Pattern Recognition 111 (2021) 107645

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Semisupervised charting for spectral multimodal manifold learning and alignment

Ali Pournemat[a], Peyman Adibi[a,*], Jocelyn Chanussot[b]

[a] Artificial Intelligence Department, Faculty of Computer Engineering, University of Isfahan, Iran
[b] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

ARTICLE INFO

Article history:
Received 30 April 2020
Revised 17 August 2020
Accepted 6 September 2020
Available online 14 September 2020

Keywords:
Semi-supervised learning
Multimodal data
Functional map
Manifold learning
Data fusion
Hyperspectral images

ABSTRACT

For one given scene, multimodal data are acquired from multiple sensors. They share some similarities across the sensor types (redundant part of the information, also called coupling part) and they also provide modality-specific information (dissimilarities across the sensors, also called decoupling part). Additional critical knowledge about the scene can hence be extracted, which is not extractable from each modality alone. For the processing of multimodal data, we propose in this paper a model to simultaneously learn the underlying low-dimensional manifold in each modality, and locally align these manifolds across different modalities. For each pair of modalities we first build a common manifold that represents the corresponding (redundant) part of information, ignoring non-corresponding (modality specific) parts. We propose a semi-supervised learning model, using a limited amount of prior knowledge about the coupling and decoupling components of the different modalities. We propose a localized version of Laplacian eigenmaps technique specifically designed to handle multimodal manifold learning, in which the ideas of local patching of the manifolds, also known as manifold charting, is combined with the joint spectral analysis of the graph Laplacians of the different modalities. The limited given supervised information is then extending on the manifold of each modality. The idea of functional mapping is finally used to align the different manifolds across modalities. The evaluation of the proposed model using synthetic and real-world multimodal problems shows promising results, compared to several related techniques.

1. Introduction

Processing of multimodal data is of critical importance [1-3]. This idea, also called information fusion, has proved its advantages, especially in pattern recognition applications [4-9]. Taking advantage of the diversity of information provided by multiple data modalities, new data representation knowledge can be extracted which is not accessible from each modality separately. In addition, more precise and robust decision-making is expected in an information fusion scenario. However, this requires an accurate understanding of how information is structured and organized within and between the different modalities. A very promising and efficient approach for high-dimensional data representation is manifold learning, which by coping with the curse of dimensionality, improves the final robustness and accuracy of pattern recognition systems [10,11]. Manifold modeling specifically designed for multimodal data offers an appealing, but challenging, framework to achieve the above mentioned information fusion goals.

As an example, consider multiple image types from one scene captured by a hyperspectral sensor, a multispectral one, a digital RGB camera, and a LiDAR imaging system. In this multimodal problem, we can learn underlying manifolds of LiDAR and hyperspectral data and align them with that of the color image data, such that information of heights and frequency bands help to segment regions with different class labels but the same color, which is not possible having only RGB modality [12-15].

In this paper, a new semisupervised multimodal learning method is proposed, that given minimal prior knowledge about corresponding and non-corresponding parts of the data in different modalities, first propagates this knowledge throughout all remaining data samples in all modalities. Then, the underlying manifolds in all modalities are simultaneously learned by a new idea which leverages propagated non-corresponding pairs of samples as well as the corresponding ones.

The proposed idea is motivated as follows. When the manifolds are simultaneously learned (aligned) in all modalities, meaningful

* Corresponding author.
E-mail addresses: ali.pournemat@eng.ui.ac.ir (A. Pournemat), adibi@eng.ui.ac.ir (P. Adibi), jocelyn.chanussot@gipsa-lab.grenoble-inp.fr (J. Chanussot).

https://doi.org/10.1016/j.patcog.2020.107645
0031-3203/

Table 1
Notations used in the paper.

