical properties. This task is not trivial, especially when the SAR imaging modality has to be combined with the optical imaging technique because the textural properties of the two images are radically different; For the SAR image, the inherent multiplicative speckle noise creates for each land cover class, a kind of macro-texture with grainy patterns (related to the back-scattered intensity of the different object surfaces). For the optical image, the noise is additive and degrades either piece-wise uniform areas or micro-textured and structured regions (representing in fact the reflection intensity of objects). A solution consists in *de-texturing* the two satellite images, *i.e.*, to create a new (grey level) mapping in which two textured areas (around pixels at distant locations) gives, in the transformed image, two pixels whose grey-

level intensity difference is proportional to a distance measure between these two textures. Otherwise said, in this new mapping, two non-adjacent or distant pixels with the same local texture (around the pixel) should have the same (grey-level) intensity.

To this end, a de-texturing approach, close to the one proposed in [68], is applied respectively on the first and second input satellite images. To this end, each pixel of an image is characterized by a feature vector gathering the values of the coarsely quantized grey level histogram followed by the values of the coarsely quantized gradient magnitude histograms in the four directions (respectively vertical, horizontal, right diagonal and left diagonal). These two histograms are computed over the set of pixels existing in an overlapping squared fixed-size (N_w) neighborhood centered around the pixel to be characterized. In our application, this local histogram is respectively quantized with q_l and q_g equidistant binnings for the grey level space and for each of the four gradient magnitude histograms. This simple texture feature extraction step thus yields to a $D = q_l + 4\,q_g$-dimensional feature vector for each pixel. This local feature descriptor turn out to be both discriminant to characterize the different grainy patterns of a SAR image or the different textural patterns specific of another imaging modality. Once this set of feature vector are extracted for each pixel, we reduce the dimensionality of this set of feature vectors to one dimension with a Multidimensional Scaling (MDS) technique [16, 106]. This allows us to project each textured image on a one-dimensional representation or concretely as a new grey-level transformed image. The interest of the MDS over other dimensionality reduction methods lies in the fact that this technique has the particularity of being able to estimate (optimally, in the least-squares sense) an embedding from the set of feature vectors in the high dimensional space (dim=D) such that the distances are faithfully preserved in the low dimensional (dim=1) target space and thus to ensure that two distant pixels (in the transformed image) will necessarily have a grey-level intensity difference proportional to a L_2 (in our application, contrary to MDS, PCA gives bad

results) distance between the two corresponding texture descriptors extracted on the input satellite image. Nevertheless, for computationally reasons, the originally proposed MDS algorithm (called *metric MDS* is not appropriate in our application and more generally for all large scale applications) because this algorithm requires a complexity of $O(N^2)$ (N being the number of pixels). Instead, we have herein used a fast alternative, called FastMap [26] whose main advantage is its linear complexity (at the price of a slightly less good approximation in the least- squares sense).

At this level, it lacks one very important aspect of the common feature space we search to build. Indeed, as already said, the L_2 distance between each textural feature vectors ($D_{<s,t>}$), at locations s and t, in the high dimensional space (dim=D) and the distances between the grey level ($d_{s,t}$) at these same locations in the low dimensional (dim=1) target space is preserved as faithfully as possible and thus the relation $D^{t_1}_{<s,t>} \equiv d^{t_1}_{<s,t>}$ is true for the pre-event satellite image (at time t_1) and for the post-event image satellite (at time t_2) $D^{t_2}_{<s,t>} \equiv d^{t_2}_{<s,t>}$ (for any $< s, t >$). Nevertheless, for two distant pixels s and t belonging to the class label *unchanged area* in satellite image t_1 and t_2, currently, nothing ensures that the grey level at location s in the first (pre-event) projected image and second (post-event) projected image are similar. The MDS technique respects the monotonicity of the grey level order (linear correlation) existing in an image, nevertheless, a nonlinear monotonic scale factor between the two transformed images could however exist. In order to correct this, we resort to a double histogram matching method [98]. More precisely, let us consider the two bi-temporal remote sensing images, y^{t_1} and y^{t_2} acquired before and after a given event and $\hat{y}^{t_1}$ and $\hat{y}^{t_2}$ their MDS projection. $\hat{y}^{t_1}$ is histogram matched to the *after* image $\hat{y}^{t_2}$ to give $\hat{y}^{t_1}_{\bigstar}$ and $\hat{y}^{t_2}$ is then histogram matched to $\hat{y}^{t_1}_{\bigstar}$ in order to finally obtain $\hat{y}^{t_2}_{\bigstar}$ (see Fig. 5.1).

Temporal Differentiation and Binarization: At this level, we can apply any monomodal CD method. In our case, we simply generate a difference image by

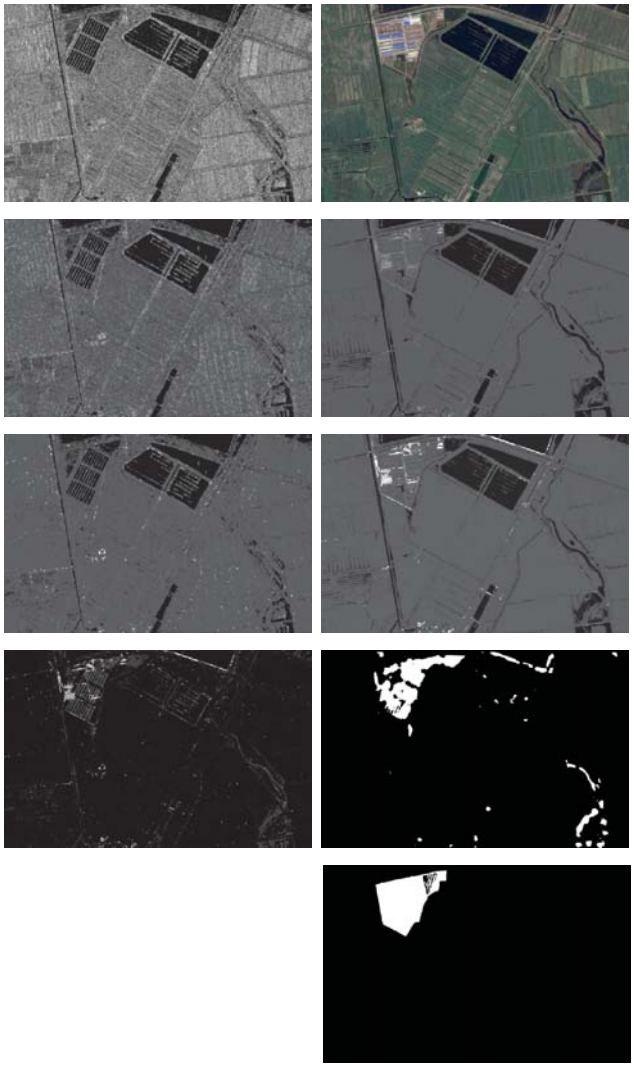

Figure 5.1. First row: SAR/Optical dataset; before and after images; Second row shows the before and after images after MDS projection; Third row represents the result of the double Histogram matching on the images of the second row. Fourth row: difference map; final segmentation result; Fifth row: ground truth.

subtracting $\hat{y}_{\star}^{t_1}$ to $\hat{y}_{\star}^{t_2}$ and taking the absolute value to obtain the difference image y^D. Finally y^D is then segmented into two classes to distinguish changes of interest of the land cover. To this end, in order to achieve more robustness, changes are then identified, from the difference image y^D, by combining the results of $T = 3$

Table 5.1. Accuracy rate of change detection on the fourth heterogeneous datasets obtained by the proposed method and the state-of-the-art *multimodal* **change detectors (supervised and unsupervised) and** *monomodal* **change detectors.**

SAR/Optical Dataset (1)	Accuracy
Proposed method	**0.967**
Liu *et al.* [51]	0.976
PCC [51]	0.821

Optical(NIR band)/Optical Dataset (2)	Accuracy
Proposed method	**0.942**
Zhang *et al.* [126]	0.975
PCC [126]	0.882

SAR/Optical Dataset (3)	Accuracy
Proposed method	**0.878**
Jorge *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

1-look SAR/5-look SAR Dataset (4)	Accuracy
Proposed method	**0.827**
Chatelain *et al.* [14]	0.732
Correlation [14]	0.521
Ratio edge [14]	0.382

different automatic thresholding algorithms ([41,121,124]). In this way, this strategy allows us to synergistically integrate multiple different criteria, for which these binary segmentation algorithms have been designed to be optimal in order to further increase the efficiency of our binarization scheme. In our application, this binary fusion process is simply achieved by using a median filter using a three dimensional window $W \times W \times T$ whose the first two dimensions are spatial and the third dimension indexes the different binary thresholded maps to be fused.

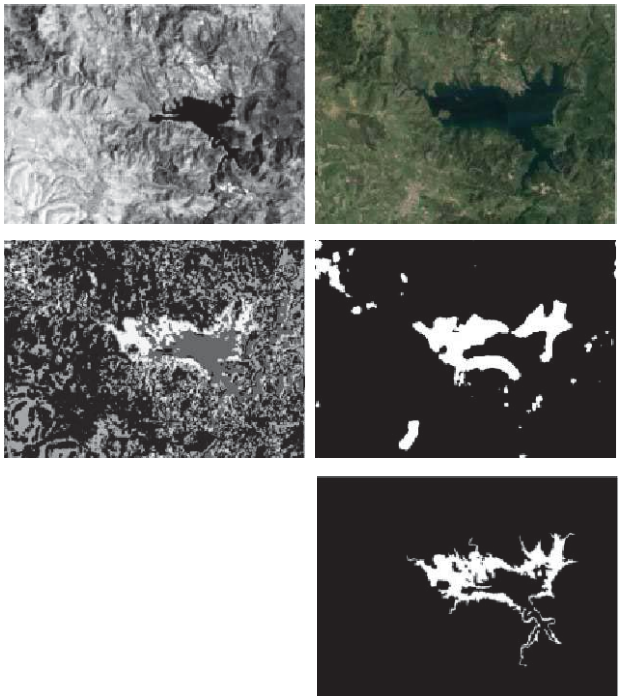

Figure 5.2. Optical(NIR)/Optical dataset. From lex. order; image t_1, t_2; difference map; final segmentation result; ground truth.

5.3 Experimental Results

In order to assess the efficiency of the proposed method to detect different types of land cover changes and to show the strength and the ability of the proposed multimodal CD method to process different remote sensing modalities, we conduct a series of tests on different real multi-source remote sensing imagery data sets. These data sets reflect the three possible change detection conditions in multimodal case. We compare the performance of our method with different state-of-the-art multimodal change detection algorithms recently proposed [14, 51, 89, 126]. The different change masks were provided by a photo interpreter.

The first data set 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.

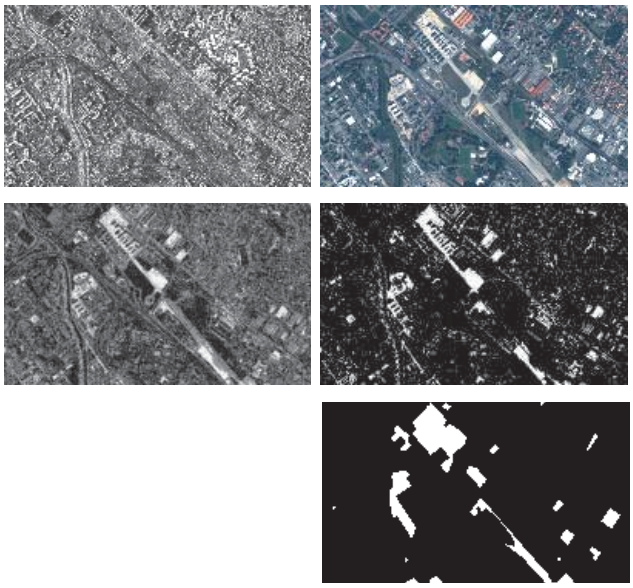

Figure 5.3. TSX/Optical dataset. From lex. order; image t_1, t_2, difference map; final segmentation result; ground truth.

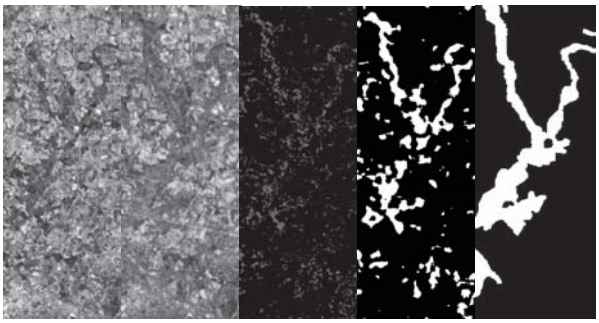

Figure 5.4. SAR 1-look/SAR 5-looks dataset. From lex. order; image t_1, t_2; difference map; final segmentation result; ground truth.

The SAR image is acquired by RADARSAT-2 (Jun. 2008) with spatial resolution of 8m. The optical image comes from Google Earth image (Sept. 2012) and its a combination of aerial photography imaging with a satellite imaging (produced respectively by QuickBird and Landsat-7) with a spatial resolution of 4m. After co-registration, they are of the same pixel-resolution 921×593 pixels.

The second dataset is composed of two heterogeneous optical images. It shows the changes of the Mediterranean in Sardinia area (Italy). This dataset is acquired by different sensor specifications, and consists of one TM (thematic mapper) image and one optical image. The TM image is the near-infrared band of the Landsat-5 (Sept. 1995 with spatial resolution of 30m.). The optical image come from Google Earth (RGB, Jul. 1996, Landsat-5) with the spatial resolution 4m. After co-registration they are of same pixel-resolution 412×300 pixels.

The third heterogeneous data set consists of one optical image and one SAR image. It shows the area of Toulouse (FR), with a size of 4404×2604 pixels. The SAR image is taken by the TerraSAR-X satellite on Feb. 2009 before a building construction. The optical image is captured by the Pleiades satellite on Jul. 2013 after the construction of a building. The optical image have a resolution of 2m. The TSX image was co-registered and re-sampled by [90] to match the optical image resolution.

The fourth multimodal dataset is composed of two heterogeneous SAR images. It shows the area of Gloucester (UK) before and after a flooding event, with a size of 762×292 pixels and with a pixel resolution of 40m. The before and after images are captured by the RADARSAT satellite with different number of looks. The numbers of looks for the before and after SAR image is one-look image (Sept. and Oct. 2000) and five-looks.

