Selection of a Similarity Measure Combination for a Wide Range of Multimodal Image Registration Cases

Mikhail L. Uss, Benoit Vozel, Sergey K. Abramov, Kacem Chehdi

▶ **To cite this version:**

Mikhail L. Uss, Benoit Vozel, Sergey K. Abramov, Kacem Chehdi. Selection of a Similarity Measure Combination for a Wide Range of Multimodal Image Registration Cases. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59 (1), pp.60-75. 10.1109/TGRS.2020.2992597 . hal-02948495

HAL Id: hal-02948495

https://univ-rennes.hal.science/hal-02948495

Submitted on 25 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Selection of a Similarity Measure Combination for a Wide Range of Multimodal Image Registration Cases

Mikhail L. Uss, Benoit Vozel, Sergey K. Abramov, and Kacem Chehdi

***Abstract*— Many similarity measures (SMs) were proposed to measure the similarity between multimodal remote sensing (RS) images. Each SM is efficient to a different degree in different registration cases (we consider visible-to-infrared, visible-to-radar, visible-to-digital elevation model (DEM), and radar-to-DEM ones), but no SM was shown to outperform all other SMs in all cases. In this article, we investigate the possibility of deriving a more powerful SM by combining two or more existing SMs. This combined SM relies on a binary linear support vector machine (SVM) classifier trained using real RS images. In the general registration case, we order SMs according to their impact on the combined SM performance. The three most important SMs include two structural SMs based on modality independent neighborhood descriptor (MIND) and scale-invariant feature transform-octave (SIFT-OCT) descriptors and one area-based logarithmic likelihood ratio (logLR) SM: the former ones are more robust to structural changes of image intensity between registered modes, the latter one is to image noise. Importantly, we demonstrate that a single combined SM can be applied in the general case as well as in each particular considered registration case. As compared to existing multimodal SMs, the proposed combined SM [based on five existing SMs, namely, MIND, logLR, SIFT-OCT, phase correlation (PC), histogram of orientated phase congruency (HOPC)] increases the area under the curve (AUC) by from 1% to 21%. From a practical point of view, we demonstrate that complex multimodal image pairs can be successfully registered with the proposed combined SM, while existing single SMs fail to detect enough correspondences for registration. Our results demonstrate that MIND, SIFT, and logLR SMs capture essential aspects of the similarity between RS modes, and their properties are complementary for designing a new more efficient multimodal SM.**

***Index Terms*— Area-based similarity measure (SM), combined SM, linear binary classifier, multimodal image registration, remote sensing (RS), structural similarity, support vector machine (SVM).**

I. Introduction

THE process of image registration brings in the same coordinate system two or more images of the same area acquired in different conditions: different time instances, viewpoints, and/or modalities [1], [2]. Accurate image registration is essential in such fields as remote sensing (RS) [2], [3], medical imaging [4]–[7], or computer vision [8], [9]. The challenging multimodal registration of RS images is important for the multisensor study of the earth's surface where each modality (e.g., visible, infrared, thermal, radar, and LIDAR) provides complementary information on a study area [10].

Image registration involves the following stages [1]: finding putative correspondences between the registered reference and template images (RI and TI), detecting outliers, estimating parameters of geometrical transformation, and image warping. The stage of finding correspondences, the focus of the present research, aims at identifying similarity between registered images. A measure of similarity between the fragments of different modalities is at the core of any multimodal registration method. Over the past decades, a number of multimodal similarity measures (SMs) have been designed in RS, medical imaging, and computer vision fields. These SMs can be roughly divided into area-based and feature-based or structural SMs [11]. In feature-based methods, the first stage is the calculation of points of interest for RI and TI. Then, a descriptor is calculated for each interesting point. The similarity between reference and template fragments is calculated as a distance between the corresponding descriptors. For area-based SMs, the similarity is measured directly by comparing two image fragments. For this group of methods, interesting points' calculation is often replaced by exhaustive search in the search region. While feature-based methods have lower computational complexity and better adaptation to structural changes between RI and TI intensities, area-based methods are more stable in the presence of sensor noise.

Representatives of the first group are normalized correlation coefficient (NCC) [12], mutual information (MI) [13], phase correlation (PC) [14], and recently proposed logarithmic likelihood ratio (logLR) [15] SMs. The second group includes scale-invariant feature transform (SIFT) [16], a version of SIFT specially adapted for radar images, scale-invariant feature transform-octave (SIFT-OCT) [17], histogram of oriented gradients (HOGs) [18], histogram of orientated phase congruency (HOPC) [19], and modality independent neighborhood descriptor (MIND) [20] SMs.

In the process of image registration, multimodal SMs are often utilized as similarity-based binary classifiers: the SM value for a particular pair of the registered image fragments is compared to a decision threshold to choose between the null hypothesis—the fragments are dissimilar, mostly attributed to a false correspondence—and alternative

Manuscript received June 3, 2019; revised September 21, 2019 and April 7, 2020; accepted April 20, 2020. *(Corresponding author: Benoit Vozel.)*

Mikhail L. Uss and Sergey K. Abramov are with the Department of Information and Communication Technologies, National Aerospace University, 61070 Kharkov, Ukraine (e-mail: uss@xai.edu.ua; s.abramov@khai.edu).

Benoit Vozel and Kacem Chehdi are with IETR UMR CNRS 6164, University of Rennes 1, Enssat, 22305 Lannion, France (e-mail: benoit.vozel@univ-rennes1.fr; kacem.chehdi@univ-rennes1.fr).

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

Digital Object Identifier 10.1109/TGRS.2020.2992597

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

hypothesis—the fragments are similar, mostly corresponding to the true correspondence. Here, the SM value represents the "score" of the associated similarity-based binary classifier [21]. In what follows, we use the terms "SM" and "SM value" to refer to the associated similarity-based binary classifier and its score, respectively.

The key requirements for multimodal SMs include high efficiency, universality, and appropriate computational complexity (can be characterized by the number of arithmetic operations needed for SM calculation) [11]. Informally, a more efficient SM should provide some benefits to the image registration process as compared to a less effective SM, for example, to allow registering more complex image pairs or improve registration accuracy [15]. These benefits are difficult to quantify in a unique way. As a binary classifier, an SM can be characterized by general-purpose measures, like true positive rate (TPR) and false-positive rate (FPR), positive predictive value (PPV), and area under the curve (AUC) [22]. The latter measure, AUC, characterizes a probability that an SM will rank a randomly chosen true correspondence higher than a randomly chosen false correspondence [23].

Universality presumes an SM to be applicable to a wide spectrum of registration cases and settings. We consider here one aspect of universality—an ability to deal with different pairs of modes. Such an aspect is important for creating fully automatic registration frameworks, as it frees practitioners from the unobvious choice of the best SM for each specific registration case. A growing number of multimodal SMs makes this choice increasingly complicated.

The performance of multimodal SMs in RS applications was compared in relatively few publications. Inglada demonstrated that MI was more suitable than NCC for registering different bands of multitemporal SPOT images [24]. In the study of Mikolajczyk and Schmid [25], the performance of interest region descriptors was comparatively evaluated for real-life images with different geometric and photometric distortions. It was shown that gradient location and orientation histogram (GLOH) and SIFT descriptors had a similar performance favorably compared to the shape context [26], steerable filters [27], and low-dimensional descriptors. Uss *et al.* [15] compared the performance of NCC, MI, SIFT, SIFT-OCT, PC, and logLR SMs for optical-to-optical, optical-to-digital elevation model (DEM), optical-to-radar and radar-to-DEM registration cases. This study revealed that MI, SIFT-OCT, and logLR had higher discriminative power than NCC, PC, and SIFT for all registration cases. However, the relative ranking of MI, SIFT-OCT, and logLR depends on the fragment size and registration case. Ye *et al.* [19] compared the performance of NCC, MI, HOG, and HOPC SMs for visible-to-infrared, visible-to-LIDAR, visible-to-SAR, and image-to-map cases. They demonstrated that NCC had the worst performance followed by MI, HOG, and HOPC. HOPC had the best performance, but the gain compared to the HOG SM for many cases was small. Rigorous comparison of multimodal SMs performance for the problem of dense multimodal stereovision was done by Yaman and Kalkan [28]. They concluded that for visible-to-infrared stereo pairs the best performing SM is MI closely followed by HOG, SIFT, and local self-similarity (LSS) [29]. Binary descriptors including fast retina keypoint (FREAK) [30], binary robust independent elementary features (BRIEF) [31], census transform (Census) [32] as well as NCC failed to provide comparable results in case of significant difference between the registered modes.

In view of many available multimodal SMs without obvious predominance of one of them, this article raises the following question: do available SMs capture different aspects of similarity between multimodal images? The contribution of the response provided in this work is to show that a combination of specially selected multimodal SMs is more universal and effective (in the sense of AUC) than each individual SM.

In the available literature, the main efforts to derive a more effective SM using positive features of several existing SMs have been in the direction of augmenting MI SM with image spatial information (e.g., gradient information). We discuss existing extensions of MI SM in Section II-B. The main limitation of such an approach of augmenting an SM is that existing feature-based and area-based SMs have complex structure; it is not clear which structure elements from different SMs are complementary and how to combine them in a new SM.

To overcome this shortage and derive a more effective SM, we propose to combine values of two or more SMs calculated for a particular pair of image fragments into a single feature vector. In this manner, each existing SM can be considered and effectively used as a part of a combined SM. A new SM, which we call later a combined SM and denote comSM, is designed as a binary classifier that separates the feature vectors into two categories corresponding to similar and dissimilar fragments. In this article, we use a linear support vector machine (SVM) classifier trained on the basis of 16 accurately registered image pairs representing visible-to-infrared, visible-to-DEM, visible-to-radar, and radar-to-DEM multitemporal/multimodal cases. The considered set of images is representative as it contains various RS mode combinations and land covers.

The main finding of this study is that the optimal combination of SMs for a wide range of RS registration cases should include an area-based SM and a structural SM. The first is responsible for reliable discrimination of image fragments under a simpler intensity transformation model (e.g., linear intensity change between RI and TI fragment). The second helps to detect similarity for fragments under complex structural changes. We also demonstrate that a rather restricted subset of existing SMs, MIND, logLR, and SIFT-OCT, is the most suitable for joint use in combined SMs. Even more important, we propose a single combined SM that is applicable for all considered registration cases (visible-to-infrared, visible-to-DEM, radar-to-DEM, and visible-to-radar) and shows the performance close to the combined SMs individually optimized for each registration case. Such an interesting feature may indicate a way of constructing a new efficient multimodal SM that borrows advantages of MIND, logLR, and SIFT-OCT SMs.

The remaining of this article is organized as follows. Section II recalls state-of-the-art SMs for multi- and mono-modal image registrations. We then introduce and analyze the combined SM based on the SVM classifier in Section III. In the experimental Section IV, the newly proposed and existing SMs are extensively compared on real data

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

according to the ROC curve and AUC criteria. This article is concluded in Section V along with comments on future research directions.