Notation	Definition
$\mathbf{x}_t^i$	the t'th data sample in the i'th modality
N_i	number of data samples in the i'th modality
N_{ij}	number of data samples in modalities i and j ($N_{ij} = N_i + N_j$)
M	number of modalities
d_i	manifold dimensionality of the i'th modality
d_i^x	maximum eigenvectors of Laplacians computed for modality i after initial information propagation
n_c, n_d	number of initially given coupling and decoupling pairs
q_c, q_d	number of coupling and decoupling pairs after initial information propagation
$\mathbf{f}_l^{i,j}$, $\mathbf{g}_l^{i,j}$	binary vectors in $\mathbb{R}^{N_i \times 1}$ indicating l'th coupling or decoupling point, respectively, in modality i w.r.t. modality j
$\mathbf{W}_i$	adjacency matrix of data kNN graph in modality i
$\tilde{\mathbf{U}}_i$	matrix with d_i^x lowest frequency eigenvectors of Laplacian of the graph represented by $\mathbf{W}_i$ as columns
$\mathbf{T}_{j,i}^C$ $\mathbf{T}_{j,i}^D$	matrices for coupling and decoupling functional mapping, respectively, from modality i to j
$\mathbf{B}_{i,j}$	matrix whose elements represent the amount of coupling or decoupling between two points of modalities i and j
$\mathbf{C}_{i,j}\mathbf{D}_{i,j}$	matrices showing the amount of coupling or decoupling of each edge of neighborhood graph, respectively, in modality i w.r.t. modality j
$\tilde{\mathbf{W}}_{i,j}$	modified adjacency matrix of data kNN graph in modality i by coupling and decoupling information w.r.t. modality j
$\mathbf{W}_{i,j}$	adjacency matrix of the modified joint within modality neighborhood graph of modalities i and j
$\mathbf{H}_{i,j}$	adjacency matrix of between modality bipartite neighborhood graph of modalities i and j
$\tilde{\mathbf{U}}_{i,j}$	matrix with columns obtained by simultaneously diagonalization of Laplacians of graphs represented by $\mathbf{H}_{i,j}$ and $\mathbf{W}_{i,j}$.
$\boldsymbol{d}_{ij}$	manifold dimensionality of the coupled part of modalities i and j
$\tilde{d}_{ij}$	a number with the condition $d_{ij} < \tilde{d}_{ij} \leq N_{ij}$

features in the common latent space of the problem are detected, which is expected to be non-empty because the same object is sensed in all modalities: they hence likely share part of the information. Also, one would expect to detect specific features for each modality, since the different modalities have their own specific characteristics and are not fully redundant. They hence offer a chance to solve some ambiguities and possible confusion limiting the performances when one single modality is used. However, we do not want this modality-specific information to interfere with the alignment process, which should only rely on the information that is indeed present in all modalities. In the semisupervised information propagation step, it is proposed to use the manifold charting and functional mapping ideas to adhere the assumed manifold structure in all modalities and the alignment between them, respectively. Also, in order to properly align the manifolds representing two given modalities, their decoupled (or non-corresponding) parts should be prevented from contributing to the learning of the common features. This is included in the formulation of the objective function proposed in this paper.

There are two main contributions in the proposed methods. The first contribution is the combined manifold charting and functional mapping idea proposed for information propagation, which respects the underlying data manifold in each modality. The second contribution is about preventing decoupling relationships to incorporate in the learning of the aligned manifolds. This learning is performed based on the updated (propagated) information, using a new optimization problem for joint diagonalization of a within-modality graph Laplacian and a between-modality bipartite graph Laplacian.

The remainder of this paper is organized as follows. In Section 2, a brief review about some recent related works is given. Section 3 discusses the detailed steps of the proposed algorithm, containing the proposed manifold charting and functional mapping ideas used for extending the given limited knowledge, and the new simultaneous diagonalization technique used for manifold alignments. The obtained results are given in Section 4. Section 5 concludes the paper and draws potential future research directions. The notations used throughout the paper are introduced in Table 1.

2. Related works

Multimodal learning based on data manifold representations has recently been considered in the machine learning community. Having different types of imaging equipment with different properties and specificities, many image processing, machine vision, and video analysis problems like image registration, image retrieval, image segmentation, image classification, etc. can be defined as multimodal problems. The problem of finding correspondences between 3D deformable non-rigid shapes (shape matching problem) has also been considered as a manifold alignment or multimodal manifold learning problem in computer graphics. In this domain, several techniques like finding a functional map between shapes (modalities) [16] and charting based deep learning models on non-Euclidean domains [17] such as geodesic, anisotropic, and mixture model convolutional neural networks (CNN) [18-20] have been proposed. In addition, considering the very good results recently reported for general deep learning models as well as the ones customized for dimensionality reduction [21], the idea of multimodal deep learning [22-24] can be specialized for manifold structured or geometric deep learning methods reviewed in [17] to achieve promising multimodal data analysis. In [25] a deep encoder-decoder model has been proposed, which uses attention mechanism to consider intra-modal relationships in addition to inter-modal ones in visual and textual modalities for image captioning application. In such deep models, nonlinear dimensionality reduction is automatically performed. Hierarchical deep prediction of click feature vectors on a word vocabulary from the visual features of an image has also been proposed [26]. Using the sparse constraints, this model reduces the semantic gap in visual recognition as well as one-shot and zero-shot learning applications. Recently, also a manifold regularized deep CNN based on low-rank representation, has been proposed for multimodal, multi-view, and multitask regression for face pose estimation [27].