In all the experimental results, we have considered $N_w = 7$, $q_g = 10$, $q_l = 40$ and $W = 7$ (see Section 7.2). Table 5.1 summarizes the different change detection accuracy rates obtained by our approach with a comparison with other state of the art approaches. We can see that the different changed-unchanged detection binary map results match fairly the different regions present in the ground truth, and that the most changed regions for the different imagery modalities are well recognized by our strategy (see Figs. 5.1-5.4).

5.4 Conclusion

In this work, the applicability of a new multimodal change detection strategy, in remote sensing, is presented. This one is based on an imaging modality-invariant transformation that projects the two multisensor satellite images to a common feature space in which the bi-temporal images share the same statistical properties and thus on which any simple monomodal change detection methods can be applied. Qualitative and quantitative results show that the proposed method offers a good compromise between simplicity of the implementation and reliability. Indeed, this method consistently performs well on different types of input satellite images and showing different kind of changes.

Acknowledgement

we would like to acknowledge the Computer Research Institute of Montreal **(CRIM)** and the Ministry of Economic Science and Innovation **(MESI)** of the Government of Québec to have supported this work

Chapitre 6

ANOMALY FEATURE LEARNING FOR UNSUPERVISED CHANGE DETECTION IN HETEROGENEOUS IMAGES: A DEEP SPARSE RESIDUAL MODEL

Dans ce chapitre, nous présentons notre article révisé dans la revue *I*EEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, intitulé: **Anomaly Feature Learning For Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model**. Nous exposons ce dernier dans sa langue originale de soumission.

Abstract

In this paper, we propose a novel and simple automatic model based on multimodal anomaly feature learning in a residual space, aiming at solving the binary classification problem of temporal change detection (CD) between pairs of heterogeneous remote sensing images. The model starts by learning from image pairs the *normal* existing patterns in the before and after images to come up with a suitable representation of the normal (non-change) class. To achieve this, we employ a stacked sparse autoencoder trained on a large number of temporal image features (training data) in an unsupervised manner. To classify pixels of new unseen image-pairs, the built anomaly detection model reconstructs the input from its representation in the latent space. First, the probe image (i.e. the bi-temporal heterogeneous image pair as the input request) is encoded in this compact *normal* space from a stacked hidden representation. The reconstruction error is assessed using the *L2* norm in what we

call the residual normal space. In which, the non-change patterns are characterized by small reconstruction errors as normal class while the change patterns are quantified by high reconstruction errors categorizing the abnormal class. The dichotomic (changed/unchanged) classification map is generated in the residual space by clustering the reconstructed errors using a Gaussian mixture model. Experimental results on different real heterogeneous images, reflecting a mixture of imaging and land surface CD conditions, confirm the robustness of the proposed anomaly detection model.

6.1 Introduction

Nowadays, detecting changes between images of the same geographical area over time is still an active topic in remote sensing image processing. A less explored problem is the multimodal change detection (CD) which is a challenging task that can be viewed as the generalization of the classical monomodal CD problem [23,33,36,77,122]. This research area became active with the launch of new satellite generations with different sensor characteristics. Definitely, the exploitation of heterogeneous multimodal data is important to increase the accuracy of any change detection system. The existing mono-modal systems are not usable as-is and need to be adapted to solve the CD problems for environmental monitoring, deforestation, geological resources survey, disaster localization and quantification, and urban planning to name a few.

Multimodal CD [55] is a data analysis procedure seeking directly to locate area of change that may have occurred between two heterogeneous satellite images acquired in the same region of interest at different times. Practical and technical advantages of this recent CD procedure have generated a growing interest, in the remote sensing research community since it should be more robust to natural changes due to environmental variables such as humidity or phenological state, that can be avoided when comparing images coming from different sources (*i.e* multimodal images). Change detection based on multimodal images (Heterogeneous) generally refers to differences

in two imaging modes in which acquired images are represented in two distinct feature spaces that do not share the same statistical properties. It is a non-trivial problem since it is subject to less stringent requirements about the source and characteristics of the acquired data. Hence, leading to radically different image statistics that cannot be compared directly from traditional change detection techniques.

To date, the multimodal CD issue has been addressed by few works, that can be grouped into five categories in which we can find parametric models [14, 65, 88, 90] that use a set of parametric multidimensional distributions (mixture), non-parametric methods [109] which aim to minimize an energy model to satisfy an overdetermined set of constraints, algorithms based on operators using spatial and temporal similarity measures as in [2, 8, 110], projection-based techniques that try to map the two heterogeneous images to a common feature space where traditional monomodal CD can be applied [29,53,111,117], and finally machine learning methods [51,66,126,127].

As the most advanced form of machine learning, deep learning was used for feature-based learning. For instance, a deep autoencoder neural network has been proposed to realize unsupervised feature learning in order to learn discriminative and effective features from a large amount of unlabeled data. The sparse autoencoders have been widely studied for feature-based deep learning methods [3] [116], as it is highly effective for finding high level representations of complex data. In our case, the multimodal change detection problem can be viewed as a binary classification task in which the *change* class or region refers to a set of pixel pairs (or instances), extracted from heterogeneous image pairs, that stand out as being different from all others. Such instances can be seen as anomalies that are indicative of a particular underlying process under the assumption that there are no errors generated from the sensor. Hence, the change class refers, practically, to different semantic regions from the same geographical area that is seen through two different imaging modalities. This anomaly detection problem can be efficiently solved using sparse autoencoder since

it has the appealing ability to uncover potential anomalies in unlabeled data [13].

In this work, we propose a new unsupervised CD model which belongs to the last family of above cited methods. Compared to the state-of-the-art methods, the proposed CD model is defined to be more robust to model the class changes as anomalies, thanks to its flexible learning architecture which is also well adapted to process new un-seen pair of heterogeneous image inputs (as anomalies) in the absence of *annotated* data. Our proposal is to modelize in a residual space the changes as anomalies. More precisely, we propose an unsupervised anomaly-based heterogeneous CD modelling based on learning image features from deep sparse autoencoder neural network as a multimodal feature extractor to gather useful image features from the usual image patterns (non-change or normal class) existing in the before and after multimodal images in the absence of labels. The built anomaly detection model utilizes a reconstruction error vector to perform anomaly detection. To analyze a new unseen image-pairs, the model projects the input into a new latent space from which it attempts to map the projected representation back to reconstructs the input. The residual difference between the original input and the reconstructed one defines our residual space. A Gaussian mixture model is then used to model the extracted features in this space to separate normal from anomalous patterns corresponding, respectively, to non-change and change class labels. The advantage of the proposed CD model lies in its flexibility to process a non-specific source type such multisensor, multisource, or multilooking SAR image pairs, avoiding the drawbacks of parametric models which require knowledge of the conditional distributions; and the disadvantages of supervised machine-learning models that require often labeled and well-balanced training data. The main advantage of our model lies on its ability to learn the underlying latent space.

The rest of this paper is organized as follows: Section 7.2 presents the proposed residual change detection model and its architecture, which allows us to learn and to

reconstruct a suitable representation (feature anomaly space), from which changed and unchanged areas are then identified as normal/abnormal classes. Section 7.3 describes the experimental framework used to evaluate the performance of the proposed CD model, and a set of experimental results compared to the state-of-the-art multimodal change detectors. Section 7.4 concludes the paper.

6.2 Proposed Change Detection Model

Let us assume two multimodal remote sensing images acquired before and after a given event in the same geographical area and also let us consider that the acquired images are co-registered. In order to estimate a binary change detction map which is supposed to represent the difference between the two temporal heterogeneous images, we rely on unsupervised reconstruction machine learning model designed especially to model the change class as anomalies in our change detection problem in order to detect different possible change events as floods, urban growing, etc....

The proposed anomaly-based CD model takes as input a combination of a variety of multimodal remote sensing images, as a combination of two optical images, SAR/optical or optical/SAR images, or SAR images with different number of looks. The pixels in those images cannot be directly compared. The model is composed of two major parts; An unsupervised learning sparse based modelling step, where a training phase is performed to learn a robust deep sparse change detector, and a binary clustering step, where a maximum a posteriori criteria is used for data clustering (see Fig. 6.1). More precisely, in the training phase, the architecture of our CD model is based on stacked sparse autoencoder with a depth of two sparse layers where each single sparse layer has an encoder layer with a corresponding decoder layer. Based on the proposed architecture, our CD model takes as input a temporal *normal* feature space and try to learn an encoder-decoder layers using a layer-wise training technique in which each sparse layer is trained independently in an unsupervised manner. The

internal and optimal values of the deep CD model parameters (prior) are predetermined using a grid search method (see subsection III.C). The temporal *normal* feature space is fed to the first single layer sparse autoencoder which was trained to extract low level feature representations from its hidden layer. The lower level features are then used to train the second sparse autoencoder where high level features are given by its hidden layer (the second layer) of the stacked sparse autoencoder. The encoder layer encodes the input in a compact representation while the decoder layer ensures to predict the encodings in order to reconstruct an estimate of the original input. Once the training phase is accomplished, the built encoder-decoder layers ensures respectively the mapping of new input feature space in a compressed space and then the reconstruction of the original space from this compact representation. The reconstruction error between the input features and their reconstructed versions is then computed using the *L2 norm*. A clustering step is achieved in the residual space to generate, as output, two clusters of data (change versus non-change) related to our bi-temporal change detection problem.

6.2.1 Unsupervised Learning Sparse Model

The anomaly-based CD problem aims at identifying the (usually rare) differences of ground features existing locally between two bi-temporal heterogeneous images, acquired over the same geographical area, with two different imaging modalities (let us assume that the two remote sensing images are co-registered). It may be considered as a binary classification task in which the (small) local spatial changes, over the time, are potential indicatives of somethings that have truly changed in the area of interest and which can thus be identified as anomalies (*i.e.*, different data seen through two different imaging modalities). More precisely, *anomalous patterns* are referred as patterns in the data that do not conform to a well-defined notion of normal behaviour [46]. A common strategy to extract anomalies is to reduce the high

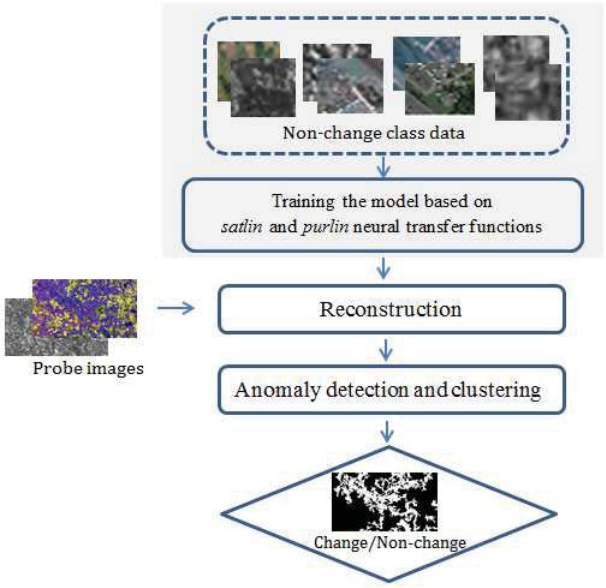

Figure 6.1. Main steps of the proposed residual space-based change detection model

dimensional input space in lower dimensional space and then apply a set of distance metrics within the reduced space in order to identify the anomalies [83].

To this end, supervised classification approaches require labeled and often well-balanced training data or more generally a pre-processing stage such as data augmentation to train a classifier model. In heterogeneous CD problem, especially in remote sensing imagery, training data are generally less available, unlabeled and often highly unbalanced. Besides, data augmentation may be harder since the binary class *change* and *non-change* are highly imbalanced over the whole acquired data.

In our CD problem, it is important to recall that the *changed* regions are smaller than the unchanged regions since a significant event (such as a flooding, earthquake, etc.) occurs rarely and are thus very localized in time and space. Consequently, we have to rely on machine learning based binary classification method in which the training phase is only performed on patterns belonging to the predominant class

(the *non-change majority* class label in our case) while keeping robust to detect the minority class *i.e.*, the rare events belonging to the *change* class as anomalies during the test phase.

Among the existing machine learning-based strategies, the reconstruction-based methods, using sparse autoencoders, seems particularly well adapted to our heterogeneous CD problem. Its main ability is to learn, in the least square sense, a compressed representation minimizing the reconstruction error of the two imaging modalities in the residual space and to estimate within this space the reconstruction error of each bitemporal input patterns from local gray level distribution as a reliable anomaly score. This score can then be exploited to identify the abnormal (rare) patterns caused by a given event (defining the *change* label) and the normal unchanged patterns belonging to the *non-change* class label.

To build our abnormal pattern-based model, we propose to learn a stacked constrained neural network model which can be trained with a layer-wise training procedure [30] in order to find a good representation for the input space [100] [92] and also to better reconstruct the normal patterns based on the learned multimodal imaging representation [22] (see Fig.6.2 and 6.3). More precisely, we propose to use a stacked sparse auto-encoder, which offers an unsupervised reconstruction framework consisting of multiple layers of sparse autoencoders, and which turns out to be robust to discover interesting structures from input image data. This allows us to build a robust anomaly change detection model to identify with a high error the unusual and abnormal features (see Fig.6.3). Let us note that deep learning methods including deep autoencoder have been applied to learn cross-modality and multimodal features. In particular, AECs are able to fuse highly heterogeneous pairs of data types, such as text mixed with images, or audio linked with video, and even combining facial expressions with sound to name a few [83] [27,34,35,38,39,43,47,52,57,84,92,112,114]. Hence, this work defines a novel application of deep networks to learn from het-

erogeneous normal patterns, a common space representation and also an appealing strategy to reconstruct or fuse different imaging modalities within an unsupervised feature-based learning strategy [83] [39] [47].