II. Multimodal SMs and Their Combined Use

This section recalls several well-known SMs previously proposed for multimodal image registration problems, their own advantages and disadvantages. We then discuss existing combined SMs that utilize the advantages of more than one SM.

A. Single SMs

Among available multimodal SMs, we consider here only those that were successfully applied to the multimodal image registration problem (preferably from RS field) and demonstrated top performance characteristics: MI, SIFT-OCT, HOG, HOPC, MIND, and logLR. The well-known NCC and PC SMs are retained for comparison as well. In what follows, we briefly introduce each of these SMs.

SMs suitable for multimodal image registration can be grouped into area-based and feature-based ones [24]. The well-known representatives of the area-based group are NCC and MI SMs. They are defined as follows:

$$\mathrm{NCC} = \frac{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(I_{\mathrm{RI}}(i,j)-m_{\mathrm{RI}})(I_{\mathrm{TI}}(i,j)-m_{\mathrm{TI}})}{\sigma_{\mathrm{RI}}\sigma_{\mathrm{TI}}} \quad (1)$$

$$\mathrm{MI} = H_{\mathrm{RI}} + H_{\mathrm{TI}} - H_{\mathrm{RITI}} \quad (2)$$

where $I_{\mathrm{RI}}(i,j)$ and $I_{\mathrm{TI}}(i,j)$ denote intensity of the RI and TI fragments of size N by N pixels, m_{RI}, $m_{\mathrm{TI}}\sigma_{\mathrm{RI}}$, and σ_{TI} are mean and standard deviation of RI and TI fragments, respectively, H_{RI} and H_{TI} denote entropy of RI and TI fragments, H_{RITI} is their joint entropy. Later, we use the absolute value of NCC, but for simplicity still refer to it as NCC.

NCC is limited to linear intensity change between RI and TI, which is generally not the case for multimodal images [15]. MI relaxes linear intensity requirement to statistical dependence between RI and TI intensities. The drawback of both NCC and MI is that they do not take into account spatial structures present in RI/TI, which could reveal important relationships between the compared modes [20].

PC SM utilizes shift property of the Fourier transform [14]. Given two images $f_1(x,y)$ and $f_2 = f_1(x-\Delta x, y-\Delta y)$ mutually shifted by $(\Delta x, \Delta y)$ and their respective Fourier transforms $F_1(u,v)$ and $F_2(u,v)$, the inverse transform of normalized cross power spectrum $C_{\mathrm{cps}}(u,v) = (F_1(u,v)\cdot F_2(u,v)^*)/(|F_1(u,v)\cdot F_2(u,v)^*|)$ between these two images have a form close to the delta-function placed at $(\Delta x, \Delta y)$. Here, * indicates the complex conjugate. Therefore, the similarity between two images can be measured by $C_{\mathrm{cps}}(0,0)$. PC can be used in multimodal cases because it relies only on the phase spectrum that is less sensitive to nonuniform illumination and nonlinear intensity change between the registered images. However, the neglecting amplitude spectrum makes PC less robust to image noise influence.

The recently proposed logLR SM [15] also belongs to the area-based group of SMs. This SM assumes that RI and TI fragments are correlated realizations of fractal Brownian field with unknown parameters. LogLR tests the null hypothesis that RI and TI are uncorrelated against the alternative hypothesis that they are correlated. All unknown RI/TI parameters are estimated at this stage (texture amplitude and roughness, the correlation between RI and TI). LogLR also assumes too simple linear transformation between RI and TI intensities. However, the advantage of this SM is in the rigorous modeling of image noise characteristics including noise signal-dependence property and spatial correlation. As a result, logLR is able to detect fewer correspondences if the real intensity transform between RI and TI deviates from the linear one, but it provides reliable discrimination if the linear model is valid.

Overcoming the area-based SM drawbacks is the core of structural SMs. The idea behind structural SMs is to find a mode-independent descriptor of the registered image fragments such that they can be compared using a simple mono-mode metric, for example, the sum of squared distances (SSD) [33]. SIFT descriptor was first introduced in [16]. This descriptor has found a great number of applications in the computer vision domain. It consists of stacked gradient orientation histograms calculated in overlapping blocks covering the full area of the analyzed image fragment. SIFT invariance to complex intensity changes is ensured by the gradient use. However, gradients are sensitive to image noise, and SIFT in its original version has been found unstable for multimodal registration. This is especially evident for radar images due to their intensive speckle noise. Suri and Reinartz [13] adopted the SIFT descriptor to optical-to-radar registration case and called it SIFT-OCT [16]. The modifications involve skipping orientation assignment stage (for RS image registration, mutual orientation can be derived from satellite platforms orbital data), and skipping the main scales of radar images to reduce speckle noise influence [17]. The similarity between two image patches is calculated as SSD between two SIFT-OCT (or SIFT) descriptors, and we will refer to it as $\mathrm{SIFT\text{-}OCT}_{\mathrm{SSD}}$.

Ye *et al.* [19] have recently proposed a novel structural SM for multimodal image registration named $\mathrm{HOPC}_{\mathrm{NCC}}$ and utilized the known HOG descriptor for the same purpose ($\mathrm{HOG}_{\mathrm{NCC}}$ SM). $\mathrm{HOG}_{\mathrm{NCC}}$ represents the NCC between two HOG descriptors calculated for RI and TI. $\mathrm{HOPC}_{\mathrm{NCC}}$ uses the framework of HOG but substitutes image gradient with more robust features: amplitude and orientation of the phase congruency [34]. Experimental results in [19] demonstrate that both descriptors outperform MI and have comparable performance. In our experiments, we also tested $\mathrm{HOG}_{\mathrm{SSD}}$ and $\mathrm{HOPC}_{\mathrm{SSD}}$ versions of $\mathrm{HOG}_{\mathrm{NCC}}$ and $\mathrm{HOPC}_{\mathrm{NCC}}$ with SSD distance between descriptors instead of NCC suggested by Ye *et al.* [19]. In all cases, $\mathrm{HOG}_{\mathrm{SSD}}$ and $\mathrm{HOPC}_{\mathrm{SSD}}$ provided slightly better results as compared to $\mathrm{HOG}_{\mathrm{NCC}}$ and $\mathrm{HOPC}_{\mathrm{NCC}}$. Therefore, in what follows we retained $\mathrm{HOG}_{\mathrm{SSD}}$ and $\mathrm{HOPC}_{\mathrm{SSD}}$ SMs.

MIND descriptor was proposed by Heinrich *et al.* [20]. It is based on "…a local representation of image structure, which can be estimated through the similarity of small image patches within one modality, is shared across modalities" [20]. MIND extracts the distinctive local structures such as corners, edges,

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

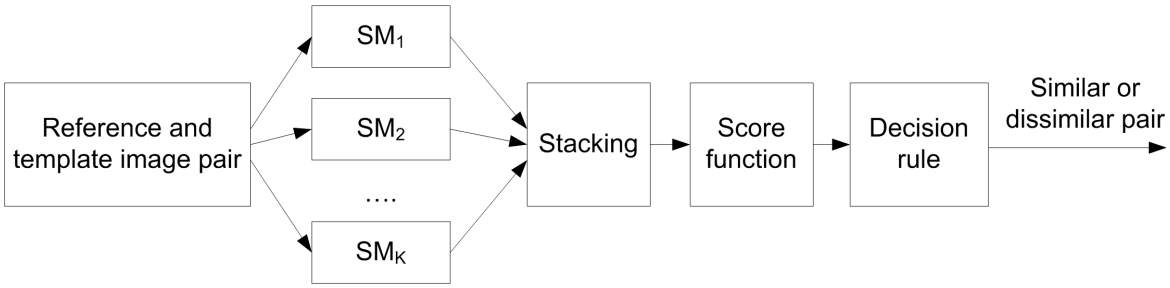

Fig. 1. Proposed combined SM.

and homogeneously textured regions in a way, which is preserved across different modalities. The strength of MIND is in its invariance to nonfunctional intensity relations, image noise, and nonuniform bias fields. MIND_{SSD} measures the similarity between two compared image patches as SSD between two MIND descriptors.

Apart from the above-mentioned SMs retained for building a combined SM in the next two sections, there exist many other SMs suitable for different RS registration cases. Registration of optical images on the basis of correlation of wavelet features was proposed in [35] and [36]. The advantage of wavelet features is that they allow multiresolution analysis and efficient parallel implementation. Shearlets were shown to be advantageous over wavelets for registration of multitemporal images with many directional edge-like features (rivers, roads, etc.) [3], [37], [38].

Deep learning (DL) is another emerging trend in designing effective SMs [39], [40]. Results obtained in this article could be useful for those DL methods that use existing SMs as a part of complex networks. For example, de Vos *et al.* [41] used the NCC SM as a part of the loss function for learning a registration convolutional neural network (CNN). The usage of NCC limits this CNN to the mono-modal case. An effective combined SM could help to extend this approach to the multimodal case.

B. Joint SMs

Considerable research has been devoted toward augmenting MI with spatial information. Pluim *et al.* [42] proposed an SM that comprises two terms, MI and a gradient term. The latter term highlights the same gradient orientations and strong gradients in both modalities. This augmented SM was successfully applied in [43] to a multimodal stereo vision system made up of an infrared camera and a color one forming together a stereo pair. A similar idea was adopted in [44] where MI is calculated between RI and TI images preliminarily modified by summing the equalized original and gradient images. Anthony and Lofffeld [45] augmented MI differently by weighting pixels with respect to the local image gradient, variance and entropy assuming that a high value of these terms indicates more utile pixels for similarity calculation. Instead of using gradient images, Mellor and Brady [46] proposed to calculate MI SM of the local phase to maximize a structural relationship between images. Sun and Ray [47] proposed to overcome the limits of MI by using a compound MI, which aggregates information from multiple marginal densities of image intensity distributions. The compound MI avoids calculating high-order histograms, which are computationally complex to deal with, but still incorporates spatial information in MI. These approaches are in some way limited as they augment MI with forms of gradient-based or phase-based structural similarity that are significantly simpler as compared to such advanced SMs as SIFT, HOG, MIND, or HOPS.

On a larger scale, Feng *et al.* [48] combined both feature- and area-based SMs for robust registration of RS images covering areas with complex terrain. The feature-based SIFT SM was used to perform in the first stage a coarse large-scale registration, robust to outliers; at the second stage, the area-based NCC SM was used for fine-tuning the geometrical transformation model. This approach is different from ours in that several SMs are not simultaneously applied to the same image patch.

III. Multimodal SMs and Their Combined Use

In this section, we propose a combined SM as the one built by learning a binary SVM classifier with all potentially optimal and selected suboptimal SMs forming a joint feature vector. We then discuss the performance criteria of such SMs. Using ROC convex hull (ROCCH) analysis, we demonstrate that among the considered SMs, there are typically two potentially optimal ones, the rest of them are suboptimal. The new combined SM performs better than each individual SM appended into the feature vector and better that ROCCH of these SMs.