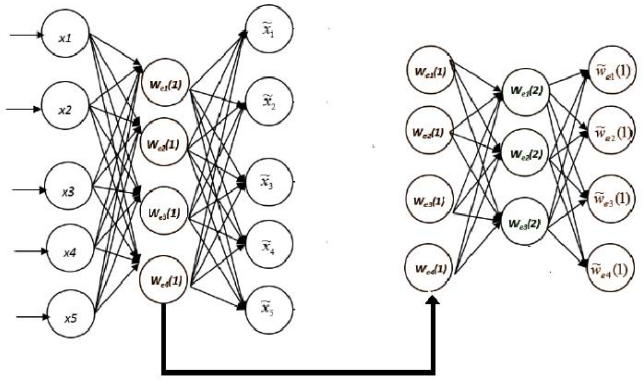

Figure 6.2. Stacked autoencoder neural network composed of two layers of sparse auto-encoders

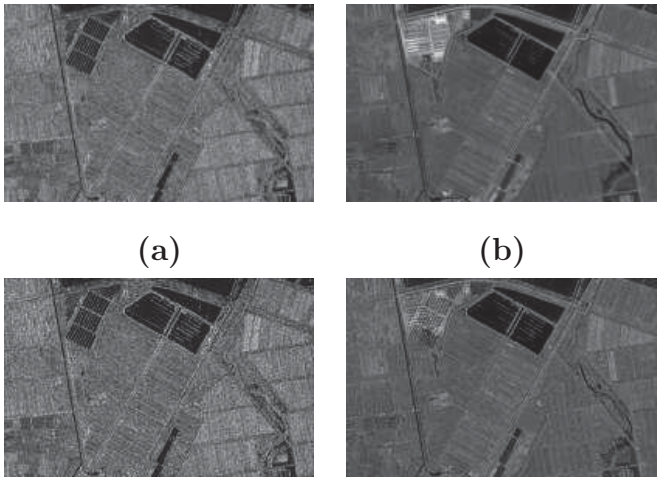

Figure 6.3. Original SAR/optical images (a) and (b); Reconstructed images (c) and (d).

It is important to remember that the intrinsic problems of the standard autoencoder model make it inefficient [79] [101]. Sparse auto-encoder is a constrained model that can learn relatively sparse features by introducing a sparse penalty term inspired by the sparse coding [79] into the autoencoder. Putting constraints on the autoencoder neural network aims to encourage the sparsity of the model [80] [79], and can improve the performance relative to the traditional autoencoders [79] [101]. This can be simply achieved by adding a sparse penalty term to the cost function of the hidden layer to control the number of *active* neurons. Hence, the cost function we used in our case for training the anomaly-based deep sparse model is composed from [101]:

Sparsity Regularization Term

Sparsity regularization tends to create specialized neurons that focus on particular subset from the training data by increasing the number of inactive neurons. The average activation of each hidden neuron $\hat{\rho}_i$ is expected to be close to a small value, and each hidden neuron activation is expected to be close to zero, thus the neurons of the hidden layer become *inactive*. To achieve this, the sparsity term is added to the objective function that penalizes $\hat{\rho}_i$ if it deviates significantly from a predefined small number ρ. The sparsity penalty term Ω_{sparsity} is employed as in [118], and attempts to enforce a constraint on the sparsity of the output from the hidden layer. It is defined by:

$$\Omega_{\text{sparsity}} = \sum_{i=1}^{D} \rho \log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1-\rho)\log\left(\frac{1-\rho}{1-\hat{\rho}_i}\right) \tag{6.1}$$

where $\hat{\rho}_i$ is the average activation value for the i^{th} hidden layer unit and D represents the number of neurons in the hidden layer. The sparsity penalty term constrains the value of $\hat{\rho}_i$ to be close to ρ according to the Kullback-Leibler divergence. This penalty function possesses the property that Kullback-Leibler divergence $\text{KL}(\rho_{\|}\hat{\rho}_i) = 0$ if $\hat{\rho}_i{=}\rho$. Otherwise, it increases monotonically as $\hat{\rho}_i$ diverges from ρ.

L2 Regularization Term

The L2 regularization term Ω_{weights} is added to keep the weight magnitudes small during the feature learning stage in order to prevent over-fitting. It is defined as follows:

$$\Omega_{\text{weights}} = \frac{1}{2}\sum_{l}^{L}\sum_{j}^{N}\sum_{i}^{k}\left(\omega_{ji}^{(l)}\right)^2 \tag{6.2}$$

where $\omega_{ji}^{(l)}$ represents a weight, L is the number of hidden layers, N is the number of observations and k is the number of variables in the training data.

Cost Function

The anomaly change detection model is based on training an unsupervised sparse neural network whose the cost function is an adjusted mean squared error function defined by equation (6.3) [101]. In our work, we propose to use a more robust encoding-decoding neural transfer functions (eq.6.4 and eq.6.5) that better mitigates the convergence problem, and improves the performance of our CD model.

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{k}(x_{kn} - \hat{x}_{kn})^2 + \lambda \cdot \Omega_{\text{weights}} + \beta \cdot \Omega_{\text{sparsity}} \tag{6.3}$$

where x_{kn} is the input vector and $\hat{x}_{kn}$ is an estimate of the input vector x_{kn}. The coefficients λ and β control, respectively, the importance of the regularization and the sparsity terms.

Transfer Functions

To make our change detector more effective for anomaly detection, we make use of the positive saturating linear transfer function for the encoding stage, and the linear transfer function for the decoding stage. Each encoder layer has a corresponding decoder layer:

$$f_{Enc}(z) = \begin{cases} 0, & \text{if} \quad z \leq 0 \\ z, & \text{if} \quad 0 < z < 1 \\ 1, & \text{if} \quad z \geq 1 \end{cases} \tag{6.4}$$

$$f_{Dec}(z) = z \tag{6.5}$$

The encoder maps the input representation x to another encoded representation as follows:

$$z_{enc} = f_{Enc}^{(l)}(W^{(l)}x + b^{(l)}) \tag{6.6}$$

where $W^{(l)}$ is a weight matrix, and $b^{(l)}$ is a bias vector of the encoding layer.

The decoder maps the encoded representation z_{enc} to reconstruct an estimate of the original input representation by:

$$\hat{x} = f_{Dec}^{(l)}(W^{(l)}z_{enc} + b^{(l)}) \tag{6.7}$$

where $W^{(l)}$ is a weight matrix, and $b^{(l)}$ is a bias vector of the decoding layer.

6.2.2 Binary Clustering

In this approach, we have formulated the heterogeneous CD problem into a learning-based reconstruction problem in which the learned constrained stacked sparse model uses its stacked hidden representation to map or reconstruct each new input image pattern. Given a new heterogeneous remote sensing image pair, we have thus to first compute the reconstruction error for each pixel (or for each feature vector centered on this pixel) occurring at the same position in the before and after image pair. The reconstruction error, between the feature vector expressed in the input feature space and the reconstructed space is then measured in the L_2 norm sense and the pixels

belonging to the *change* class label are then simply identified by their high abnormal reconstruction error.

Based on the reconstruction error, the automatic clustering of the residual space, can be performed by a thresholding technique or a k-means based classification strategy ($k = 2$). Another strategy, less sensitive to false alarms or the *a priori* assumption of two spherical class label datasets with the same radius (in the case of the k-means procedure) consists in estimating the parameters of a mixture of two Gaussians in the residual space with the EM algorithm. The MAP rule based on these mixture parameters is used as final binary decision to assign a normal class label to the *non-change* class *and* the abnormal class label to the *change* class. Algorithm 6 shows the predictions of the CD model on the new unseen data.

Step 1:

• $\hat{x} \leftarrow$ reconstruct a new input feature space (test) x using the built deep sparse model with the optimal parameter

foreach $\hat{x_i} \in$ *reconstructed space* $\hat{x}$ **do**

  • $e_i \leftarrow$ compute the reconstruction error between x_i and $\hat{x_i}$ using the L2 norm

end

Step 2:

• Perform a clustering stage on e_i

Algorithme 6: Prediction steps of the CD model.

6.3 Experimental Results

In order to validate and to show the strength of the proposed model to process both different imaging modality cases and change detection conditions along with differ-

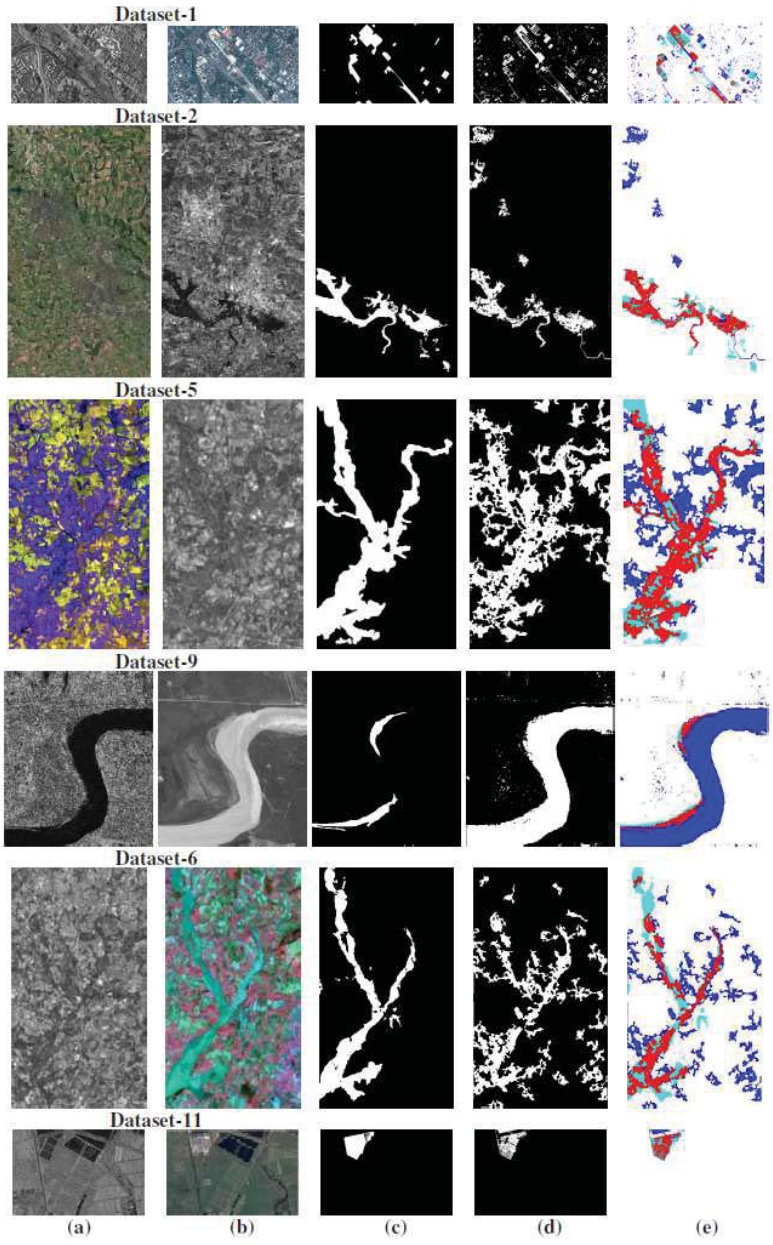

Figure 6.4. Heterogeneous (multisource) Optical/SAR and SAR/Optical datasets: (a-c) image t_1, t_2, ground truth; (d-e) final (changed-unchanged) clustering result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

ent spatial resolutions, we have conducted our study on 11 real heterogeneous image pairs with different kinds of modalities; namely multi-sensor (heterogeneous optical images), multi-source (optical and SAR images) and multi-looking (heterogeneous SAR images) in which the change mask (ground-truth) is provided, for each heterogeneous dataset by a photo-interpreter.

In our application, we use the leave-one-out test scenario to evaluate the performance of the proposed CD model. In this well known procedure, we remove one entire dataset from the eleven heterogeneous datasets and we train the model on the remaining heterogeneous datasets. The output of the trained model is then used to classify the removed dataset. We repeated this process 11 times and at each time we resort to the two heterogeneous images to be our test example.

6.3.1 Heterogeneous Dataset Description

• The first multimodal dataset is a pair of SAR/optical satellite images (Toulouse, France), with size 4404×2604 pixels, before and after construction. The SAR image was taken by the TerraSAR-X satellite (Feb. 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of CNES, Centre National d'Etudes Spatiales) satellite (July 2013). The TSX image was co-registered and re-sampled by [87] with a pixel resolution of 2 meters to match the optical image.

• The second one is a pair of optical/SAR satellite images (Gloucestershire region, in southwest England, near Gloucester), with size 2325×4135 pixels, before and after a flooding taking place in an urban and rural area. The optical image comes from the Quick Bird 02 (QB02) VHR satellite (15 July 2006) and the SAR image was acquired by the TerraSAR-X satellite (July 2007). The TSX image presents a resolution of 7.3 meters and the QB02 image (with resolution of 0.65 meter and 0% cloud cover) was co-registered and re-sampled by [87] to match this resolution.

• The third dataset shows two Heterogeneous optical images acquired in Toulouse (Fr) area by different sensor specifications (size 2000×2000 pixels with a resolution of

Table 6.1. Accuracy rate of change detection on the eleven 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	Accuracy (%)
Proposed method	**0.892**
Prendes *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

Optical/SAR Dataset	Accuracy (%)
Proposed method	**0.961**
Prendes *et al.* [87, 90]	0.918
Prendes *et al.* [88]	0.854
Copulas [65, 88]	0.760
Correlation [65, 88]	0.688
Mutual Inf. [65, 88]	0.768
Pixel Dif. [88, 104]	0.782
Pixel Ratio [88, 104]	0.813

Optical/Optical Dataset	Accuracy (%)
Proposed method	**0.880**
Prendes *et al.* [87, 89]	0.844
Correlation [87, 89]	0.679
Mutual Inf. [87, 89]	0.759
Pixel Dif. [87, 104]	0.708
Pixel Ratio [87, 104]	0.661

SAR 1-look / SAR 5-looks Dataset	Accuracy (%)
Proposed method	**0.814**
Chatelain *et al.* [14]	0.732
Correlation [14]	0.521
Ratio edge [14]	0.382

VHR Optical/SAR Dataset	Accuracy (%)
Proposed method	**0.780**
Gregoire *et al.* [64]	0.70

ERS/Spot Dataset	Accuracy (%)
Proposed method	**0.836**
Liu *et al.* [29]	0.818
Liu *et al.* [29]	0.655

SAR/Optical Dataset	Accuracy (%)
Proposed method	0.767
PCC [51]	0.961
SCNN without pre-training [51]	0.958
SCNN with 1 coupling layer [51]	0.964
SCNN with 2 coupling layer [51]	0.969
SCNN with 3 coupling layer [51]	**0.977**
Zhao *et al.* [127]	0.974

SAR/Optical Dataset	Accuracy (%)
Proposed method	**0.980**
Zhao *et al.* [127]	0.979
Liu *et al.* [51]	0.976
SCNN [127]	0.952
PCC [51]	0.821

Optical(NIR band)/Optical Dataset	Accuracy (%)
Proposed method	0.929
Zhang *et al.* [126]	**0.975**
PCC [126]	0.882

Quickbird/IKONOS Dataset	Accuracy (%)
Proposed method	0.847
Yuqi *et al.* [103]	0.986
Multiscale [103]	**0.991**

Quickbird/IKONOS Dataset	Accuracy (%)
Proposed method	0.817
Yuqi *et al.* [103]	0.959
Multiscale [103]	**0.966**

0.5 meter). 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 (11 July 2013) after the construction of a building. The WorldView2 VHR-image was co-registered by [87] to

Table 6.2. Confusion matrix in terms of number of pixels and percentage for the eleven multimodal datasets *i.e.*, **[TSX/Pleiades] (**4404×2604 **pixels), [QB02/TSX] (**2325×4135 **pixels), [Pleiades/WorldView 2] (**2000×2000 **pixels), [SAR 1-look / SAR 5-looks] (**762×292 **pixels), [Spot VHR/ ERS] (**1318×2359 **pixels), [ERS/Spot] (**330×590 **pixels), [MS (NIR) / MS] (**412×300 **pixels), [QB02 /IKONOS] (**240×240 **pixels), [SAR/Optical] (**291×343 **pixels), [QB02 /IKONOS] (**400×400 **pixels), [SAR/Optical] (**921×593 **pixels).**

Multimodal image pairs	TP	TN	FP	FN
TSX/Pleiades	440211 (48.2%)	9791031 (92.8%)	764001 (7.2%)	472773 (51.8%)
QB02/TSX	419342 (68.0%)	8819894 (98.0%)	177191 (2.0%)	197448 (32.0%)
Pleiades/WorldView 2	339464 (56.0%)	3183160 (93.8%)	210542 (6.2%)	266834 (44.0%)
SAR 1-look/SAR 5-looks	26544 (68.1%)	154679 (84.3%)	28871 (15.7%)	12410 (31.9%)
VHR Spot/ERS	480846 (70.4%)	1946913 (80.2%)	479675 (19.8%)	201728 (29.6%)
ERS/spot	13703 (57.2%)	149187 (87.4%)	21555 (12.6%)	10255 (42.8%)
MS (NIR band) /MS	6353 (83.9%)	108577 (93.6%)	7451 (6.4%)	1219 (16.1%)
Quickbird/IKONOS	4689 (54.3%)	44096 (90.1%)	4863 (9.9%)	3952 (45.7%)
SAR/Optical	2317 (73.4%)	74217 (76.8%)	22440 (23.2%)	839 (26.6%)
QuickBird /IKONOS	13450 (52.2%)	117384 (87.4%)	16876 (12.6%)	12290 (47.8%)
SAR/Optical	14746 (66.3%)	520632 (99.4%)	3286 (0.6%)	7489 (33.7%)

match the Pleiades image.

• The fourth dataset [14] is a pair of SAR/SAR satellite images (Gloucester, U.K.) before and during a flood event caused by intense and prolonged rainfall, overwhelming the drainage capacity, on a urban and agricultural/rural areas, with size 762×292 pixels, acquired by RADARSAT satellite with different number of looks. The number of looks for the before SAR image is 1-look image (Sept. 2000) and the number of looks for the after image is 5-looks (Oct. 2000). These two SAR images have a resolution of about 40 meters.

• The fifth dataset [63, 64] 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 on Sept. 1999 before a flooding event. The SAR image is captured by the European Remote Sensing (ERS) satellite (around Nov. 2000) during the flooding event. The resolution of these two images are about 10 meters [63].

• The sixth dataset consists of one SAR image and one SPOT image with the same size of 330×590 pixels. The ERS image is acquired on November 16, 1999 before the flood in Gloucester U.K, and the optical image combined with 3 bands is acquired on October 21, 2000 during the flood in Gloucester U.K.

•The seventh dataset is composed of two heterogeneous optical images. 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 and one optical image. The TM image is the near-infrared band of the Landsat-5 (Sept. 1995 with spatial resolution of 30 meters). The optical image comes from Google Earth (RGB, Jul. 1996, Landsat-5) with a spatial resolution of 4 meters. After co-registration they are of same pixel-resolution 412×300 pixels.

• The eighth dataset shows two Heterogeneous optical images from another area in the south campus of Hubei province of China, were respectively acquired by the QuickBird satellite in May 2002 and the IKONOS satellite in July 2009, with a size of 240×240 pixels. The images after preprocessing have the same spatial resolution of 3.28 meters.

• The ninth dataset is a pair of SAR/Optical satellite images with a size of 291×343 pixels. The before image is acquired by RADARSAT-2 in June 2008 over the River of China. The optical image comes from Google Earth (September 2010), acquired after a flooding event, and which integrates imagery from both Quickbird US VHR satellite and SPOT5 satellite. After, co-registration, they are of the same spatial resolution of 8 meters.

• The tenth dataset shows two heterogeneous optical images covering the campus of Wuhan University in Hubei province of China. They were respectively acquired by the QuickBird satellite in April 2005 and the IKONOS satellite in July 2009, and correspond to 4-bands (red, green, blue, and NIR band) with a size of 400×400 pixels. The resolution of these images is of 2.44 and 3.28 meters. After re-sampling the after image have the same spatial resolution as the before image 2.44 meters.

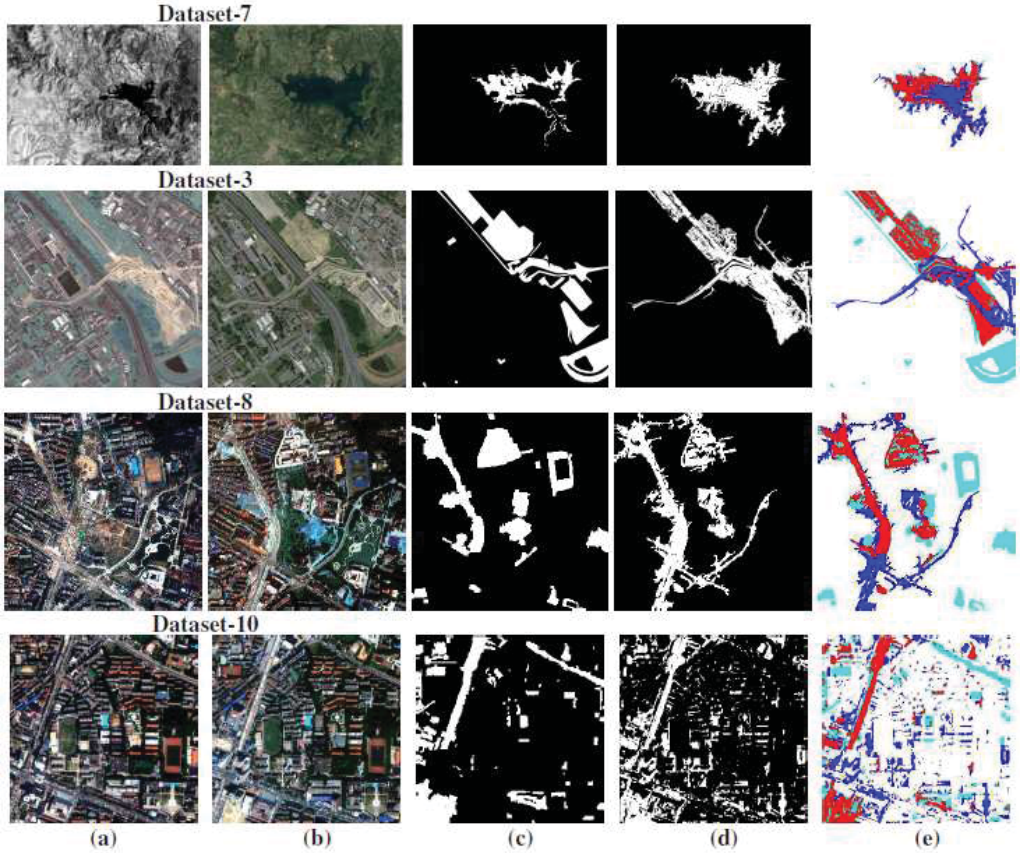

Figure 6.5. Heterogeneous (multisensor) Optical/Optical dataset: (a-c) image t_1, t_2, ground truth; (d-e) final (changed/unchanged) clustering result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

• The eleventh data set 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 meters. The optical image comes from Google Earth image (Sept. 2012) with a spatial resolution of 4 meters [51]. After co-registration, they are of the same pixel-resolution to give a size of 921×593 pixels.

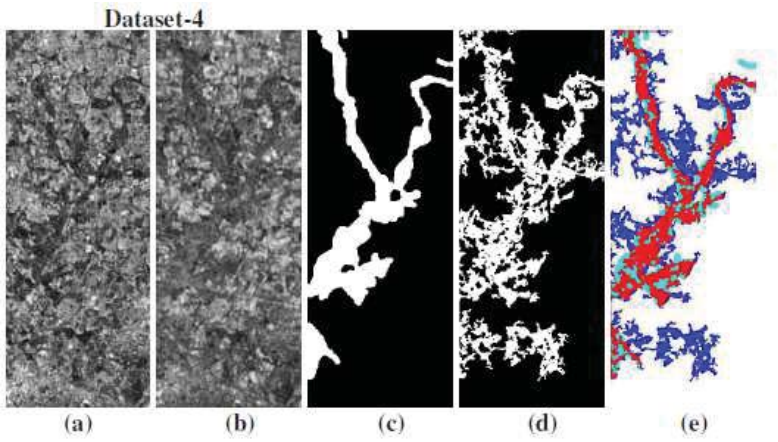

Figure 6.6. Heterogeneous (multilooking) SAR/SAR datasets: (a-c) image t_1, t_2, ground truth; (d-e) final (changed/unchanged) clustering result and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

6.3.2 Results & Evaluation

In our anomaly-based CD problem, we first convert the multi-bands image to a grayscale image, the temporal feature image space is simply done by collecting the local gray level intensities using a squared window of size Sw ($Sw = 9$ in our case).

We have used a learning architecture composed of a stacked sparse auto-encoder and consisting of two layers of sparse auto-encoders, where each encoder layer has a corresponding decoder layer, a deep sparse auto-encoder with a number of hidden layers $L_h = 2$, that takes a bi-temporal feature vector input of dimension $D_{inp} = 162 (= 2 \times 9 \times 9)$. The learned encoder layers compress the input space into a low-dimensional representation, first into a number of dimensions $d_{hl_1} = 80$ and then into a number of dimensions $d_{hl_2} = 40$.

The reconstruction of this compact representation of dimension $d_{hl_2} = 40$ is done by using the two previously learned decoder layers, respectively from $d_{hl_2} = 40$ to $d_{hl_1} = 80$ and from 80 to the original input dimension $\hat{D}_{inp} = 162$. We recall that in

this learning architecture, we use the *satlin* function for the encoding stage and the *purelin* function for the decoding stage.

Our anomaly-based CD model can be optimized *via* a layer-wise training technique [30], using a scaled conjugate gradient descent algorithm [76], by starting to train the first layer to learn to encode the normal representation $D^{(N)}$ to $d^{(hl_1)}$ and to decode $D^{(N)}$ from $d^{(hl_1)}$, and then to train the second layer to learn to encode $d^{(hl_1)}$ to $d^{(hl_2)}$ and to decode $d^{(hl_2)}$ from $d^{(hl_1)}$.

In our application, the coefficients λ and β for the $L2$ regularization and the sparsity regularization terms, were fixed, respectively, to 0.01 and 4.0. The value of the sparsity proportion ρ was set to 0.10 and the maximum number of training epochs for each of the sparse autoencoder architecture was set to 1000 and 400 epochs.

In order to discuss the obtained results, from the conducted experiments, we compare our results to the state-of-the-art methods in terms of classification rate, *i.e.*, the accuracy that measure the percentage of the correct changed and unchanged pixels.

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \tag{6.8}$$

Where TP and TN denote to the number of pixels that are correctly classified, FN and FP denote to the number of misclassified pixels

Table 7.2 summarizes the different change detection accuracy rates obtained by our approach and draws a comparison with both supervised and unsupervised state of the art approaches.

Based on the leave-one-out evaluation strategy, we can notice that the accuracy rate of the proposed method outperforms the most state of the art approaches and remains comparable to the other supervised and unsupervised state of the art methods. The strength of our model is its ability to process a wide variety of satellite imaging modalities, *i.e.*, multi-sources, multi-sensor, and multi-looking SAR images,

Table 6.3. Parameters of the stacked sparse autoencoder.

Parameter name	Min	Step	Factor	Max
Hidden layer 1	80	10	-	120
Hidden layer 2	30	10	-	50
Hidden layer 3	10	5	-	20
Hidden layer 4	3	2	-	7
ρ	0.00625	-	2	0.8
λ	0.0001	-	10	0.1
β	0.5	-	2	8

Table 6.4. The Stacked Sparse Autoencoder Hyper parameters obtained on the subset multimodal dataset with the mean squared reconstruction error (MSE).

Number of layers	ρ	λ	β	Size of first layer	Size of second layer	Size of third layer	Size of fourth layer	MSE
2	0.1	0.01	4	80	40	-	-	0.0640
3	0.05	0.01	2	110	50	10	-	0.1385
4	0.05	0.01	4	100	40	10	3	0.1409

under different resolutions. The method can effectively process images corrupted by different noise types and different degradation levels (see Fig. 6.6 where SAR images are corrupted by different speckle noise levels).

From Table 7.3, we can see also that the changed and un-changed area are well detected and that the different resulting binary maps match fairly the different regions shown in the ground truth for the different satellite imagery sources (see Figs. 6.4-7.2-6.6).

The global accuracy rate obtained by our unsupervised anomaly detection model, over the 11 heterogeneous image pairs, using the leave-one-out evaluation scenario, is 0.863%.

Table 6.5. Average classification accuracy and the Stacked Sparse Autoencoder Hyperparameters used with the first and second hidden layers.