A. Proposed Combined SM

The above-introduced SMs leverage different aspects of RS images structure to catch similarity between modes, either intensity correspondence between RI and TI, or different forms of structural resemblance between RI and TI. The natural question arises whether these aspects are complementary. The positive answer to this question would indicate the existence of a more powerful SM somehow incorporating several aspects of multimodal similarity.

As mentioned above, in this article, for building a combined SM we consider stacking values of two or more SMs in a combined feature vector (Fig. 1)

$$\mathbf{f} = (\text{SM}_1, \text{SM}_2, \ldots, \text{SM}_K)^{\text{T}} \tag{3}$$

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

where K denotes the number of SMs used in our combined SM, SM_k is the value of the kth SM, $k = 1, \ldots, K$. The combined SM represents a binary classifier that takes a decision of absence (the null hypothesis) or existence (the alternative hypothesis) of similarity between the two fragments from different modalities according to the following rule:

$$y_{\mathbf{f}} = \begin{cases} 1, & d(\mathbf{f}) > d_{\text{th}} \\ -1, & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\mathbf{f}$ is a query (combined) feature vector, $d(\mathbf{f})$ is a scalar score function separating similar and dissimilar fragment classes, d_{th} is a decision threshold, label y value 1 for the null hypothesis, and -1 for the alternative hypothesis. We understand both single and combined SMs as binary classifiers. The difference is that the classifier score for a single SM is its value, whereas for a combined SM, it is a scalar function $d(\mathbf{f})$ of the combined feature vector.

The proposed combined SM can be understood from the viewpoint of ensemble learning methodology that builds "… a predictive model by integrating multiple models" [49]. As we consider a fixed set of methods that are not individually modified for building a combined model, our proposed method falls into an independent framework of ensemble learning. Using the output of many single SMs for learning one combined classifier instead of learning it directly from original training data is known as meta-combination stacking technique. This technique is known to offer high generalization accuracy when stacked classifiers are "mutually orthogonal" [50]. The main difference with the proposed method is that the single SMs considered in it are rule-based classifiers that are not optimized to a given training set. In this context, we have experimentally found that a combined SM performs better than every single SM, provided their output is complementary to each other.

B. SM Performance Criteria Specific to Image Registration Problem. Baseline Combined Classifier

We leverage two objective criteria to compare the performance of multimodal SMs: ROC curve and AUC. Let us briefly recall the meaning of each criterion.

A binary classifier that uses a particular SM operates by calculating the SM value for a pair of RI/TI fragments and comparing this value to a decision threshold. The decision that RI/TI fragments are similar is undertaken if the calculated SM value exceeds a threshold (we assume that a higher SM value corresponds to a higher similarity between the registered images). For a given decision threshold, the binary classifier is characterized by TPR and FPR values. ROC curve represents TPR against FPR at various decision threshold settings. AUC is an integral measure of classifier performance, in the sense of its ability to assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. AUC reduces the ROC curve to a single number and can be used to rank classifiers: an increase of AUC toward unity indicates a more powerful classifier.

When applied to the same data, each SM is characterized by its own ROC curve in the ROC space. The convex hull of the set of points in ROC space is called the ROCCH of the corresponding set of classifiers. Depending on their position with respect to ROCHH, ROC curves and respective SMs can be divided into potentially optimal and suboptimal ones [22]. A classifier is called potentially optimal if and only if it lies on ROCHH but not necessarily coincides with it. A classifier that lies below ROCCH is called suboptimal. In Fig. 2, classifiers A and B are potentially optimal, whereas C is suboptimal.

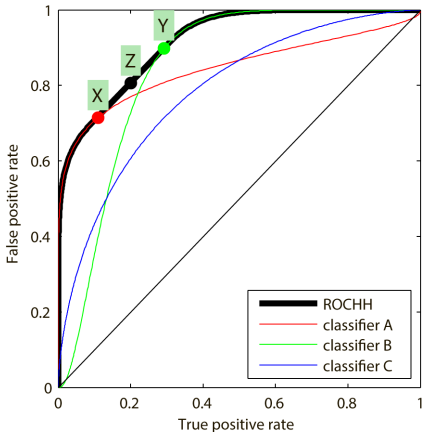

Fig. 2. Illustration of ROCCH of a set of classifiers.

Having more than one potentially optimal classifier, the simplest way to create a more effective SM is to interpolate between them [22]. Interpolation means sampling decisions of each classifier. For example, the classifier with (TPR, FPR) corresponding to point Z at ROCHH can be derived by choosing a decision of classifier A at point X with a probability or sampling rate of r and decision of classifier B at point Y with a probability of $1 - r$. The sampling rate $0 \leq r \leq 1$ determines the position on the ROCHH curve between X and Y. The ROC curve of such an interpolated classifier will coincide with the ROCHH and have AUC higher than that of each potentially optimal and suboptimal classifier (A, B, and C in the example).

C. Selection of Classifier. Comparison to the Baseline Classifier

To find a suitable separating surface $d(\mathbf{f})$, the existing binary classification methods can be used including SVM, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision trees, random forests, and neural networks [51]. The choice of a particular classifier depends upon the data characteristics and complexity of the separating surface. In this subsection, we use experimental data to demonstrate that a linear SVM classifier is the most suitable one for designing the proposed multimodal combined SM.

For this analysis, 2500 positive and 2500 negative samples are collected from visible-to-infrared pair of images (introduced later in Experimental Section IV) for each SM. The RI represents band #60 (central wavelength of 955.93 nm, spectral width of 11.3871 nm) of the Hyperion sensor; TI is band #3 (wavelength range 525–600 nm) of the Landsat8 OLI sensor.

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

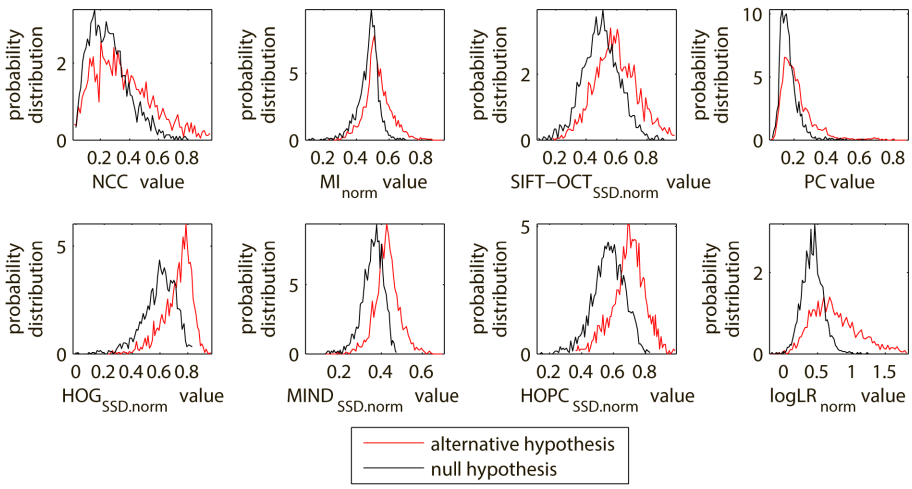

Fig. 3. Experimental distributions of normalized SM values for similar and dissimilar fragments of visible-to-infrared image pair.

The selection of the fragment size for calculating an SM is governed by two main factors: the spatial resolution of the SM output and the SM efficiency. The smaller the fragment size, the finer details of the displacement field between RI and TI can be estimated. However, the reliability of an SM is reduced for smaller fragments, due to a smaller amount of information available. As the reference fragment size, we could consider fragments of size 16 × 16 pixels as chosen in the original SIFT article [16]. In the RS literature, a typical choice for SMs calculation is a fragment size of 32 × 32 pixels. As a compromise, we have selected an intermediate fragment of size 21 × 21 pixels for the main experiments performed (an odd linear size is required by logLR SM). Then, for better supporting the results obtained in the main experiments with such an intermediate fragment size, we considered a smaller fragment size of 13 × 13 pixels as well.

For comparison utility, MI, SIFT-OCT$_{\mathrm{SSD}}$, HOG$_{\mathrm{SSD}}$, MIND$_{\mathrm{SSD}}$, HOPS$_{\mathrm{SSD}}$, and log LR SMs values were preliminarily normalized to the range [0, 1] (NCC and PC are originally within this range). This normalization is valid for a fragment size of 21 × 21 pixels and settings of SMs as specified in Experimental Section IV: $\mathrm{MI}_{\mathrm{norm}} = \mathrm{MI}/1.75$, $\text{SIFT-OCT}_{\mathrm{SSD.norm}} = 1 - \text{SIFT-OCT}_{\mathrm{SSD}}/4000$, $\mathrm{HOG}_{\mathrm{SSD.norm}} == 1 - \mathrm{HOG}_{\mathrm{SSD}}/0.02$, $\mathrm{MIND}_{\mathrm{SSD.norm}} = 1 - \mathrm{MIND}_{\mathrm{SSD}}/0.25$, $\mathrm{HOPC}_{\mathrm{SSD.norm}} = 1 - \mathrm{HOPC}_{\mathrm{SSD}}/0.003$, $\mathrm{logLR}_{\mathrm{norm}} == F_N^{-1}(F_{\chi^2}(\mathrm{logLR}, 3), 0, 1)/6 + 0.5$, where $F_{\chi^2}(x, \upsilon)$ denotes cumulative distribution function (CDF) of $\chi^2(\upsilon)$ distribution with υ degrees of freedom, and $F_N^{-1}(x, m, \sigma)$ denotes inverse CDF of the normal distribution $N(m, \sigma)$. LogLR normalization takes into account that its distribution for zero hypothesis is $\chi^2(3)$ [15]; $F_N^{-1}(F_{\chi^2}(\mathrm{x}, 3), 0, 1)$ function transforms $\chi^2(3)$ to the standard normal distribution. Distributions of normalized SM values for the above-mentioned test image pair are shown in Fig. 3. We will refer to normalized SMs by using additional index "norm" only when relevant.

For all SMs, the distribution of normalized values for the null and alternative hypothesis exhibit an apparent non-Gaussianity. Deviation from the standard normal distribution, measured by Pearson's moment coefficient of skewness and kurtosis [52], is the least significant for SIFT-OCT$_{\mathrm{SSD}}$ and HOPC$_{\mathrm{SSD}}$ (with the absolute skewness not exceeding 0.4 and kurtosis not exceeding 3.5). For logLR, MIND$_{\mathrm{SSD}}$, HOG$_{\mathrm{SSD}}$, NCC, and MI, deviation from the normal distribution is significant with the absolute skewness reaching 0.7–1.1 and kurtosis varying from 3.5 to 6.5. For PC, distributions are strongly not normal with the skewness about 2.2 and kurtosis about 11.

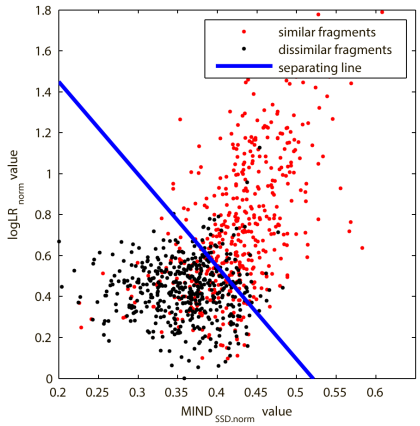

Fig. 4. Values of ($\mathrm{MIND}_{\mathrm{SSD.norm}}$, $\mathrm{logLR}_{\mathrm{norm}}$) feature vectors for visible-to-infrared registration case. Separating line is obtained using an SVM classifier.

The structure of the separating surface between different SMs is illustrated in Fig. 4 for the pair (MIND$_{\mathrm{SSD}}$, logLR). In this case, the separating line has a simple linear form. Similarly, a linear separating line was found suitable for other SM combinations. For linearly separable classes, the usage of complex classifiers such as QDA, decision trees, random forests, neural networks, and nonlinear SVM classifier is unnecessary [51]. Among LDA and linear SVM classifiers, the latter is preferable as LDA requires data to follow the normal distribution [53], and for many SMs, this requirement is violated even after normalization. Therefore, we selected linear SVM to calculate separating hyperplane between combined

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

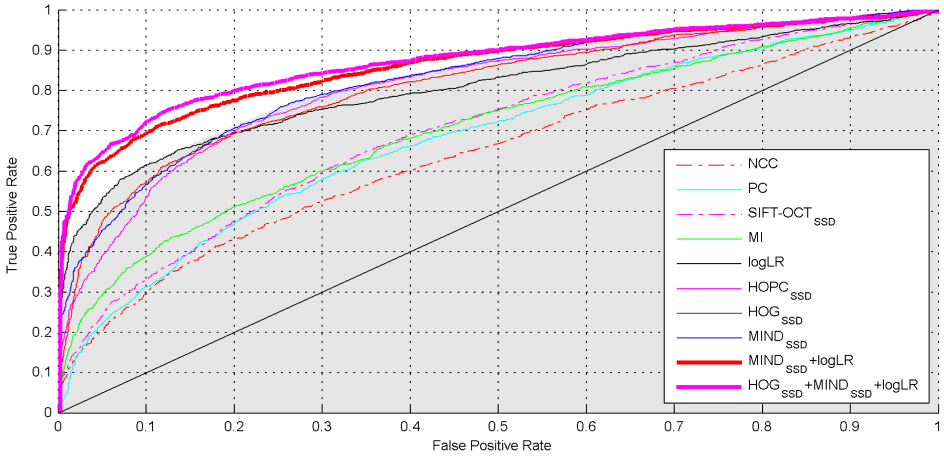

Fig. 5. ROC curves of individual SMs for visible-to-infrared registration case (best in color). Area below ROCCH is shown in gray.

feature vectors corresponding to similar and dissimilar image fragment pairs.

A binary SVM classifier separates feature vectors into two classes by the maximum-margin hyperplane according to the following decision rule [51]:

$$y_{\mathbf{f}} = \text{sign}\left(\left[\sum_{i=1}^{N} c_i y_i g(\mathbf{f}_i, \mathbf{f})\right] + b\right) \tag{5}$$

where $\mathbf{f}$ is a query feature vector, $\mathbf{f}_i$ are training feature vectors with labels y_i ($y_i = 1$ for the null hypothesis, and $y_i = -1$ for the alternative hypothesis), N is the number of training vectors, $g(\cdot,\cdot)$ denotes the reproducing kernel (e.g., polynomial reproducing kernels in the form $g(\mathbf{f}_1, \mathbf{f}_2) = (1 + \mathbf{f}_1^{\mathrm{T}} \cdot \mathbf{f}_2)^p$, where p is some positive integer), c_i and b are coefficients estimated during the training of the SVM classifier. For a linear classifier, the decision rule (5) simplifies to

$$y_{\mathbf{f}} = \text{sign}(\beta^{\mathrm{T}} \cdot \mathbf{f} + b) \tag{6}$$

where β is a $K \times 1$ vector of the maximum-margin hyperplane coefficients, b is a scalar hyperplane offset.

To prevent SVM classifier overfitting, k-fold cross-validation [51] is used with $k = 10$. According to this approach, the training sample is randomly partitioned into k subsamples of approximately equal size. Training is performed k times; in each round of training, one subsample is retained for validation, the rest $k - 1$—for training. The value of a combined SM for each sample is therefore calculated when this sample is not included in the training subset.

For the considered visible-to-infrared case, ROC curves for each SM retained for comparison are shown in Fig. 5. The area below ROCCH is shown in gray color in Fig. 5. Two potentially optimal SMs can be identified: MIND_{SSD} and logLR. The rest of SMs are suboptimal. For the considered example, AUC takes the value of 0.8245 for MIND_{SSD}, 0.8050 for logLR, and 0.8371 for ROCCH. The latter value corresponds to AUC of classifier interpolated between MIND_{SSD} and logLR.

Training SVM classifier for a combined feature vector is more efficient than interpolating between individual classifiers because even suboptimal classifiers can be used to improve the combined SM performance. In Fig. 5, ROC curves for two combined SMs are shown, one combining potentially optimal MIND_{SSD} and logLR SMs (red curve) and another additionally considering suboptimal HOG_{SSD} classifier (pink curve). The first combined SM has AUC of 0.8616 that is already higher than AUC for ROCCH. The second combined SM has an even higher AUC of 0.8720 gained by using suboptimal HOG_{SSD} SM.

IV. Comparison of Single and Combined SMs Performance on Real Multimodal RS Images

This section compares the performance of NCC, MI, PC, $\text{SIFT-OCT}_{\text{SSD}}$, MIND_{SSD}, HOG_{SSD}, HOPC_{SSD}, logLR, and the SM combining MI and image gradient [42] denoted as MIgrad, and the proposed combined SMs on real multimodal image pairs, which represent typical registration cases in RS field. The primary objective of this comparison is to verify that a joint usage of several SMs has a better performance than that of single involved SMs and to identify the minimum combination of SMs suitable for different multimodal registration cases and for general-purpose multimodal registration.

A. Training and Test Data

The comparison of retained SMs is based on sixteen real-life image pairs covering four different multimodal/multitemporal registration cases: visible-to-infrared, visible-to-DEM, visible-to-radar, and radar-to-DEM. Test images are collected from the following sensors: Hyperion [54], Landsat-8 (OLI sensor) [55], Sentinel-2 [56] for visible/infrared data, ASTER GDEM2 [57], [58] and ALOS World 3D with 30-m resolution [59] for DEM data, SIR-C [60] and Sentinel-1 [61] for radar data. This set of images is representative in that it covers the main RS modes—visible, infrared, radar, DEM—various land covers (including forestry, rural and urban areas, agricultural areas, rivers, lakes), seasons (images taken in both summer and winter time), sensors, and acquisition time lags (time between RI and TI varies from 1 to 22 years implying that all image pairs are multitemporal ones). Optical images

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
SOURCES OF RI AND TI

Pair index	Modality	Sensors	RI image (Band, Lat°, Lon°, Time)	TI image (Band, Lat°, Lon°, Time)
1	visible-to-infrared	Hyperion – Landsat-8	EO1H1800252002116110KZ, band #25, (49.43, 32.06), 2002, Apr	LC81800262014070LGN00, band #1, (48.86, 31.63), 2014, Mar
2	-//-	Hyperion – Landsat-8	EO1H2010262006218110PZ, band #155, (48.39, -1.16), 2002, Aug	LC82010262013342LGN00, band #5, (48.88, -0.82), 2013, Dec
3	-//-	Sentinel-2 - Landsat-8	L1C_T36TXT_A017650_20181108T084334, band #4, (47.3393, 35.0438), 2018, Nov	LC08_L1TP_177027_20141203_20170416_01_T2 band #2, (47.47, 35.74), 2014, Dec
4	-//-	Sentinel-2 - Landsat-8	L1C_T19PHK_A019198_20190224T145721, band #5, (8.5367, -65.7776), 2019, Feb	LC08_L1TP_003054_20140319_20170425_01_T1 band #2, (8.66, -65.83), 2014, Mar
5	visible-to-radar	Landsat-8 – SIR-C	LC81990262014363LGN00, band #8, (48.84, 2.31), 2014, Dec	PR41419 Band HV, (48.96, 2.87), 1994, Oct
6	-//-	SIR-C - Landsat-8	pr43020 Band HV, (49.82, 36.75), 1994, Oct	LC81770252016039LGN00, Band #8, (50.23, 36.88), 2016, Feb
7	-//-	Sentinel-2 - Sentinel-1	L1C_T36TXT_A017650_20181108T084334, band #4, (47.3393, 35.0438), 2018, Nov	S1A_IW_SLC__1SDV_20190314T153629_20190314T153656_026336_02F1DC_AA57, polarization VH, swath #3, tile #6, (47.21, 34.88), 2019, Mar
8	-//-	Sentinel-2 - Sentinel-1	L1C_T19PHK_A019198_20190224T145721, band #5, (8.5367, -65.7776), 2019, Feb	S1A_IW_GRDH_1SDV_20180818T100859_20180818T100928_023299_0288AE_14BE, polarization VV, swath #1, tile #1, (8.35, -65.39), 2018, Aug
9	visible-to-DEM	Hyperion - GDEM2	EO1H1800252002116110KZ, band #25, (49.43, 32.06), 2002, Apr	ASTGTM2_N49E031, ASTGTM2_N49E032
10	-//-	Landsat-8 – GDEM2	LC82010262013342LGN00, band #5, (48.88, -0.82), 2013, Dec	ASTGTM2_N48W001
11	-//-	Sentinel-2 - GDEM2	L1C_T36TXT_A017650_20181108T084334, band #4, (47.3393, 35.0438), 2018, Nov	N046E034, N047E034, N046E035, N047E035
12	-//-	Landsat-8 – GDEM2	LC08_L1TP_003054_20140319_20170425_01_T1 band #3, (8.66, -65.83), 2014, Mar	N008W066, N008W067, N009W066, N009W067
13	radar-to-DEM	GDEM2 – SIR-C	N048E002, N048E003, N049E002, N049E003	PR41419 polarization HV, (48.96, 2.87), 1994, Oct
14	-//-	SIR-C - GDEM2	pr43020 polarization HV, (49.82, 36.75), 1994, Oct	N049E036, N049E037, N050E036, N050E037
15	-//-	GDEM2 – Sentinel-1	N046E034, N046E035, N047E034, N047E035	S1A_IW_SLC__1SDV_20190314T153629_20190314T153656_026336_02F1DC_AA57, polarization VH, swath #3, tile #5, (47.04, 34.93), 2019, Mar
16	-//-	GDEM2 – Sentinel-1	N008W066, N008W067 N009W066, N009W067	S1A_IW_GRDH_1SDV_20180818T100859_20180818T100928_023299_0288AE_14BE, polarization VV, swath #1, tile #1, (8.35, -65.39), 2018, Aug

have a scarce cloud and water cover. Details of all image pairs are given in Table I.