ρ	λ	β	Average accuracy (%)
0.00625	0.01	4	0.579
0.05	0.01	4	0.822
0.4	0.01	4	0.764
0.8	0.01	4	0.715
0.1	0.0001	4	0.801
0.1	0.001	4	0.808
0.1	0.1	4	0.823
0.1	0.01	0.5	0.830
0.1	0.01	1	0.837
0.1	0.01	2	0.829
0.1	0.01	8	0.832

Table 6.6. Impact of the Square Window size on the Average classification accuracy.

Sw	Average accuracy (%)
9	0.863
11	0.849
13	0.838
15	0.844

Step 1:

- Set of hyperparameters from a defined space
- Normal training and validation subsets

Step 2:

foreach *combination of the model parameters* $\in$ *defined space* **do**

- Train the first and second layers of the sparse AEC model using (Eq.3)
- Compute the MSE on the validation subset.

end

- Optimal hyper-parameters outputs with the least squares.

Algorithme 7: Grid search based hyper-parameter optimization of the proposed CD model.

6.3.3 Architecture configuration and Experimental Settings

In all our experiments, we choose the best architecture as the one having the least mean reconstruction error (MSE) on the validation set containing only normal patterns. For parameters settings, we note that our training/validation dataset is a subset of each multimodal pair image and having dimension $d_s = 162$. The dataset is randomly subdivided into two subsets: (2/3) for the training set and (1/3) for the testing set. We inject in our normal training dataset a proportion of 3.0% anomalous (change) patterns to form the final training dataset. We present empirical results produced by our anomaly CD model on this data subset. We use a simple stacked sparse neural network model with normal class. The network parameter settings are described in Table 6.3. In order to fix the neural network architecture and to find optimal hyper-parameters, we rely on a grid search method performed in a defined space with a fixed step/factor and using the following hyper-parameters space: num-

ber of hidden units per layer for the first, second, third and fourth hidden layers, the coefficient of the sparsity term β, the coefficient of the regularization term λ, and the sparsity proportion ρ. Once a layer-wise training strategy was adopted, each layer was trained independently from the others and the parameter values (ρ, λ, β) were varied by exploring different combinations of optimization parameters for each of the four layers with the corresponding number of hidden units. Indeed, we gradually increase the hidden layer number starting by two layers and choose the architecture, giving the best parameter values that minimized the MSE. Algorithm 7 shows the estimation step (with a grid search based optimization technique) of the internal parameters of the stacked sparse neural network reconstruction model.

When the number of hidden-layers was set to 3 and 4, the mean squared error is, respectively, 0.1385% and 0.1409%, which are greater than the MSE value of 0.0640% obtained only with two hidden layers. Therefore, the number of the hidden layers in our anomaly based CD model was set to 2 in our application. Table 6.4 shows the optimal parameters and the MSE obtained by the grid search method for different architectures depth.

6.3.4 Discussion

Before all, it is important to recall that this type of deep autoencoder will necessarily be well adapted to our multimodal CD detection task, since this one has already proven its efficiency to learn and fuse highly heterogeneous pairs of data types in a common space representation [39] [100] [52] [34] [84] and also has proven to be effective in modeling/fusing highly heterogeneous data/sources supported in the multimedia domain (such as words/images [27] [114], speech/images [57] [47] [35], audio/video [39], facial expressions/sound [52] [43], or multimodal DCE/MRI medical images [100], two MRI medical images modalities [112]). In this study (the first study to our knowledge), we confirm the relevance of this type of deep autoencoder in dealing/fusing heterogeneous data (or heterogeneous imaging modalities) used in

remote sensing.

We now discuss the influence of the different parameter settings for our anomaly CD model on 11 benchmark multimodal datasets using the leave-one-out evaluation strategy. To this end, we vary the parameter to be evaluated and fix the others to their optimal values (see Tab. 6.4), and quantify the average accuracy. In our application, the parameter ρ plays a crucial role because it conditions the level of sparsity which may affect considerably our analysis. More precisely, ρ is used to optimize false alarm rates in our unsupervised anomaly CD detection problem and its tuning is based only on normal class images. Indeed, a small ρ induces an over classification of many normal class patterns as anomalous/outliers. In the opposite case, a large ρ discourages normal data patterns from being classified as anomalous/outliers. Thereby, a bad choice of the value of ρ classifies many normal patterns as anomalous and increases the false positive rate or classify many abnormal patterns as normal and increases the false negative rate, which decreases the performance of the anomaly CD model (see Tab. 6.5). Accordingly, the optimal ρ (in our case $\rho = 0.1$) balances both false-positive and false-negative rates (see Tab. 7.3). We can notice that the weight decay λ and the regularization parameter β affect less the behavior of the autoencoder compared to the sparsity parameter ρ (see Tab. 6.5). Also, experiments conducted on different numbers of hidden layers show that augmenting the number of layers does not effectively increase the average classification accuracy. The average classification rate obtained using 3 and 4 hidden layers with a number of nodes set to 10 and to 3 are respectively equal to 0.847%, 0.845% which are lower than our average classification rate 0.863% that corresponds to the optimal number (= 2) of layers. Varying the number of nodes of the hidden layers also does not enhance necessarily the average accuracy. Different combinations were tested giving very close values to the optimal average accuracy which is obtained by 80 nodes for the first and 40 nodes in the second hidden layer. In the same way, the impact of the squared

window size (Sw) is assessed by a comparison study done on the average classification accuracy of the anomaly CD model using different sizes. Table 6.6 demonstrates that the average classification accuracy is not much significantly influenced by the size (Sw). To conclude, the results obtained from different experiments have shown that the choice of the optimization hyper-parameters are a crucial task in the features network setting, particularly the ρ parameter which is the key parameter of the network contrary to the other parameters such as the depth of the network that does not significantly influence the anomaly CD model performances. The main quality of our model is that it achieves a better classification rate accuracy under different change detection conditions reflecting a variety of imaging modalities with different noise types and levels, where the sensitivity of different parameters is analyzed (see Tab. 6.5). This justifies the fact that it can also be less accurate than some specific supervised/unsupervised multimodal CD models, dealing only with a specific type of noise and a specific imaging modalities such as PCC and SCNN methods [51] which also use denoising algorithms to reduce the speckle noise of the SAR images and/or the Gaussian noise of the optical images (particularly when the SAR images are too much corrupted by the multiplicative speckle noise degrading their quality and creating for each texture class, a kind of macro texture with grainy patterns (see dataset-9 Fig.6.4)).

6.4 Conclusion

In this paper, we have proposed a new anomaly-based CD model for heterogeneous remote sensing image pairs. This model exhibits quite interesting properties. First, the proposed model is based on unsupervised training stage in which a stacked multimodal sparse auto-encoder model employing a *satlin* and *purlin* neural transfer functions is trained to learn and infer a suitable latent representation of the normal image patterns existing in the before and after multimodal images. This is done in

order to identify and to disentangle from the normal image patterns (belonging to the *non-change* class label) the change class as unusual from abnormal feature patterns in the residual space; the trained anomaly-based CD model tries to reconstruct the feature space for each new un-seen image pair by encoding and decoding the image pair inputs using its stacked hidden representation. The reconstruction error between the original input feature and its reconstruction is quantified to generate (for each pixel) an anomaly-based error score that highlights the usual and unusual (rare) patterns that belongs to the abnormal class (*change* class label) or to the normal class (*non-change* class label). Finally, a Gaussian mixture model (GMM) assigns a class label to each pixel (change vs non-change) in the MAP sense. The different experimentation conducted on the proposed CD model, in the leave-one-out test scenario, demonstrates its effectiveness in processing new-unseen input heterogeneous image pairs. Besides, the model seems to be flexible enough to process heterogeneous image pairs with both different spatial resolution, covering different heterogeneous CD conditions (as multi-source, multi-sensor, and multi-looking image pairs). It accurately determines different kinds of natural and/or man-made changes (e.g. major urban construction and changes resulting from different types of natural phenomenon).

Acknowledgement

We would like to acknowledge the Computer Research Institute of Montreal **(CRIM)** and the Ministry of Economic Science and Innovation **(MESI)** of the Government of Québec to have supported this work. We would like also to acknowledge all other researchers that made at our disposal the change detection dataset in order to validate the proposed anomaly change detection model.

Chapitre 7

PAIRWISE DESCRIPTORS LEARNING FOR MULTIMODAL CHANGE DETECTION USING PSEUDO-SIAMESE CNN NETWORK MODEL

Dans ce chapitre, nous présentons notre article soumis dans la revue *I*EEE Geoscience and Remote Sensing Letters (GRSL), intitulé: **Pairwise Descriptors Learning For Multimodal Change Detection using Pseudo-Siamese CNN Network Model**. Nous exposons ce dernier dans sa langue originale de soumission.

Abstract

This paper addresses the problematic of detecting changes in bi-temporal heterogeneous remote sensing image pairs. In different disciplines, multimodality is the key solution for performance enhancement in a collaborative sensing context. Particularly, in remote sensing imagery there is still a research gap to fill with the multiplication of sensors, data sharing capabilities, and multi-temporal data availability. This study was aimed to explore to some extent the multimodality in a multi-temporal set-up for a better understanding of the collaborative sensor-wide information completion and error elimination. In this context, we propose a pairwise learning approach consisting on a pseudo-siamese network architecture based on two uncoupled parallel network streams. Each stream represents itself a convolutional neural network (CNN) that encodes each input patch. The overall change detector (CD) model includes a fusion stage that concatenates the two encodings in a single multimodal feature representation which is then reduced to a lower dimension using fully connected layers and finally a loss function based on the binary cross entropy is used as the final layer.

Thanks to the pseudo-siamese pairwise learning architecture the CD model is able to capture the spatial and the temporal dependencies between multimodal input image pairs. The model processes the two multimodal input patches at one-time under different spatial resolutions. The evaluation performances on different real multimodal datasets reflecting a mixture of CD conditions with different spatial resolutions, confirm the effectiveness of the proposed CD architecture.

7.1 Introduction

In remote sensing imagery, change detection is the process of computing differences in a geographical area by analyzing it at different times. Change detection problems can be divided into two main types: the monomodal CD problem assumes that the change area occurred between two/multiple images over time under the assumption that the two/multiple images share the same characteristics *i.e.* acquired by the same satellite sensor with the same specifications. The multimodal CD problem assumes that the bi-temporal images are acquired by different sensors or with the same sensor but with different specifications. Detecting changes between heterogeneous images is a non-trivial problem as it must take into account multiple sources and characteristics of the acquired data. This problem is still less explored, although it has recently generated a growing interest in the remote sensing research community. The technical and practical advantages enable to increase the system performances, and especially to avoid detecting natural changes due to environmental variables such as humidity or phenological state. This challenging task can be viewed as the generalization of the classical monomodal CD problem [109] which is less used as-is for solving the same CD problems.

Nowadays, few research works have been addressed in the multimodal CD issue. Nevertheless, these can be divided into five categories in which we can find parametric models [14, 65, 88, 90], non-parametric methods [109], algorithms based on operators

using spatial and temporal similarity measures as in [2,8,110], projection-based techniques [29,53,111], and machine learning methods [51,126,127].

Deep learning has become a methodology of choice as the most advanced form of machine learning for image classification, object detection, segmentation and other applications. In particular, convolutional neural network (CNN) is a descriptor learning framework with a deep architecture that transforms the input data through many layers to extract high level representations from the inputs. Invariant feature representation learning is a type of descriptor learning framework, which can be build on CNN e.g. siamese-CNN network [125]. The siamese CNN architecture is used for *patch* comparison and refers to two coupled network streams with the same CNN architecture and the same parameters applied to a pair of input data at the same time. In our case, the multimodal CD problem can be viewed as a binary classification task in which the siamese-CNN architecture takes as inputs the two heterogeneous images.

In this work, we are concerned with a heterogeneity problem. We propose a CD model principally designed to deal with different imaging sources with different spatial resolutions and which is well adapted for representing and detecting temporal changes between two heterogeneous remote sensing images. Our CD model learns directly a binary classification function from various types of pair patches coming from different sources, which are processed through two network CNN streams that share the same architecture configuration but with uncoupled weights between them, in order to extract descriptors independently, for each input patch. The final stage of the proposed model consists to combine the two output descriptors from each stream in a single multimodal representation, which is then used to learn the binary classification cost function. The built model ensures the classification of new temporal input images by processing the input patch pairs in parallel using the learned duplicated convolutional streams and the decision network for binary classification.

The rest of this paper is organized as follows: section 7.2 describes the designed

CD model and its architecture to identify change or non-change input pairs as similar/dissimilar classes. Section 7.3 presents the evaluation strategy used to assess the performance of our CD model and the obtained results compared to the state-of-the-art multimodal techniques. Finally, section 7.4 concludes the paper.

7.2 Proposed Change Detection Model

The two/multiple remote sensing input images that correspond to the same geographic area are acquired and co-registered at different time by two/multiple different sensors. Dealing with the characteristics of the different sources of image represents the main challenging issue.

One interesting solution is to design a multimodal CD model with two branches that take as input a pair of images instead of one input, in which the image before and after are fed to two branches allowing us to capture both the spatial and the temporal inter-dependencies. Formally, the task of multimodal change detection can be viewed as a pairwise identification problem, where a pair of non-change/non-change images (samples) are called similar pair, and a pair of non-change/change are called dissimilar or different pair which represents the difference (in the land use) caused by the event and not by the different source of data.

In this case, the pairwise learning approach is more appropriate to verify whether a pair of temporal images corresponds to the similar pair or to the dissimilar pair, i.e. corresponds to the non-change class in the case of similar pair and to the change class in the other case. This can be achieved by training a network based on the similarity of images in order to learn the similarity between pair of images. Among metric learning approaches, siamese network has already been successfully used in several applications [125] such as signature verification, one-shot image recognition, face verification, learning image descriptors, and image ranking to name a few. The siamese architecture consists of two identical subsystems sharing the same set of

parameters and a cost function module to quantify the pairwise relationship. The cost function can be defined via a distance metric or a similarity measure. The goal consists to increase the similarity score or to decrease the distance between similar pairs, and dually, to reduce the similarity score or to increase the distance between two dissimilar image patches.

In our case, siamese network architecture is able to support as input a pair of images. Since the image pair is multimodal, i.e., composed of two different imaging modalities (acquired from different sources), the siamese network architecture is less effective when the weights are shared between the parallel network streams (parallel subsystems) [129]. Ultimately, using a cost function based on a distance metric or a similarity measure to distinguish between similar and dissimilar pair images is less suitable for evaluating similar pair images coming from two different sources due to the fact that there was not a strong enough correlation between heterogeneous similar pair images.