All pairs were registered using the recently proposed registration with accuracy estimation (RAE) registration method [62] that is able to provide subpixel registration accuracy even for the most complex multimodal cases. Fragments of some of the registered images are available at [62]. Examples of a registered visible-to-DEM and visible-to-radar image pairs are shown in Section IV-E.

Overall, 50 000 truly corresponding and 50 000 falsely corresponding RI/TI fragments were collected from the registered RI and TI images. True correspondences uniformly cover the intersection area of RI and TI images. False correspondences were obtained by randomly shifting TI fragment at a distance exceeding the TI size. Test fragments uniformly represent all registration cases. In all cases, when a combined SM is compared to an individual SM for a particular case, the SVM classifier is trained on the data corresponding to this case. The fragment size used in our analysis is 21 × 21 pixels.

The parameters of SIFT-OCT$_{SSD}$, MIND$_{SSD}$, HOG$_{SSD}$, and HOPC$_{SSD}$ SMs were set according to the default values suggested in their respective original articles. Specifically, for the MIND descriptor, the parameter $\sigma = 0.5$, the variance of image noise is calculated in six-neighborhood (default settings). For the SIFT descriptor, the sigma of the Gaussian weighting function is equal to one half of the width of the descriptor window, the latter coincides with the TI size, the cell size is 4 × 4 pixels, and the number of bins is eight. For HOG and HOPC descriptors, the block overlap is 0.5, the cell size is 4 × 4 pixels, the block size is 3 × 3 cells, and the number of bins is eight. MI SM was implemented using the Feature Selection Toolbox for C and MATLAB (FEAST-v1.1.4); HOG descriptor—using MATLAB extractHOGFeatures function, SIFT-OCT descriptors—using VLFeat library (v 0.9.20), MIND and HOPC descriptors—using implementations provided by the authors (see references in [19] and [20]). MI was calculated between RI/TI fragments that had been preliminarily normalized to zero mean and unit variance. The number of bins was set to 30.

B. Analyzed SM Combinations and SMs Ordering

To find the best combination of the considered eight SMs, we have trained all possible combinations of two (28 combinations in total), three (56 combinations in total), four (70 combinations in total), five (56 combinations in total),

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
SMs ORDERING FOR THE GENERAL REGISTRATION CASE

Position in ordering	1	2	3	4	5	6	7	8
SM	MIND_{SSD}	logLR	$\text{SIFT-OCT}_{\text{SSD}}$	PC	HOPC_{SSD}	NCC	HOG_{SSD}	MI
AUC	0.6835	0.6731	0.6398	0.5942	0.6614	0.6092	0.6409	0.5907

six (28 combinations in total), seven (8 combinations in total), and eight (1 combination) single SMs, and calculated AUC for each combined SM. Overall, we have analyzed 247 combined SMs. For simplicity, in what follows, a combination of k SMs is denoted as comSMk (e.g., comSM2 for a combination of two SMs).

The contribution of every single SM to the performance of a combined SM is obviously unequal. To establish an objective ordering of the eight single SMs according to their ability to distinguish between similar and dissimilar multimodal image patches, we propose to measure the quality of an SMs ordering $i_1, i_2, \ldots, i_N$ by the cumulated sum of AUCs achieved for an increasing number of combined SMs from one SM to the maximum value of eight SMs

$$\text{AUC}_{i_1 i_2 i_3,\ldots,i_N} = \text{AUC}(\text{SM}_{i_1}) + \text{AUC}(\text{SM}_{i_1 i_2}) + \text{AUC}(\text{SM}_{i_1 i_2 i_3}) + \cdots + \text{AUC}(\text{SM}_{i_1 i_2 i_3,\ldots,i_N}). \quad (7)$$

We define the best ordering as the one found by maximizing $\text{AUC}_{i_1 i_2 i_3,\ldots,i_N}$

$$(i_{1\cdot\text{best}}, i_{2\cdot\text{best}}, i_{3\cdot\text{best}}, \ldots, i_{N\cdot\text{best}}) = \arg\max_{i_1,i_2,i_3,\ldots,i_N}(\text{AUC}_{i_1 i_2 i_3,\ldots,i_N}). \quad (8)$$

The underlying idea behind such an ordering is that the most important (informative) SM improves the most all combined SM it is included in. The second SM improves the most all combination in which the best SM was already included. Thus, by maximizing (8), we are able to find such an SMs ordering that all its subsets $\text{SM}_{i_{1\cdot\text{best}}}$, $\text{SM}_{i_{1\cdot\text{best}},i_{2\cdot\text{best}}}$,…, $\text{SM}_{i_{1\cdot\text{best}},i_{2\cdot\text{best}},i_{3\cdot\text{best}},\ldots,i_{N\cdot\text{best}}}$ are the best or close to the best combined SM for the corresponding number of SMs.

C. Combined SM Performance in General Case

We define the general multimodal case as any combination of RS modes considered in this study. The combined SM performing the best irrespective of a considered registration case is found by training an SVM classifier using measurements collected from all cases. The found ordering of the eight SMs analyzed in this work is given in Table II with the most contributing SMs starting from the left. The first observation is that feature-based SMs are better than area-based SMs except for logLR and HOPC_{SSD}. Interestingly, HOG_{SSD} has the fourth higher AUC value, but it is almost the least important SM in the ordering. To check the independence of HOG_{SSD} against other considered SMs, we have calculated the Spearman rank correlation between values of all pairs of SMs. It was found that the correlation between HOPC_{SSD} and HOG_{SSD} takes a very high value of 0.92. For the rest of SMs, correlation varies from -0.6 to 0.6, and for the selected order of SMs from -0.36 to 0.28. We conclude that HOPC_{SSD} and HOG_{SSD} are not complementary to each other and their joint usage is redundant and useless. Analysis of particular registration cases (see Section IV-D) confirms this hypothesis (either HOPC_{SSD} or HOG_{SSD} is among the least important SMs in the ordering). The only area-based SM the contribution of which is comparable to the feature-based SMs in the general multimodal case is logLR.

For the general case, ROC curves for $\text{comSMk}_{\text{order}}$ and the first five SMs in the best ordering are shown in Fig. 6. The ROCHH is formed by the first two SMs in the ordering: MIND_{SSD} and logLR. These two SMs are potentially optimal, and the rest are suboptimal. It is interesting to notice that potentially optimal SMs involve both feature- and area-based SMs that probably capture complementary aspects of multimodal image similarity.

For each combination of k SMs, we have formed the combined SM using the ordered SMs as given in Table II and have called it $\text{comSMk}_{\text{order}}$. This combined SM is then compared to all other combinations of SMs. In Table III, AUC for the $\text{comSMk}_{\text{order}}$, the difference Δ AUC between this AUC and the best AUC for k SMs and the gain of $\text{comSMk}_{\text{order}}$ over ROCHH are reported. It is seen that for all k, $\text{comSMk}_{\text{order}}$ corresponds to the best SM combination.

Overall, the gain provided by $\text{comSMk}_{\text{order}}$ over ROCHH increases with k and reaches about 3.9% when six considered SMs are combined together. Combination of three feature-based SMs and two area-based SM—MIND_{SSD}+logLR+$\text{SIFT-OCT}_{\text{SSD}}$+PC+$\text{HOPC}_{\text{SSD}}$—is responsible for a gain of 3.83%. The rest of SMs (MI, NCC, and HOG_{SSD}) add little to the performance of the combined $\text{comSM5}_{\text{order}}$. For combinations of 6…8 SMs, AUC is fluctuating about the same value and even slightly decreases due to the random nature of the k-fold cross-validation approach used for training. This means that MI, NCC, HOPC_{SSD}, and HOG_{SSD} do not complement SMs in $\text{comSM5}_{\text{order}}$. Considering the tradeoff between performance and computational complexity, combining five SMs seems to be a reasonable choice: AUC is only 0.1% less than the one reached with a combination of all SMs. Therefore, in the following, we analyze essentially the combined $\text{comSM5}_{\text{order}}$ trained for the general case.

Experimental distributions of the normalized SM values for similar and dissimilar fragments are shown in Fig. 7 for the combined $\text{comSM5}_{\text{order}}$ and the first three SMs in the ordering.

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

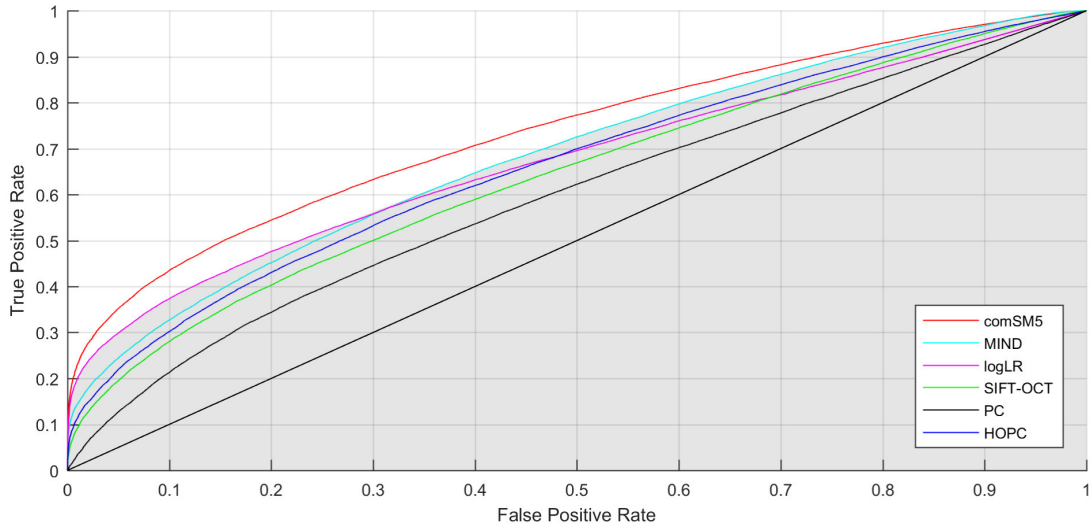

Fig. 6. ROC curves of individual SMs and the best comSM4. Line of no-discrimination is shown in all subfigures for reference. Index "SSD" specifying distance metric in structural SMs is omitted in legends for compactness. Area below ROCCH is shown in gray.

TABLE III
PARTIAL COMBINED SMS FOR THE BEST ORDERING

k	$comSMk_{order}$	AUC, %	ΔAUC, %	AUC-ROCHH, %
1	$MIND_{SSD}$	68.357	0.0	-1.082
2	$MIND_{SSD}$+logLR	71.380	0.0	1.940
3	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$	72.481	0.0	3.041
4	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$+PC	72.980	0.0	3.540
5	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$+PC+$HOPC_{SSD}$	73.277	0.0	3.837
6	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$+PC+$HOPC_{SSD}$+NCC	73.352	0.0	3.912
7	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$+PC+$HOPC_{SSD}$+NCC+HOG_{SSD}	73.284	0.0	3.844
8	$MIND_{SSD}$+logLR+$SIFT\text{-}OCT_{SSD}$+PC+$HOPC_{SSD}$+NCC+HOG_{SSD}+MI	73.220	0.0	3.780

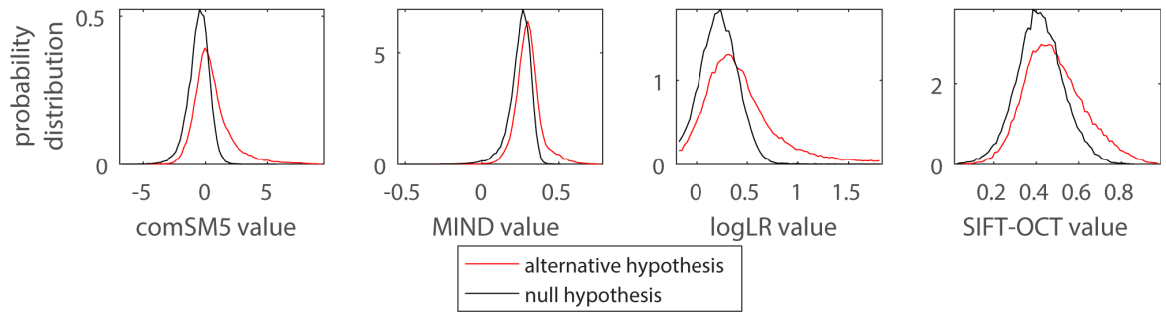

Fig. 7. Experimental distributions of the normalized SM values for similar and dissimilar fragments for $comSM5_{order}$, $MIND_{SSD}$, logLR, and $SIFT\text{-}OCT_{SSD}$.

It is seen that for $comSM5_{order}$ the distribution for the null hypothesis is close to normal (skewness is about 1.23 and kurtosis is about 6.68).

The gain of 3.8% over ROCHH is obtained with a linear SVM classifier. If the polynomial kernel of the second order is used, AUC increases additionally by 0.6%–4.4%. An additional increase of polynomial order does not improve AUC. Using LDA or QDA classifiers also shows no gain as compared to SVM.

We have also performed experiments with smaller fragments size of 13 × 13 pixels and obtained consistent results: an optimal combination of SMs includes both area-based and structural SMs. The obtained SMs ordering is slightly different but consistent with the one obtained in the 21 × 21 pixel case:

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
DIFFERENCE BETWEEN AUC FOR COMSM5$_{order}$ (TRAINED FOR THE GENERAL CASE) AND INDIVIDUAL AND COMBINED SMS FOR PARTICULAR REGISTRATION CASES, $*100\%$

Similarity measure	Registration case				
	visible-to-infrared	visible-to-radar	visible-to-DEM	radar-to-DEM	general
MIND$_{SSD}$	1.0621	9.2831	1.4402	10.0568	5.0341
logLR	8.0207	6.9449	7.8370	4.0535	6.0771
SIFT-OCT$_{SSD}$	17.8468	4.0223	7.4550	6.0704	9.4052
HOG$_{SSD}$	8.3564	11.5585	3.1737	9.5064	9.2938
PC	15.1188	13.6104	11.2642	13.2347	13.9636
MI	13.8051	12.6027	8.1365	16.0239	14.3204
HOPC$_{SSD}$	8.0760	10.1009	2.4337	7.7291	7.2481
NCC	21.4287	7.1859	9.6641	9.5947	12.4632
MIgrad	7.6431	8.2251	3.5052	9.5838	7.7991
The best comSM5$_{order}$ for particular case	-0.5213	-0.6103	-0.7455	-0.7913	0
The best comSM5$_{order}$ for general case (MIND$_{SSD}$+logLR+SIFT-OCT$_{SSD}$+PC+HOPC$_{SSD}$) trained for particular case	-0.2239	-0.6103	-0.5063	-0.7913	0
ROCHH	0.9536	3.8484	1.2051	3.2559	3.8373

logLR, MIND$_{SSD}$, HOPC$_{SSD}$, SIFT-OCT$_{SSD}$, PC, NCC, MI, and HOG$_{SSD}$. The comSM5$_{order}$ classifier derived for fragments of 21 $\times$ 21 pixels and applied to smaller fragments of 13 $\times$ 13 pixels results in a 3.7% gain with respect to the best single SM (logLR). This value is only 0.61% smaller than AUC provided by the combined SM specifically trained for fragment size of 13 $\times$ 13. Therefore, we can suggest that the results for a fragment of the size of 21 $\times$ 21 pixels may be extended to other sizes. For the sake of paper clarity and length, the results for fragments of 13 $\times$ 13 pixels are not included.

D. Comparative Analysis of the Best Combined SM, Single SMs and Combined SMs Trained for Each Particular Registration Case

Let us analyze the efficiency of the combined comSM5$_{order}$ trained for the general case in comparison to single SMs and combined SMs trained for each particular registration case. The numerical results are given in Table IV. Each cell of Table IV represents the difference of AUCs between a particular SM and comSM5$_{order}$. Note that comSM5$_{order}$ is trained for the general case and applied to particular cases (as well as to the general case).

In all cases, comSM5$_{order}$ outperforms each of the single SMs in comparison. The gain varies in a wide range from 0.4% for MIND$_{SSD}$ to 20.76% for NCC, both for the visible-to-infrared case. On average, the gain in AUC provided by comSM5$_{order}$ over single SMs is significant and takes the value of about 9%. Also, notice that the SM MIgrad used for comparison (that also represents a combined SM) in all cases performs worse than comSM5$_{order}$ (AUC is lower by 7.3% on average). These results underline a more effective behavior of the derived combined multimodal SM.

Another ongoing promising research direction for the construction of effective SMs is deep learning [21]. We could consider learned invariant feature transform (LIFT) SM [39] for the comparison with the proposed combined SM. We decided not to include LIFT results because of the following two reasons: first, LIFT is not suitable for fragments with a smaller size (applicable to 64 $\times$ 64 fragments); second, we have found LIFT performance to decrease for multitemporal and multimodal cases. Therefore, comparison with LIFT within the settings of our experiment is not fair. However, we consider deep learning as a very promising way to design more effective multimodal SMs and suggest seeing the current results as an additional benchmark for such methods. It would allow analyzing comparatively learned SMs and existing rule-based SMs, or their combinations.

It has been found that SMs ordering for the general case and particular cases differ (Table V). However, the first five SMs comprise almost the same SMs including MIND$_{SSD}$, logLR, SIFT-OCT$_{SSD}$, and PC. The HOG$_{SSD}$ and HOPC$_{SSD}$ appear to have the capability to replace each other depending on the considered registration case. The gain of the best-combined SM trained for a particular registration case (combining five SMs) and the general comSM5$_{order}$ is given in Table V. It does not exceed 0.8% and constitutes 0.66% on average. However, this gain cannot be attributed to the particular ordering of

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
SMs Ordering for General and Particular Registration Cases

case	1	2	3	4	5	6	7	8
general	$MIND_{SSD}$	logLR	SIFT-OCT_{SSD}	PC	$HOPC_{SSD}$	NCC	HOG_{SSD}	MI
visible-to-infrared	$MIND_{SSD}$	logLR	PC	HOG_{SSD}	NCC	SIFT-OCT_{SSD}	MI	$HOPC_{SSD}$
visible-to-radar	SIFT-OCT_{SSD}	$MIND_{SSD}$	logLR	$HOPC_{SSD}$	PC	MI	NCC	HOG_{SSD}
visible-to-DEM	$MIND_{SSD}$	$HOPC_{SSD}$	logLR	MI	PC	HOG_{SSD}	NCC	SIFT-OCT_{SSD}
radar-to-DEM	logLR	$HOPC_{SSD}$	SIFT-OCT_{SSD}	$MIND_{SSD}$	PC	HOG_{SSD}	NCC	MI

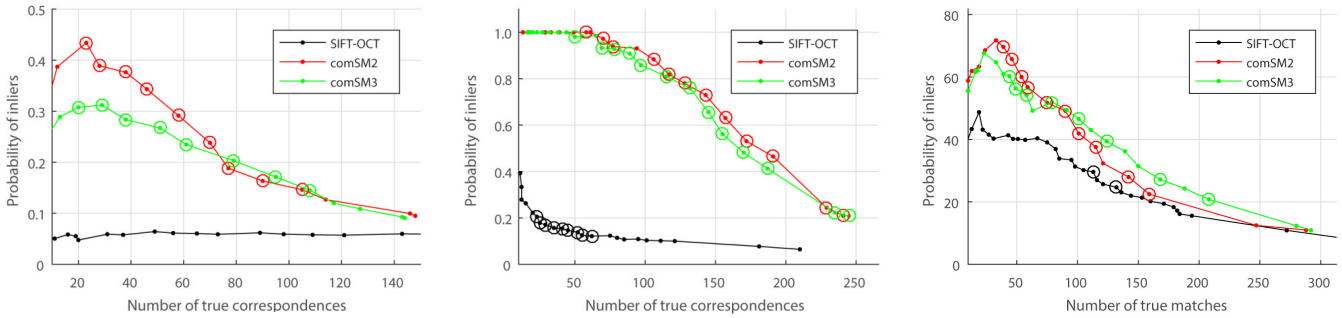

Fig. 8. Number of true correspondences versus probability of inliers for SIFT-OCT, $comSM2_{order}$, and $comSM3_{order}$ similarity metrics for different decision thresholds: the number of true correspondences versus probability of inliers for SIFT-OCT, $comSM2_{order}$, and $comSM3_{order}$ similarity metrics for different decision thresholds: (a) visible-to-DEM registration case, (b) visible-to-infrared case, and (c) visible-to-radar.

SMs optimized for a particular registration case. Considering $comSM5_{order}$ consisting of SMs optimal for the general case, almost the same gain is seen between versions trained for particular and general cases.

From this analysis, we can conclude that such a combination of existing SMs can be found that demonstrates a quite uniform advantage for the main multimodal RS registration cases including combinations of visible, infrared, radar, and DEM modalities.

E. Combined SM in Complex Multimodal Registration Case

The following experiment demonstrates the benefits provided by the combined SM in challenging visible-to-DEM [Fig. 8(a)], visible-to-infrared [Fig. 8(b)], and visible-to-radar [Fig. 8(c)] multimodal registration cases. The first experiment is conducted for test pair #10, the second one—for test pair #2 (see Table I for details), and the third one—for additional pair of Landsat8-Sentinel1 images (Landsat 8: LC08_L1TP_177025_20140509_20180526_01_T1, band #B1, (50.26°, 36.83°), May 2014; Sentinel1: S1A_IW_SLC_1SDV_20181125T034705_20181125T034735_024739_02B8B2_59AB, polarization VH, swath #2, tile #5, (49.96°, 36.17°), August 2018).