Inspired by the siamese network, we propose an adapted pseudo siamese network model that handles multimodal pair of images.

Pseudo-siamese architecture is a variant closely linked to the basic siamese architecture [125], well adapted to our multimodal CD problem since it is a less restricted network in terms of weights which are not shared between the two network branches (Fig. 7.1). This leads to increase the number of parameters to be adjusted during the training phase, giving a more flexible network than the original siamese network [125].

We recall that training pseudo-siamese network is accomplished using a pairwise learning approach that involves a loss function depending on pairs of input examples. In our application, the pairwise learning task is formalized as classification of temporal multimodal image pairs into two categories change/non-change. More precisely, our pseudo-siamese network based CD model that takes as input a pair of heterogeneous patches, performs both a supervised multimodal dimensionality reduction and a binary classification tasks.

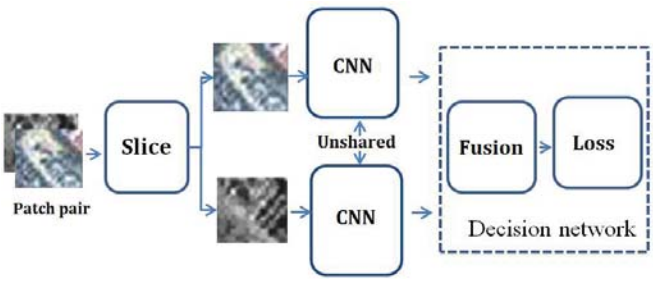

Figure 7.1. Network architecture of the pseudo-siamese based change detector model.

Our multimodal CD model architecture is mainly based on pseudo-siamese network architecture, having two branches that share exactly the same configuration architecture, but with less restrictions on the set of weights, i.e. with uncoupled weights between the two branches. Each branch acts as a feature extractor/descriptor that takes as input one of the two multimodal patches, which can be also a multichannel patches with respect to the number of bands in the input patches.

Our overall CD model includes a decision network, a top network that forms a descriptor within a lower dimensional space and a loss function to learn a decision function from the compact feature space.

The input to the CD model is considered to be a pair of image patches, from which descriptors are first computed independently using two parallel streams and then concatenated with a top network module that decide if the two multimodal input patches present a similar or dissimilar pair corresponding to change vs. non-change class.

Inspired also by the recent advances in neural architectures and deep learning, the structure of our descriptor is represented in terms of a deep convolutional neural network. Indeed, we explore and propose a CNN network architecture that addresses the issue of our multimodal CD problem. Our CNN architecture network is composed of a set of convolutional, ReLU, max-pooling, and fully connected layers, that takes

patches as input and apply on them a three convolutional, max-pooling, ReLU operations and one concatenation operation in the last layer which is a fully connected layer. Our proposed CNN architecture is inspired by the MatchNet network architecture, but with a few layers. The main difference comes from the layers setting. This means that our architecture favors sparse-dense features and does not favor sparse-sparse features produced by the ReLU. Note also that performing a mean-pooling operation instead of max-polling, does not significantly increase the performance of the CD model. The structure of the CNN architecture uses small filters of 5×5 for all convolutional layers, that effectively increases the model performance and reduces the number of filter parameters to be learned. ReLU function is used after the three convolutional layers, which helps to generate sparse features. The last layer is a fully connected layer that acts as a linear dimension reduction layer, and project convolutional features in lower dimensions. The ReLU function is removed after this layer to favorize dense representation. The output of the fully connected layer is the feature representation of the input patch. The spatial padding of convolutional layer input is 2 pixels for the three convolutional layers with 5×5 filter size. The convolution stride is set to 1 pixel. Three max-pooling are performed using 3×3 spatial pooling kernel with a stride of 2. Table 7.1 summarizes the details of our CNN architecture settings.

In the fusing stage, the two output descriptors of each CNN stream are concatenated using a fusion layer that merges the two input features in one single 128-dimensional feature representation, which is then reduced using 2 fully connected (FC) layers but without ReLU function. The first FC layer contains 16 features and the second has 2 outputs corresponding to the change/non-change binary mapping.

In the proposed approach, the CD model takes a single input which is a pair of patches stacked along the depth dimension that requires to be splitted to feed each patch into the corresponding CNN stream. This is ensured using a slice layer that splits the single input into two patches which are in fact the original patches. Fig.7.1

Table 7.1. DETAILS OF THE MODEL ARCHITECTURE FOR CNN.

Name	Type	Input size	Filter number	Filter Size conv	Filter Size pool	Stride	Pad	Stride	ReLU
Conv1/Pool1	conv/max pool	32×32	32	5×5	3×3	1	2	2	Yes
Conv2/Pool2	conv/max pool	$32 \times 16 \times 16$	32	5×5	3×3	1	2	2	Yes
Conv3/Pool3	conv/max pool	$32 \times 8 \times 8$	64	5×5	3×3	1	2	2	Yes
FC1	fully-conn	$64 \times 4 \times 4$	64	N/A	N/A	N/A	N/A	N/A	No

shows the overall pseudo-siamese CD model.

Loss Function

As mentioned earlier, the input of the CD model is considered to be a pair of patches. Learning similarity function between the pair of descriptor outputs is possible, but remains less effective than combining them. Thus, we propose to use the binary cross-entropy loss for training our multimodal pseudo-siamese network.

7.3 Experimental Results

In order to validate and to show the strength of the proposed model, we conduct the experimentations on five realistic multimodal datasets, reflecting different imaging modalities cases under different change detection conditions with different spatial resolutions, namely multi-sensor (heterogeneous optical images) and multi-source (optical and SAR images), showing construction and destruction of buildings in different area. For each multimodal dataset, the change mask (ground-truth) is provided by a photo-interpreter.

In our application, the classification performance of the proposed CD model is assessed using the leave-one-out test procedure. In this well known evaluation strategy, one entire multimodal dataset is removed from the whole training multimodal images, whereas the training phase is performed on the remaining heterogeneous

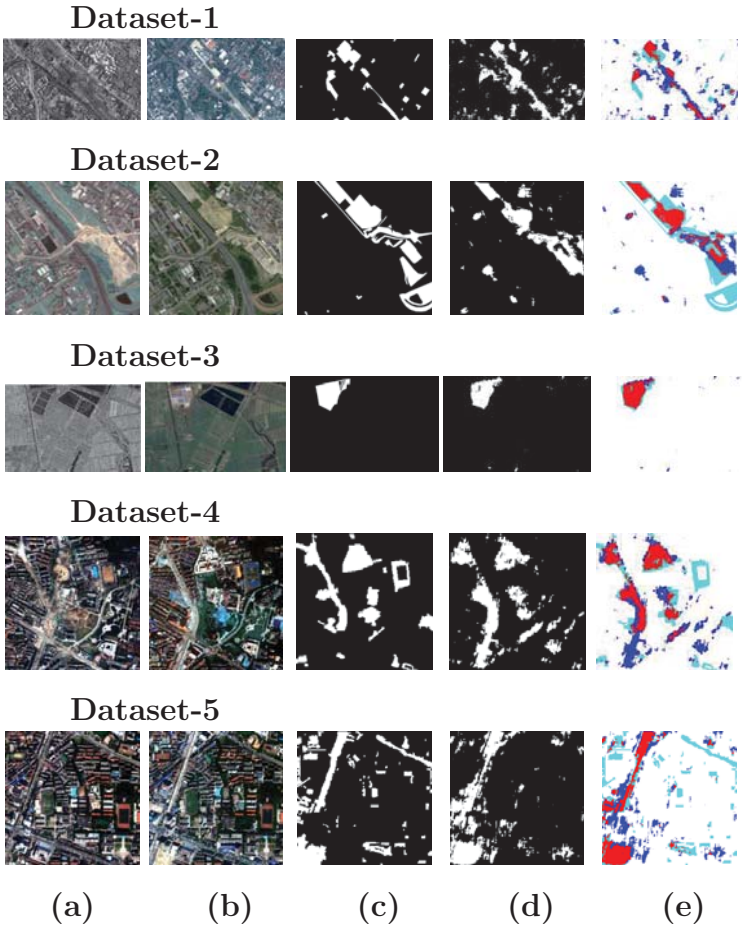

Figure 7.2. Heterogeneous dataset: (a-c) image t_1, t_2, ground truth; (d-e) final (changed/unchanged) binary classification and confusion map (white: TN, red: TP, blue: FP, Cyan: FN) obtained by the proposed approach.

datasets. The built CD model is then evaluated on the removed dataset to generate the binary map. This process is repeated 5 times and at each time two multimodal images were retained as validation data.

7.3.1 Heterogeneous Dataset Description

Table 7.2. Accuracy rate of change detection on the five 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	Accuracy (%)
Proposed method	**0.870**
Prendes *et al.* [89]	0.844
Correlation [89]	0.670
Mutual Inf. [89]	0.580

Optical/Optical Dataset	Accuracy (%)
Proposed method	**0.865**
Prendes *et al.* [87, 89]	0.844
Correlation [87, 89]	0.679
Mutual Inf. [87, 89]	0.759
Pixel Dif. [87]	0.708
Pixel Ratio [87]	0.661

SAR/Optical Dataset	Accuracy (%)
Proposed method	**0.987**
Liu *et al.* [51]	0.976
PCC [51]	0.821

Quickbird/IKONOS Dataset	Accuracy (%)
Proposed method	0.877
Yuqi *et al.* [103]	0.986
Multiscale [103]	**0.991**

Quickbird/IkONOS Dataset	Accuracy (%)
Proposed method	0.837
Yuqi *et al.* [103]	0.959
Multiscale [103]	**0.966**

• The first multimodal dataset is a pair of SAR/optical satellite images (Toulouse, France), with size 4404×2604 pixels, before and after construction. The SAR image was taken by the TerraSAR-X satellite (Feb. 2009) and the optical image by the Pleiades (High-Resolution Optical Imaging Constellation of CNES, Centre National d'Etudes Spatiales) satellite (July 2013). The TSX image was co-registered and re-sampled by [87] with a pixel resolution of 2 meters to match the optical image.

• The second dataset shows two heterogeneous optical images acquired in Toulouse (Fr) area by different sensor specifications (size 2000×2000 pixels with a resolution of 0.5 meter). 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

Table 7.3. Confusion matrix for each of the five multimodal datasets $i.e.$**, [TSX/Pleiades] (**4404×2604 **pixels), [Pleiades/WorldView 2] (**2000×2000 **pixels), [QB02 /IKONOS] (**240×240 **pixels), [QB02 /IKONOS] (**400×400 **pixels).**

Multimodal image pairs	TP	TN	FP	FN
TSX/Pleiades	0.50%	0.90 %	0.10%	0.50%
Pleiades/WorldView 2	0.47%	0.94 %	0.06 %	0.53%
SAR/Optical	0.81%	0.99%	0.01%	0.19%
Quickbird/IKONOS	0.52%	0.94%	0.06%	0.48%
QuickBird /IKONOS	0.49%	0.90%	0.10%	0.51%

satellite from three (Red, Green and Blue) spectral bands (11 July 2013) after the construction of a building. The WorldView2 VHR-image was co-registered by [87] to match the Pleiades image.

• The third multimodal data set 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 meters. The optical image comes from Google Earth image (Sept. 2012) with a spatial resolution of 4 meters. After co-registration, they are of the same pixel-resolution to give a size of 921×593 pixels.

• The fourth dataset shows two heterogeneous optical images from another area in the south campus of Hubei province of China, were respectively acquired by the QuickBird satellite in May 2002 and the IKONOS satellite in July 2009, with a size of 240×240 pixels. The images after preprocessing have the same spatial resolution of 3.28 meters.

• The fifth dataset shows two heterogeneous optical images covering the campus of Wuhan University in Hubei province of China. They were respectively acquired

by the QuickBird satellite in April 2005 and the IKONOS satellite in July 2009, and correspond to 4-bands (red, green, blue, and NIR band) with a size of 400×400 pixels. The resolution of these images is of 2.44 and 3.28 meters. After re-sampling the after image have the same spatial resolution as the before image 2.44 meters.

7.3.2 Training details

In this work, the CD model was trained using the scaled conjugate gradient descent algorithm [76], with a fixed learning rate of 0.001. The momentum and the weight decay were set to 0.9 and 0.004 respectively. The Training was conducted on GPU clusters with batches of 64 pairs of 32×32 patches using balanced classes with leave-one-out evaluation strategy, i.e. the training takes around five rounds. Each time a completely different datasets is used for evaluation.

7.3.3 Evaluation Results and Discussion

We summarize respectively in Table 7.2 and 7.3 the accuracy rate and the confusion matrix obtained using the leave-one-out evaluation strategy. Table 7.2 and 7.3 demonstrates that the proposed CD model outperforms some state-of-the-art methods, it is able to process new probe image pairs under different change detection conditions and without favouring (overfitting) neither of the two classes. The multimodal CD described in this paper turns out to be interesting for multiresolution change detection. Indeed, the CD model learned to fuse features of the two multimodal patches which help to factorize the differences (e.g. land cover changes) and the imaging modalities, but also makes use of standard max-pooling layers to deal with the multi-resolution nature of the data. The model can be also less accurate than some specific CD models that are dedicated to a restricted number of specific imaging modalities.

7.4 Conclusion

In this paper, we presented a CD model based on an uncoupled parallel learning architecture for change detection from bi-temporal multimodal remote sensing images. The model that combines a pseudo-siamese CNN encoder, a fusion layer and a cost classification module, is able to properly capture the spatial and the temporal dependencies between the multimodal input image pairs thanks to its ability to process input data pairs in parallel. Experiments using the leave-one-out test strategy demonstrate that the proposed CD model presents an effective way to process new-unseen heterogeneous input image pairs with different spatial resolutions and under different heterogeneous CD conditions such as multi-source and multi-sensor image pairs.

Acknowledgement

We would like to acknowledge the Computer Research Institute of Montreal **(CRIM)** and the Ministry of Economic Science and Innovation **(MESI)** of the Government of Québec to have supported this work. We would like also to acknowledge all other researchers that made at our disposal the change detection dataset in order to validate the proposed change detection model.