Positions of RI and TI fragments were found by applying to original images the difference of Gaussian (DoG) scale-space pyramid. We assume that the initial geometrical transformation between RI and TI restricts to a translation only. The initial translation uncertainty was set as $L_{uncert} = 30$ pixels with respect to both horizontal and vertical directions. Prior to the calculation of SM between RI and TI fragments, TI was transformed into RI coordinate system to eliminate rotation and scale differences. For each RI fragment, all TI fragments within $\pm L_{uncert}$ uncertainty range were selected. The similarity between the RI fragment and selected TI fragments was calculated using SIFT-OCT, $comSM2_{order}$ ($MIND_{SSD}$+logLR), and $comSM3_{order}$ ($MIND_{SSD}$+logLR+SIFT-OCT_{SSD}) SMs. The pair of fragments with the maximum SM value was selected as putative correspondence. All correspondences with SM exceeding a decision threshold were fed into the locally linear transforming (LLT) registration method [63] to find the nonrigid geometrical transformation between RI and TI. Parameters of LLT were set as suggested in the original article.

For all compared SMs, LLT registration was performed for a set of decision thresholds to change the balance between inliers probability and the number of found true correspondences. The results are presented in Fig. 8 where each testing point is shown by "o" marker if registration was successful and "." otherwise. By successful registration, we mean results with more than ten registered correspondences

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

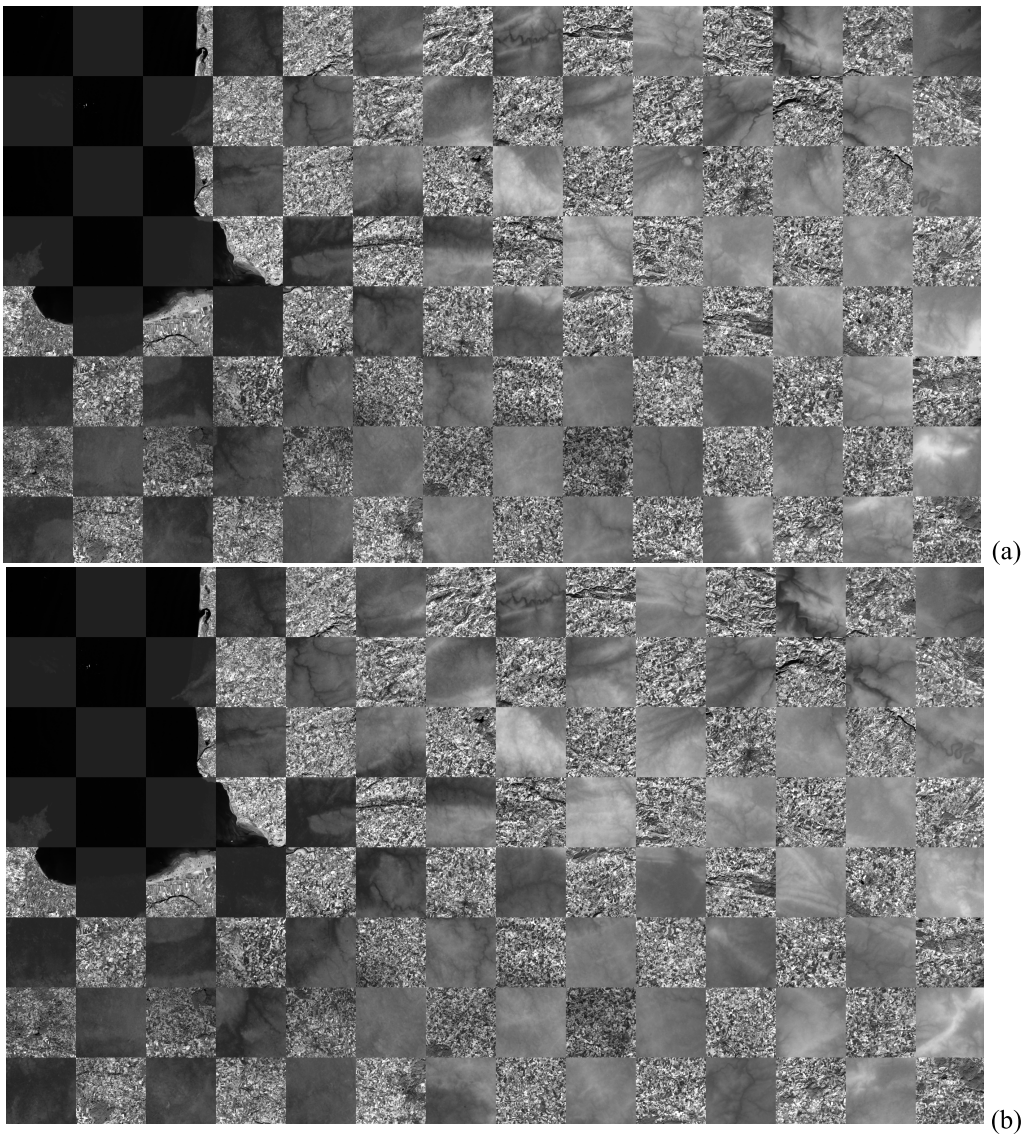

Fig. 9. Registered test pair #10. Visible-to-DEM registration case. (a) LLT method with SIFT. (b) LLT method with comSM3.

and registration RMSE less than 4 pixels. The points with successive decision thresholds are connected by straight lines.

From Fig. 8(a), it is seen that for the visible-to-DEM case, the SIFT-OCT SM provides an inliers probability of about 0.05. The LLT method fails to find a valid geometrical transform for such a low inliers probability. Both $\text{comSM2}_{\text{order}}$ and $\text{comSM3}_{\text{order}}$ perform similarly providing inliers probability from 0.1 to about 0.4. The LLT method successfully registered the tested pair of images for inliers probability exceeding 0.15 and the number of true correspondences exceeding 20. For the visible-to-infrared case [Fig. 8(b)], LLT registration with SIFT-OCT is possible for the number of true correspondences less than 60 when inliers probability exceeds 0.15. For the same number of true correspondences (60), both $\text{comSM2}_{\text{order}}$ and $\text{comSM3}_{\text{order}}$ provide an inliers probability of about 0.95. With $\text{comSM2}_{\text{order}}$ and $\text{comSM3}_{\text{order}}$, it is possible to find a significantly higher number of true correspondences: LLT successfully finds about 200 correspondences at a point corresponding to 250 true correspondences and inliers probability of 0.2. For the visible-to-radar case [Fig. 8(c)], using comSM2 and comSM3 LLT successfully registers images in a wider range of conditions as compared to SIFT. As in the previous case, LLT with comSM3 finds a larger number of true correspondences.

To better illustrate the complexity of the considered registration cases, Fig. 9 demonstrates registration results for the visible-to-DEM case in checkerboard representation: failed registration by LLT method using SIFT SM in Fig. 9(a) and successful registration by LLT using comSM3 SM in Fig. 9(b). To assist visual interpretation of the registration result, Fig. 10 shows a close up view of several representative fragments of Fig. 9(b). Correct alignment of structures in visible and DEM images is seen. Fig. 11 shows a successful registration result obtained with comSM3 for the visible-to-radar case.

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

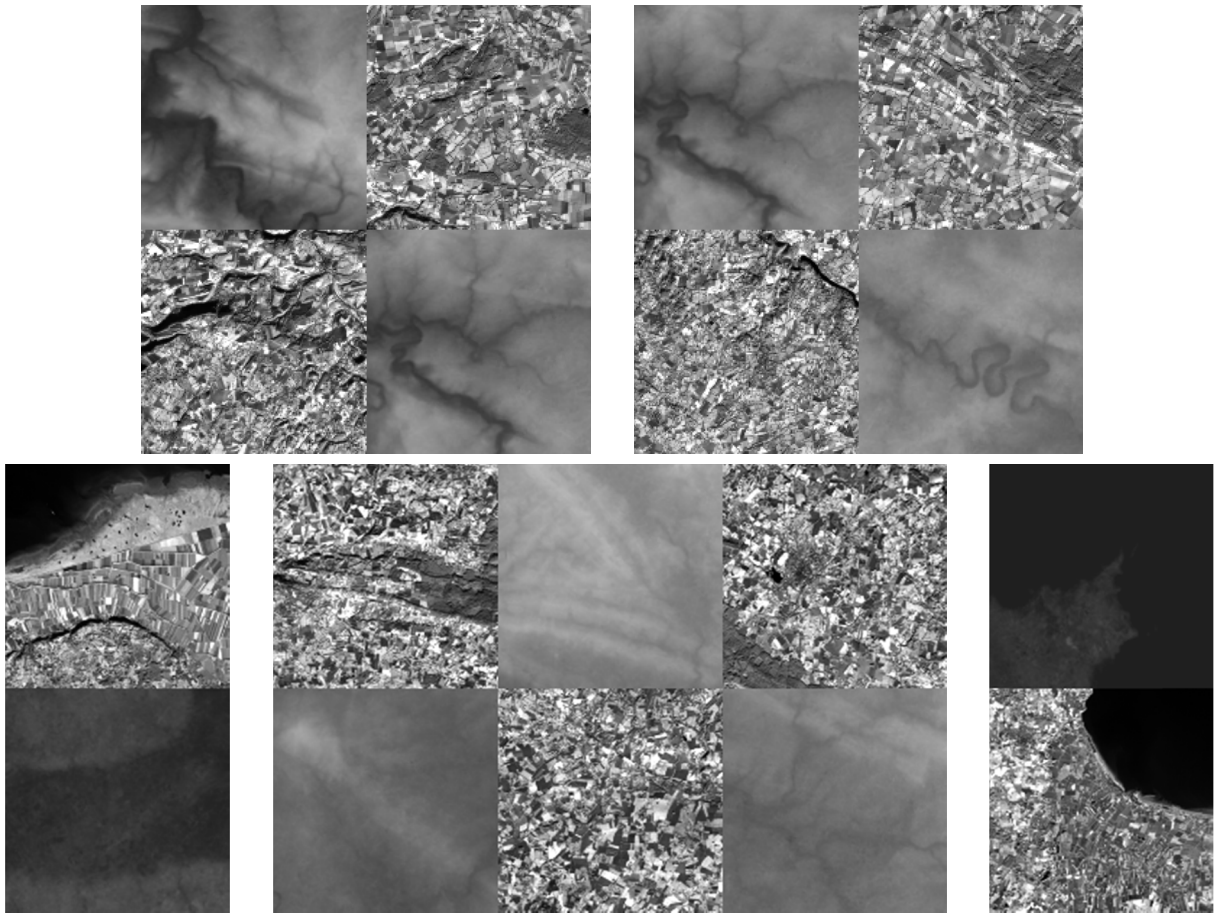

Fig. 10. Enlarged fragments of test image pair #10 from Fig. 9 registered with the LLT method using comSM3 SM.

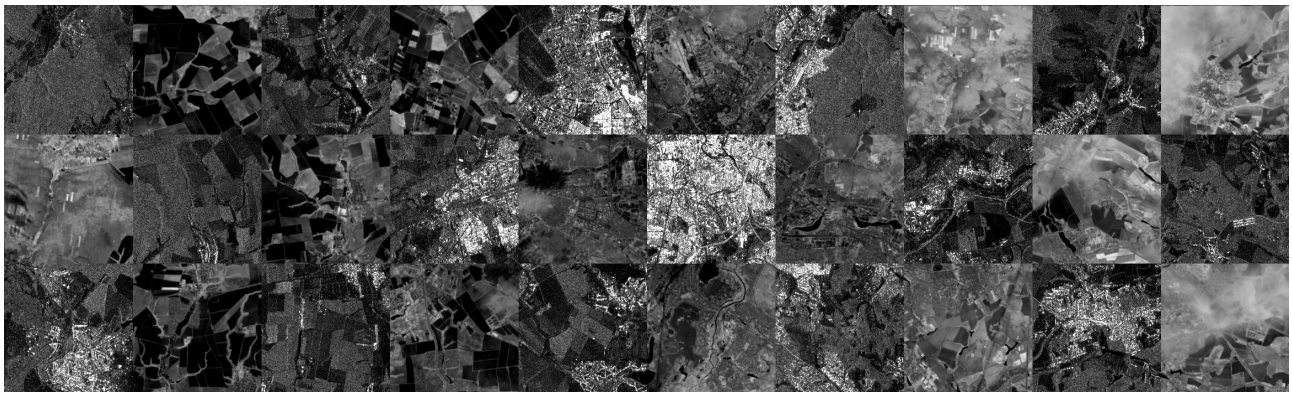

Fig. 11. Registered pair of images for visible-to-radar case: the LLT method with comSM3.

These registration examples demonstrate that for complex cases, when a single SM may not be effective enough, a suitable combined use of several SMs could be decisive for successful image registration.

V. Conclusion

The problem of measuring local similarity between multimodal pairs of RS images is considered in this article. With many available multimodal SMs, both classical and recently proposed ones, and the absence of one SM with predominant performance, we have investigated the possibility of deriving more powerful SMs by combining two or more SMs.

The combined SMs have been obtained by training a linear SVM classifier of a feature vector formed by the stacking values of several SMs. For training, we have used sixteen pairs of the registered multimodal RS images representing four registration cases: visible-to-infrared, visible-to-radar, visible-to-DEM, and radar-to-DEM ones.

Our main findings can be summarized as follows. Existing multimodal SMs are complementary for discriminating similar and dissimilar pairs of different modalities. We have proposed a method to order SMs according to their importance for this task. The three most important SMs are MIND_{SSD}, logLR, and $\text{SIFT-OCT}_{\text{SSD}}$. The three least important among the considered

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ones are PC, MI, and NCC. Two SMs, HOG_{SSD} and $HOPC_{SSD}$, are interchangeable. The combined SMs that include five existing SMs have significantly higher discriminative power than existing single SMs in general as well as in particular registration cases: for combined SM AUC increases from 1% to 21% as compared to single SMs. The advantage of the combined SM can be determinant for the successful registration of complex multimodal images, especially with complex geometrical transformations and large initial registration errors.

Apart from deriving a more powerful combined multimodal SM for RS applications, our findings have wider implications. They indicate that, for the considered set of RS images, the existing multimodal SMs are complementary in the sense they capture different aspects of similarity between different modes. Area-based approaches with rigorous modeling of noise affecting RI/TI modes provide high discriminative power, but it is increasingly more difficult to apply them for non-linear and nonfunctional intensity transformations between modes. On the contrary, structural SMs are better suited for complex transformations of mode intensities, but they have less discriminative power because of this flexibility (trend to find a higher number of false similarities between modes). A combination of these two approaches may in the future lead to more powerful multimodal SMs.

REFERENCES

[1] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.

[2] J. Le Moigne, N. S. Netanyahu, and R. D. Eastman, *Image Registration for Remote Sensing*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[3] J. M. Murphy, J. Le Moigne, and D. J. Harding, "Automatic image registration of multimodal remotely sensed data with global shearlet features," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1685–1704, Mar. 2016.

[4] J. Antoine Maintz, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, Mar. 1998.

[5] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.

[6] W. Rui and L. Minglu, "An overview of medical image registration," in *Proc. 5th Int. Conf. Comput. Intell. Multimedia Appl. (ICCIMA)*, Sep. 2003, pp. 385–390.

[7] K. Modin, A. Nachman, and L. Rondi, "A multiscale theory for image registration and nonlinear inverse problems," *Adv. Math.*, vol. 346, pp. 1009–1066, Apr. 2019.

[8] R. Szeliski, "Image alignment and stitching: A tutorial," *FNT Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.

[9] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[10] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015.

[11] A. Goshtasby and J. Le Moign, *Image Registration: Principles, Tools and Methods*. London, U.K.: Springer-Verlag, 2012.

[12] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cambridge, MA, USA: Springer, 1998, pp. 1115–1124.

[13] S. Suri and P. Reinartz, "Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 939–949, Feb. 2010.

[14] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 188–200, Mar. 2002.

[15] M. Uss, B. Vozel, V. Lukin, and K. Chehdi, "Statistical power of intensity- and feature-based similarity measures for registration of multimodal remote sensing images," *Proc. SPIE*, vol. 10004, pp. 1000403–1000414, Oct. 2016.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[17] S. Suri, P. Schwind, J. Uhl, and P. Reinartz, "Modifications in the SIFT operator for effective SAR image matching," *Int. J. Image Data Fusion*, vol. 1, no. 3, pp. 243–256, Sep. 2010.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005, pp. 886–893.

[19] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[20] M. P. Heinrich *et al.*, "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, Oct. 2012.

[21] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Comput. Methods Biomech. Biomed. Eng., Imag. Visualizat.*, vol. 6, no. 3, pp. 248–252, May 2018.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[23] X. Zhang, X. Li, Y. Feng, and Z. Liu, "The use of ROC and AUC in the validation of objective image fusion evaluation metrics," *Signal Process.*, vol. 115, pp. 38–48, Oct. 2015.

[24] J. Inglada, "Similarity measures for multisensor remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jun. 2002, pp. 104–106.

[25] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[26] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[27] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, Sep. 1991.

[28] M. Yaman and S. Kalkan, "Performance evaluation of similarity measures for dense multimodal stereovision," *J. Electron. Imag.*, vol. 25, no. 3, Jun. 2016, Art. no. 033013.

[29] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[30] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 510–517.

[31] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. ECCV*, Berlin, Germany, 2010, pp. 778–792.

[32] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. ECCV*, Berlin, Germany, 1994, pp. 151–158.

[33] C. Wachinger and N. Navab, "Entropy and Laplacian images: Structural representations for multi-modal registration," *Med. Image Anal.*, vol. 16, no. 1, pp. 1–17, Jan. 2012.

[34] P. Kovesi, "Image features from phase congruency," *J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.

[35] J. Le Moigne, W. J. Campbell, and R. F. Cromp, "An automated parallel image registration technique based on the correlation of wavelet features," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 8, pp. 1849–1864, Aug. 2002.

[36] I. Zavorin and J. Le Moigne, "Use of multiresolution wavelet feature pyramids for automatic registration of multisensor imagery," *IEEE Trans. Image Process.*, vol. 14, no. 6, pp. 770–782, Jun. 2005.

[37] J. M. Murphy and J. Le Moigne, "Shearlet features for registration of remotely sensed multitemporal images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1084–1087.

[38] J. M. Murphy, O. N. Leija, and J. Le Moigne, "Agile multi-scale decompositions for automatic image registration," *Proc. SPIE*, vol. 9840, May 2016, Art. no. 984011.

[39] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. ECCV*, Heidelberg, Germany, 2016, pp. 467–483.

ACCEPTED MANUSCRIPT

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[40] M. Uss, B. Vozel, V. Lukin, and K. Chehdi, "Efficient discrimination and localization of multimodal remote sensing images using CNN-based prediction of localization uncertainty," *Remote Sens.*, vol. 12, no. 4, p. 703, Feb. 2020.
[41] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Med. Image Anal.*, vol. 52, pp. 128–143, Feb. 2019.
[42] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–814, 2000.
[43] F. Barrera Campo, F. Lumbreras Ruiz, and A. D. Sappa, "Multimodal stereo vision system: 3D data extraction and algorithm evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 437–446, Sep. 2012.
[44] T. Chye Cheah, S. Anandan Shanmugam, and K. A. L. Mann, "Medical image registration by maximizing mutual information based on combination of intensity and gradient information," in *Proc. Int. Conf. Biomed. Eng. (ICoBE)*, Feb. 2012, pp. 368–372.
[45] A. Anthony and O. Lofffeld, "Image registration using a combination of mutual information and spatial information," in *Proc. 2006 IEEE Int. Symp. Geosci. Remote Sens.*, Jul. 2006, pp. 4012–4016.
[46] M. Mellor and M. Brady, "Phase mutual information as a similarity measure for registration," *Med. Image Anal.*, vol. 9, no. 4, pp. 330–343, Aug. 2005.
[47] Z. Sun and L. A. Ray, "Multimodal image registration based on compound mutual information," *Proc. SPIE*, vol. 5747, pp. 1274–1282, Apr. 2005.
[48] R. Feng, Q. Du, X. Li, and H. Shen, "Robust registration for remote sensing images by combining and localizing feature- and area-based methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 15–26, May 2019.
[49] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, pp. 1–39, Feb. 2010.
[50] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.
[51] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), 2nd ed. New York, NY, USA: Springer-Verlag, 2009.
[52] M. Kendall and A. Stuart, *The Advanced Theory of Statistics: Distribution Theory*, vol. 1, 4th ed. London, U.K.: Griffin, 1977.
[53] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, vol. 544. Hoboken, NJ, USA: Wiley, 2004.
[54] J. S. Pearlman, P. S. Barry, C. C. Segal, J. Shepanski, D. Beiso, and S. L. Carman, "Hyperion, a space-based imaging spectrometer," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 6, pp. 1160–1173, Jun. 2003.
[55] D. P. Roy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, Apr. 2014.
[56] M. Drusch *et al.*, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.
[57] T. Tachikawa, M. Hato, M. Kaku, and A. Iwasaki, "Characteristics of ASTER GDEM version 2," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 3657–3660.
[58] H. Fujisada, M. Urai, and A. Iwasaki, "Technical methodology for ASTER Global DEM," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3725–3736, Oct. 2012.
[59] Japan Aerospace Exploration Agency (JAXA). *ALOS Global Digital Surface Model 'ALOS World 3D—30m' (AW3D30)*. Accessed: May 9, 2020. [Online]. Available: http://www.eorc.jaxa.jp/ALOS/en/aw3d30/
[60] R. L. Jordan, B. L. Huneycutt, and M. Werner, "The SIR-C/X-SAR synthetic aperture radar system," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 4, pp. 829–839, Jul. 1995.
[61] R. Torres *et al.*, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, May 2012.
[62] M. L. Uss, B. Vozel, V. V. Lukin, and K. Chehdi, "Multimodal remote sensing image registration with accuracy estimation at local and global scales," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 11, pp. 6587–6605, Nov. 2016. [Online]. Available: https://www.researchgate.net/publication/306917490_Registered_Reference_and_Template_Image_Pairs
[63] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

ACCEPTED MANUSCRIPT