Chapitre 8

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Dans cette thèse, nous avons présenté notre contribution pour résoudre le problème de la détection de changement dans une série d'images satellitaires multimodales. Notre principal objectif est de développer des modèles de détection semi-supervisés ou non supervisés, en utilisant non seulement les techniques de traitement d'images mais aussi en les combinant avec des techniques d'apprentissage automatique.

Nous avons tout d'abord présenté l'état de l'art des différentes méthodes de détection de changement multimodal existantes, que nous avons classées en cinq catégories principales; à savoir les méthodes paramétriques, non paramétriques, celles utilisant des mesures de similarité invariantes, les méthodes utilisant les techniques de projection dans un espace commun et enfin celles exploitant l'apprentissage machine.

Nous avons proposé pour chacune de ces cinq catégories, un modèle original, robuste et largement non supervisé et le plus souvent en complexité linéaire (premier, deuxième, troisième, et quatrième modèle). Plus précisément, nous avons proposé en premier une nouvelle modélisation par paires de pixels intégrée dans une fonction d'énergie dont la solution peut être trouvée en temps linéaire. Ensuite, nous avons développé un nouvel opérateur de détection de changement invariant aux modalités d'imagerie, exprimé par des normes duales, qui détectent, à différentes échelles, les différences en termes de hautes fréquences de chaque région structurelle. Nous avons aussi proposé un modèle paramétrique basé sur un modèle Markovien dont la nouveauté réside dans l'utilisation d'un champ d'observation construit à partir d'une modélisation par paires de pixels. Nous avons développé une méthode basée sur une représentation à l'échelle multidimensionnelle (MDS) qui transforme les images avant

et après dans un espace de caractéristiques commun. Nous avons proposé également un modèle de détection basé sur la détection des changements comme des anomalies en apprenant un modèle qui reconstruit avec erreur des motifs hétérogènes de la classe rare (changée). Dans la même optique, nous avons développé un modèle supervisé qui repose sur une approche d'apprentissage par paires (spatial) de pixels et qui utilise comme architecture deux flux de réseaux convolutifs parallèles et partiellement non-couplés suivis par un réseau de décision.

Dans cette étude, nous avons validé notre modèle sur une base de données constituée de plusieurs paires d'images multimodales et mono-modales de modalité d'imagerie différente (multi-senseur, multi-source, et multi-looking SAR images) représentant des images de zones terrestres de nature variables en relief, en urbanisation, et obtenues sous différentes conditions d'acquisition avec des résolutions et des tailles différentes afin de considérer un maximum de cas et ainsi tester au mieux la flexibilité et précision des différents modèles de détection proposés.

Comparaisons des différents modèles :

Les six modèles de détection de changement présentés dans cette thèse sont de nature assez différentes et possèdent donc des caractéristiques différentes en terme d'adaptabilité ou de flexibilité (aux différents types de modalités d'imagerie possibles associées aux paires d'images satellitaires utilisées), d'efficacité, de robustesse, de généralisation (des résultats sur les images monomodales), de complexité calculatoire et d'améliorations possibles. Le premier et troisième modèle utilisent une modélisation par paires de pixels et sont tous deux des modèles à base d'énergie (dont le but est d'estimer une solution qui correspond au minimum d'une fonction d'énergie ou de coût généralement associé à un critère statistique). Le premier modèle est non paramétrique et au sens des moindres carrés (MC) et le troisième est paramétrique (Markovien) et est presque son équivalent, en terme de modélisation, mais au sens du maximum de vraisemblance (MV). Le critère du MC du premier modèle rend celui-ci

assez robuste au bruit et donne des résultats assez stables pour différents types de modalité d'imageries différentes. Le critère du MV du troisième modèle est peut-être moins stable que le premier modèle (plus sensible au bruit) et donne de meilleurs résultats lorsque les distributions (posées à priori) s'adaptent bien aux données et/ou ne sont pas trop bruitées ou mélangées et inversement donne de résultats moins meilleurs lorsque les lois de distribution sont très mélangées et/ou les données sont fortement bruitées. Le troisième modèle qui fait appel à un optimiseur stochastique est le plus lent des six modèles mais aussi celui qui, grâce à son cadre mathématique très étudié (inférence Bayésienne), à la plus grande faculté d'amélioration possible. Le premier modèle de complexité rendue linéaire, est l'une des techniques les plus rapides des six modèles proposés. Le deuxième modèle (opérateur de gradient textural spatio-temporel invariant par modalité d'imagerie) et le quatrième modèle (projection des deux images satellitaires dans un espace de caractéristiques commun) sont les plus simples, mathématiquement parlant, et aussi les plus locales, en terme de modélisation (modélisation associée à un voisinage). Ce sont les moins complexes, algorithmiquement parlant mais ils ont l'avantage d'offrir un bon compromis en terme de simplicité et coût calculatoire versus efficacité et robustesse obtenue. Du fait de leurs modélisation, ils ont peut-être le désavantage de rendre plus difficile de concevoir des améliorations possibles et de comprendre comment ces modifications permettraient de les améliorer. Les deux derniers modèles reposent sur une modélisation utilisant l'apprentissage machine. Dans ces deux cas, les inconvénients de ces modèles sont connus. Ils résident essentiellement dans la base d'apprentissage de départ que l'on espère être représentatif des différents types de modalités d'imagerie que l'on sera à même de rencontrer. L'inclusion de paires d'images supplémentaires dans cette base d'apprentissage de départ pourrait changer assez sensiblement les résultats et il est malheureusement très difficile de prévoir quelles modifications dans les résultats obtenus seraient influencés par un changement donné dans la base d'apprentissage (i.e., l'inclusion d'une nouvelle paire d'image satellitaire dans la base d'apprentissage).

Perspectives :

Pour la suite de travaux de recherche, nous comptons étudier d'autres contraintes par paires de pixels utilisant des distances basées sur des caractéristiques locales texturales par paires de pixels. Nous envisageons aussi d'autres techniques d'apprentissage profond, de débruitage spatial ou fréquentiel sur la carte de similarité afin d'améliorer le taux de détection tout en essayant de constituer une base de données mono et multimodale, plus importante, que nous espérons rendre publique pour le développement de la recherche dans ce domaine.

De plus, comme nous l'avons déjà dit, les algorithmes de détection de changement multimodal, appliqué dans cette thèse en télédétection, pourraient être aussi presque directement applicables en imagerie médicale pour détecter automatiquement (ou semi-automatiquement) les changements intervenant entre deux radiographies successives issues de deux modalités d'imagerie différentes (par exemple) d'un même patient et pour l'éventuelle détection et quantification (puis le suivi et le traitement) d'une anomalie entre ces deux images médicales multidatés. Nous aimerions explorer ce domaine de recherche médicale et adapter nos algorithmes à cette fin.

De même, il serait intéressant d'étudier nos algorithmes de détection de changement multimodale utilisés dans cette thèse dans le cadre de la détection de changement ou comme cadre mathématique de départ pour la fusion de données issues de caméras de nouvelle génération fusionnant l'infrarouge et l'optique traditionnelle ou encore les caméras du futur combinant l'optique avec les rayons T (utilisés dans les téléphones portables et capables de voir à travers la peau, les vêtements, la fumée et même les murs).

RÉFÉRENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, et S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, November 2012.

[2] V. Alberga. Similarity measures of remotely sensed multi-sensor images for change detection applications. *Remote Sensing*, 1(3):122–143, 2009.

[3] R. M. Aurelio, B. Y-Lan, et L. Yann. Sparse feature learning for deep belief networks. Dans *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 1185–1192, USA, 2007. Curran Associates Inc.

[4] S. Banks. *Signal processing, image processing and pattern recognition*. Prentice Hall, 1990.

[5] J. C. Baritaux et M. Unser. A priori guided reconstruction for fdot using mixed norms. Dans *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 692–695, April 2010.

[6] I. Bayram. Mixed norms with overlapping groups as signal priors. Dans *2011 IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)*, pages 4036–4039, May 2011.

[7] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, B-48:259–302, 1986.

[8] D. Brunner, G. Lemoine, et L. Bruzzone. Earthquake damage assessment of buildings using vhr optical and sar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2403–2420, May 2010.

[9] L. Bruzzone et D. Fernandez-Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Image Processing*, 38(3):1171–1182, 2000.

[10] L. Bruzzone et D.F. Prieto. An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. *IEEE Transactions on Image Processing*, 11(4):452–466, 2002.

[11] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Luis Rojo-Álvarez, et M. Martínez-Ramón. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Trans. Geoscience and Remote Sensing*, 46(6):1822–1835, 2008.

[12] S-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

[13] R. Chalapathy, A. K. Menon, et S. Chawla. Anomaly detection using one-class neural networks. *CoRR*, abs/1802.06360, 2018.

[14] F. Chatelain, J. Y. Tourneret, et J. Inglada. Change detection in multisensor sar images using bivariate gamma distributions. *IEEE Transactions on Image Processing*, 17(3):249–258, March 2008.

[15] X. Chen, J. Li, Y. Zhang, et L. Tao. Change Detection with Multi-Source Defective Remote Sensing Images Based on Evidential Fusion. *ISPRS Annals*

of Photogrammetry, Remote Sensing and Spatial Information Sciences, pages 125–132, Juin 2016.

[16] T.F. Cox et M.A.A. Cox. Multidimensional scaling. *Chapman & Hall, London*, 1994.

[17] H. Dankert, J. Horstmann, S. Lehner, et W. Rosenthal. Detection of wave groups in sar images and radar image sequences. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1437–1446, June 2003.

[18] A.P. Dempster, N.M. Laird, et D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society*, pages 1–38, 1976.

[19] F. Destrempes et M. Mignotte. A statistical model for contours in images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(5):626–638, 2004.

[20] F. Destrempes et M. Mignotte. Localization of shapes using statistical models and stochastic optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(9):1603–1615,, 2007.

[21] T.G. Dietterich. Ensemble methods in machine learning. Dans Lecture Notes In Computer Science, editeur, *Proceedings of the First International Workshop on Multiple Classifier Systems, LNCS, Multiple Classifier Systems*, volume 1857, pages 1–15. Springer, 2000.

[22] A. Droniou, S. Ivaldi, et O. Sigaud. Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems*, 71:83 – 98, 2015. Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence.

[23] P. Du, S. Liu, P. Gamba, K. Tan, et J. Xia. Fusion of difference images for change detection over urban areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1076–1086, Aug 2012.

[24] P. Du, S. Liu, J. Xia, et Y. Zhao. Information fusion techniques for change detection from multi-temporal remote sensing images. *Inf. Fusion*, 14(1):19–27, Janvier 2013.

[25] M. T. Eismann, J. Meola, et R. C. Hardie. Hyperspectral change detection in the presenceof diurnal and seasonal variations. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1):237–249, January 2008.

[26] C. Faloutsos et K.-I. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. Dans *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 163–174, San Jose, California, june 1995.

[27] F. Fangxiang, W. Xiaojie, et L. Ruifan. Cross-modal retrieval with correspondence autoencoder. Dans *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 7–16, New York, NY, USA, 2014. ACM.

[28] H. Fu, M. K. Ng, M. Nikolova, et J. L. Barlow. Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. *SIAM Journal on Scientific Computing*, 27(6):1881–1902, 2006.

[29] Z. g. Liu, G. Mercier, J. Dezert, et Q. Pan. Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning. *IEEE Geoscience and Remote Sensing Letters*, 11(1):168–172, January 2014.

[30] H. Geoffrey, S. Osindero, et T. Yee-Whye. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, Juillet 2006.

[31] M. Gong, J. Zhao, J. Liu, Q. Miao, et L. Jiao. Change detection in synthetic aperture radar images based on deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 27:125–138, 2016.

[32] M. Gong, Z. Zhou, et J. Ma. Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering. *IEEE Transactions on Image Processing*, 21(4):2141–2151, 2012.

[33] R. Hedjam, M. Kalacska, M. Mignotte, H. Ziaei Nafchi, et M. Cheriet. Iterative classifiers combination model for change detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):6997–7008, December 2016.

[34] C. Hong, J. Yu, J. Wan, D. Tao, et M. Wang. Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing*, 24(12):5659–5670, Dec 2015.

[35] W. N. Hsu, Y. Zhang, et J. R. Glass. Learning latent representations for speech generation and transformation. *CoRR*, abs/1704.04222, 2017.

[36] X. Huang, L. Zhang, et T. Zhu. Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):105–115, Jan 2014.

[37] Y. Huang, W. Zhang, et J. Wang. 1 and infinite norm support vector machine. Dans *2012 IEEE International Conference on Information Science and Technology*, pages 476–484, March 2012.

[38] N. Jaques, S. Taylor, A. Sano, et R. Picard. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better

mood prediction. Dans *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 202–208, Oct 2017.

[39] N. Jiquan, K. Aditya, K. Mingyu, N. Juhan, L. Honglak, et Y. Andrew. Multimodal deep learning. Dans *ICML*, pages 689–696. Omnipress, 2011.

[40] P.M. Jodoin et M. Mignotte. Markovian segmentation and parameter estimation on graphics hardware. *Journal of Electronic Imaging*, 15:033015–1–15, 2006.

[41] J.N. Kapur, P.K. Sahoo, et A.K.C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3):273–285, 1985.

[42] A. Khlif et M. Mignotte. Segmentation data visualizing and clustering. *Multimedia Tools and Applications,*, 76(1):1531–1552, 2017.

[43] Y. Kim, H. Lee, et E. M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. Dans *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691, May 2013.

[44] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303 – 324, 2009.

[45] D. Lahat, T. Adali, et C. Jutten. Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. *Proceedings of the IEEE*, 103(9):1449–1477, Aug 2015.

[46] R. Laxhammar et G. Falkman. Online learning and sequential anomaly detection in trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1158–1173, June 2014.

[47] K. Leidal, D. Harwath, et J. R. Glass. Learning modality-invariant representations for speech and images. *CoRR*, abs/1712.03897, 2017.

[48] H. Li, M. Gong, et J. Liu. A local statistical fuzzy active contour model for change detection. *IEEE Geoscience and Remote Sensing Letters*, 12:582–586, 2015.

[49] Y. Li, M. Gong, L. Jiao, L. Li, et R. Stolkin. Change-detection map learning using matching pursuit. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4712–4723, August 2015.

[50] J. Liu, M. Gong, Q. Miao, L. Su, et H. Li. Change detection in synthetic aperture radar images based on unsupervised artificial immune systems. *Appl. Soft Comput.*, 34:151–163, 2015.

[51] J. Liu, M. Gong, K. Qin, et P. Zhang. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE Transactions Neural Netw. Learn. Syst.*, 2018, in press.

[52] Wei Liu, Wei-Long Zheng, et Bao-Liang Lu. Multimodal emotion recognition using multimodal deep learning, 2016.

[53] Z. Liu, G. Li, G. Mercier, Y. He, et Q. Pan. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. *IEEE Transactions on Image Processing*, 27(4):1822–1834, April 2018.

[54] S. P. Lloyd. Least squares quantization in PCM. 28(2):129–136, 1982.

[55] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, et Q. Du. Multi-modal change detection, application to the detection of flooded areas: Outcome of

the 2009-2010 data fusion contest. *IEEE J. Sel. Topics Appl. Earth Observ.*, 5(1):331–342, Feb. 2012.

[56] J. Lu, J. Li, G.Chen, L. Zhao, B. Xiong, et G. Kuang. Improving pixel-based change detection accuracy using an object-based approach in multitemporal sar flood images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8:3486–3496, 2015.

[57] T. Lu, M. Liou, Y. Yuan, L. Hung, et L. Lin-Shan. Semantic retrieval of personal photos using a deep autoencoder fusing visual features with speech annotations represented as word/paragraph vectors. Dans *INTERSPEECH*, pages 140–144. ISCA, 2015.

[58] L. T. Luppino, S. N. Anfinsen, G. Moser, R. Jenssen, F. M. Bianchi, S. B. Serpico, et G. Mercier. A clustering approach to heterogeneous change detection. Dans *Image Analysis - 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12-14, 2017, Proceedings, Part II*, pages 181–192, 2017.

[59] J. Ma, M. Gong, et Z. Zhou. Wavelet fusion on ratio images for change detection in sar images. *IEEE Geoscience and Remote Sensing Letters*, 9:1122–1126, 2012.

[60] L. Mascolo, J. M. Lopez-Sanchez, F. V.-Guijalba, F. Nunziata, M. Migliaccio, et G. Mazzarella. A complete procedure for crop phenology estimation with polsar data based on the complex wishart classifier. *IEEE Trans. Geoscience and Remote Sensing*, 54(11):6505–6515, 2016.

[61] P. Masson et W. Pieczynski. SEM algorithm and unsupervised statistical segmentation of satellite images. *IEEE Trans. on Geoscience and Remote Sensing*, 31(3):618–633, 1993.

[62] F. Melgani et Y. Bazi. Robust unsupervised change detection with markov random fields. Dans *IEEE International Geoscience & Remote Sensing Symposium, IGARSS 2006, July 31 - August 4, 2006, Denver, Colorado, USA, Proceedings*, pages 208–211, 2006.

[63] G. Mercier et J. Inglada. Change detection with misregistration errors and heterogeneous data through the Orfeo Toolbox. Rapport technique, Lab-STICC(TB) - Laboratoire en sciences et technologies de l'information, de la communication et de la connaissance (UMR CNRS 6285 - Télécom Bretagne - Université de Bretagne Occidentale - Université de Bretagne Sud - ENSTA Bretagne - Ecole Nationale d'ingénieurs de Brest), CNES - Centre national d'études spatiales (.), 2008.

[64] G. Mercier, G. Moser, et S. Serpico. Conditional copula for change detection on heterogeneous sar data. Dans *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 2394–2397, July 2007.

[65] G. Mercier, G. Moser, et S. Serpico. Conditional copulas for change detection in heterogeneous remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1428–1441, May 2008.

[66] N. Merkle, P. Fischer S. Auer, et R. Muller. On the possibility of conditional adversarial networks for multi-sensor image matching. Dans *Proceedings of IGARSS 2017, IGARSS 2017*, pages 1–4, Fort Worth, Texas, USA, July 2017.

[67] M. Mignotte. Fusion of regularization terms for image restoration. *Journal of Electronic Imaging*, 19(3):333004–, July-September 2010.

[68] M. Mignotte. Mds-based multiresolution nonlinear dimensionality reduction

model for color image segmentation. *IEEE Trans. on on Neural Networks*, 22(3):447–460, March 2011.

[69] M. Mignotte. A bi-criteria optimization approach based dimensionality reduction model for the color display of hyperspectral images. *IEEE Trans. on Geoscience and Remote Sensing*, 50(2):501–513, January 2012.

[70] M. Mignotte. An energy based model for the image edge histogram specification problem. *IEEE Trans. on on Image Processing*, 21(1):379–386, January 2012.

[71] M. Mignotte. Mds-based segmentation model for the fusion of contour and texture cues in natural images. *Computer Vision and Image Understanding*, 116(9):981–990, 2012.

[72] M. Mignotte. Non-local pairwise energy based model for the hdr image compression problem. *Journal of Electronic Imaging*, 21(1), January-March 2012.

[73] M. Mignotte, C. Collet, P. Pérez, et P. Bouthemy. Sonar image segmentation using an unsupervised hierarchical MRF model. *IEEE Trans. on Image Processing*, 9(7):1216–1231, 2000.

[74] M. Mignotte, J. Meunier, et J.-P. Soucy. DCT-based complexity regularization for em tomographic reconstruction. *IEEE trans. on Biomedical Engineering*, 55(2):801–805, 2008. tomographic reconstruction.

[75] A. Moevus, M. Mignotte, J.A. de Guise, et J. Meunier. A perceptual map for gait symmetry quantification and pathology detection. *BioMedical Engineering OnLine (BMEO)*, 14(1):99, 2015.

[76] M. F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525 – 533, 1993.

[77] A. A. Nielsen, K. Conradsen, et H. Skriver. Change detection in full and dual polarization, single- and multi-frequency sar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(8):4041–4048, aug 2015.

[78] E. J. Nyström. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica*, 54(1):185–204, 1930.

[79] B. Olshausen et D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[80] B. A. Olshausen et D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

[81] T. Ooura. General purpose fft (fast fourier/cosine/sine transform) package. http://momonga.t.u-tokyo.ac.jp/ ooura/fft.html.

[82] K. Ozkan et E. Seke. Combination of l1 and l2 norms for image deconvolution problems. Dans *2011 7th International Conference on Electrical and Electronics Engineering (ELECO)*, pages II–148–II–151, December 2011.

[83] D. Park, Y. Hoshi, et C. C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, July 2018.

[84] W. Pengcheng, C.H. Steven, X. Hao, Z. Peilin, W. Dayong, et M. Chunyan. Online multimodal deep similarity learning with application to image retrieval. Dans *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 153–162, New York, NY, USA, 2013. ACM.

[85] P. Pérez, A. Blake, et M. Gangnet. JetStream: probabilistic contour extraction with particles. Dans *Proc. IEEE Int Conf. Computer Vision, ICCV'01*, Vancouver, Canada, July 2001.

[86] W. Pieczynski. Convergence of the iterative conditional estimation and application to mixture proportion identification. Dans *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 49–53, Aug 2007.

[87] J. Prendes. *New statistical modeling of multi-sensor images with application to change detection*. PhD thesis, Toulouse, 2015.

[88] J. Prendes, M. Chabert, F. Pascal, A. Giros, et J. Tourneret. A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors. *IEEE Transactions on Image Processing*, 24(3):799–812, March 2015.

[89] J. Prendes, M. Chabert, F. Pascal, A. Giros, et J. Tourneret. Performance assessment of a recent change detection method for homogeneous and heterogeneous images. *Revue Française de Photogrammétrie et de Télédétection*, 209:23–29, 2015.

[90] J. Prendes, M. Chabert, F. Pascal, A. Giros, et J. Y. Tourneret. Change detection for optical and radar images using a Bayesian nonparametric model coupled with a Markov random field. Dans *Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP'15*, Brisbane, Australia, April 2015.

[91] J. Prewitt et M.L. Mendelsohn. The analysis of cell images. *Annals of the New York Academy of Sciences*, 128(3):1035–1053, 1996.

[92] Y. Qi, Y. Wang, X. Zheng, et Z. Wu. Robust feature learning by stacked autoencoder with maximum correntropy criterion. Dans *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6716–6720, May 2014.

[93] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[94] F. Salzenstein et W. Pieczynski. Parameter estimation in hidden fuzzy markov random fields and image segmentation. *CVGIP: Graphical Model and Image Processing*, 59(4):205–220, 1997.

[95] A. Schaum. Local covariance equalization of hyperspectral imagery: advantages and limitations for target detection. Dans *2005 IEEE Aerospace Conference*, pages 2001–2011, March 2005.

[96] J. Sexton, X. Song, M. Feng, P. Noojipady, A. Anand, C. Huang, D. Kim, K. Collins, S. Channan, C. DiMiceli, et J. Townshend. Global 30-m resolution continuous fields of tree cover: Landsat-based rescaling of modis vegetation continuous fields with lidar-based estimates of error. *Int. Journal of Digital Earth*, 6(5):427–448, 2013.

[97] A.G. Shanbhag. Utilization of information measure as a means of image thresholding. *CVGIP: Graphical Models and Image Processing*, 56(5):414–419, 1994.

[98] D. Shapira, S. Avidan, et Y. Hel-Or. Multiple histogram matching. Dans *IEEE International Conference on Image Processing, ICIP'13*, pages 2269–2273, September 2013.

[99] A.J.C. Sharkey. *Combining artificial neural nets ensemble and modular multi-net systems*. Springer-Verlag, New York, Inc., ISBN:185233004X, 1999.

[100] H. C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, et M. O. Leach. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1930–1943, Aug 2013.

[101] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, et X. Chen. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, 89:171 – 178, 2016.

[102] B. Tan, J. T. Morisette, R. E. Wolfe, F. Gao, G. A. Ederer, J. Nightingale, et J. A. Pedelty. An enhanced timesat algorithm for estimating vegetation phenology metrics from modis data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(2):361–371, June 2011.

[103] Y. Tang et L. Zhang. Urban change analysis with multi-sensor multispectral imagery. *Remote Sensing*, 9(3), 2017.

[104] OTB Development Team. The orfeo toolbox software guide. Available at http://orfeo-toolbox.org/., 2014.

[105] J. Tian, A. A. Nielsen, et P. Reinartz. Improving change detection in forest areas based on stereo panchromatic imagery using kernel MNF. *IEEE Transactions on Geoscience and Remote Sensing*, 52(11):7130–7139, November 2014.

[106] W.S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[107] R. Touati et M. Mignotte. MDS-based multi-axial dimensionality reduction model for human action recognition. Dans *Eleventh conference on Computer and Robot Vision, CRV'2014*, pages 262–267, Montréal, Quebec, Canada, May 2014.

[108] R. Touati et M. Mignotte. A multidimensional scaling optimization and fusion approach for the unsupervised change detection problem in remote sensing images. Dans *6th IEEE International Conference on Image Processing Theory, Tools and Applications, IPTA'16*, Oulu, Finland, December 2016.

[109] R. Touati et M. Mignotte. An energy-based model encoding non-local pairwise pixel interactions for multi-sensor change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1), January 2018.

[110] R. Touati, M. Mignotte, et M. Dahmane. A new change detector in heterogeneous remote sensing imagery. Dans *7th IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA 2017)*, Montreal, Canada, Qc, December 2017.

[111] R. Touati, M. Mignotte, et M. Dahmane. Change detection in heterogeneous remote sensing images based on an imaging modality-invariant mds representation. Dans *25th IEEE International Conference on Image Processing (ICIP'18)*, Athens, Greece, October 2018.

[112] G. van Tulder et M. de Bruijne. Learning cross-modality representations from multi-modal images. *IEEE Transactions on Medical Imaging*, pages 1–1, 2018.

[113] M. Volpi, F. de Morsier, G. Camps-Valls, M. Kanevski, et D. Tuia. Multi-sensor change detection based on nonlinear canonical correlations. Dans *2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS*, pages 1944–1947, July 2013.

[114] W. Wei, O. Beng Chin, Y. Xiaoyan, Z. Dongxiang, et Z. Yueting. Effective multi-modal retrieval based on stacked auto-encoders. *Proc. VLDB Endow.*, 7(8):649–660, Avril 2014.

[115] N. Widynski et M. Mignotte. A multiscale particle filter framework for contour detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(10):1922–1935, 2014.

[116] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, et S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, June 2010.

[117] C. Wu, B. Du, et L. Zhang. Slow feature analysis for change detection in multispectral imagery. *IEEE Trans. Geoscience and Remote Sensing*, 52(5):2858–2874, 2014.

[118] J. Xie, L. Xu, et E. Chen. Image denoising and inpainting with deep neural networks. Dans *In NIPS*, 2012.

[119] B. Xiong, Q. Chen, Y. Jiang, et G. Kuang. A threshold selection method using two SAR change detection measures based on the markov random field model. *IEEE Geoscience and Remote Sensing Letters*, 9:287–291, 2012.

[120] M. Xu, C. Cao, H. Zhang, Y. Xue, Y. Li, J. Guo, C. Chang, Q. He, M. Gao, et X. Li. Change detection of the tangjiashan barrier lake based on multi-source remote sensing data. Dans *2009 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages IV–303–IV–306, July 2009.

[121] J.C. Yen, F.J. Chang, et S. Chang. A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, 4(3):370–378, 1995.

[122] O. Yousif et Y. Ban. Improving sar-based urban change detection by combining map-mrf classifier and nonlocal means similarity weights. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(10):4288–4300, Oct 2014.

[123] Y.L. Yu. Arithmetic duality for norms. 2012.

[124] G.W. Zack, W.E.Rogers, et S.A. Latt. Automatic Measurement of Sister Chromatid Exchange Frequency. *Journal of Histochemistry and Cytochemistry*, 25(7):741–753, 1977.

[125] S. Zagoruyko et N. Komodakis. Learning to compare image patches via convolutional neural networks. *CoRR*, abs/1504.03641, 2015.

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

[127] W. Zhao, Z. Wang, M. Gong, et J. Liu. Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7066–7080, Dec 2017.

[128] J. Zhou, C. Kwan, B. Ayhan, et M. T. Eismann. A novel cluster kernel rx algorithm for anomaly and change detection using hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(11):6497–6504, November 2016.

[129] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 784–788, May 2018.