Other visual recognition problems, such as quality control in industrial vision, multi-view action recognition, 3D robot vision, and 3D pose estimation, etc., take advantage of multimodal manifold learning, which are based on the intrinsic geometric property of the images or image patches on underlying manifolds [28-33]. The approach of [32] improves sparse coding by local similarity preserving which respects the underlying data manifolds, and integrating multi-view data, for 3D human pose estimation. A semi-supervised local patch alignment [33] is also proposed for multi-view facial animation, which applies LLE and LTSA manifold learning techniques to find intrinsic embedding.

Biological and microscopy image registration problems are generally involved with multimodal data [34], for which manifold

learning and alignment technics are applied Laplacian eigenmap technique [35] has been used to embed the image patches of three different types of magnetic resonance (MR) images as different modalities, to have proper structural representations, and then have been aligned to find the final representation [36]. This work has recently been improved to develop a multimodal manifold learning method for medical image registration application that is more robust to transformations [37]. A cross-modal manifold learning method for multimodal medical image retrieval application has also been proposed [38], in which a perturbed minimal spanning tree representation for data of each modality is found, and then a joint feature (latent) space is learnt by alignment of these representations using a limited number of corresponding instances. The required classification step for retrieval is performed in this common latent space. Manifold learning has also been applied in other types of medical multimodal data processing problems [39].

In remote sensing, learning manifolds in multimodal images captured by different types of sensors or acquisition tools, allows to achieve better image classification performances. Optimizing a joint graph Laplacian objective function, such that the geometry of underlying manifolds in each modality is respected, while maximizing similarities and minimizing dissimilarities of the modalities in a common latent space through a semisupervised paradigm has been proposed in [40]. An improved kernel based version of this method has also been developed [41], and applied on other domains like multimodal object recognition problems [42]. Similar ideas for multimodal image segmentation, having two RGB and LiDAR modalities have also been proposed in [43]. Recently, a research given in [44] tried to show that a limited amount of hyperspectral remote sensing images as highly-discriminative training data, can improve the classification of a model trained on a large amount of multispectral images as poorly-discriminative data [44]. Much higher cost of collecting hyperspectral data than that of the multispectral ones, motivates this idea. The proposed semisupervised cross-modality learnable manifold alignment (LeMA) technique, learns a joint graph from the data instead of using a fixed one, and applies a graph-based label propagation technique on this graph.

An interesting idea in multimodal manifold learning lies in the joint diagonalization of Laplacian matrices of multiple modalities. With these low dimensional obtained representations, existing information about correspondence between different modalities are respected, in addition to the manifold geometry of each modality. In a recent work, inspired by the ideas presented in [45-47], the simultaneous diagonalization of Laplacian matrices constructed from different modalities is proposed [48]. A limited number of coupling and decoupling points between modalities are used, and the problem is formulated accordingly, and successfully evaluated in multimodal object recognition and multimodal spectral clustering problems. This model was also applied in a medical domain to improve the registration quality of multi-subject functional magnetic resonance imaging (fMRI) data [49].

There are also several works related to multimodal manifold learning presented under different titles such as domain adaptation, transfer learning, and multi-view learning. For example, a recent algorithm [50] learns aligned eigen basis of the graph Laplacians in the source and target domains as two modalities, using the class labels of a sparse set of corresponding data pairs, and based on a formulation which has similarities to the one presented in [48]. They evaluated their algorithm as a domain adaptation technique for transfer learning in face recognition problems under different variations. Another work [51] proposed multi-view diffusion maps by defining a random walk process hop between different views, to learn coupled embedding of different modalities. This idea also results in a formulation having similarities with [48]. A multi-view manifold regularized classifier learning is also proposed through a vector-valued reproducing kernel Hilbert space [52]. It can also be considered as a form of multiple kernel learning (MKL) which is different from typical MKL approaches (e.g. see [53]) in considering supervised as well as semi-supervised settings, incorporating existing information about correspondence between views or modalities, and learning how to combine weights without mathematical constraints like non-negativity.

3. Proposed method

In a multimodal learning problem, given a set of corresponding and non-corresponding pairs of data samples between different modalities, the proposed multimodal manifold learning method is briefly described as follows. For any pair of modalities i and j, the given initial information about the correspondence and non-correspondence (coupling and decoupling) between the data points are first propagated around them in their data spaces. In this way, more realistically, the data regions are considered as the elements of information fusion, instead of the initial data points (section III-A). Then, an embedding for the common manifold of the two modalities are learned. To this end, a bipartite graph between two modalities is defined, and the Laplacian of this graph (between-modality graph) and the Laplacians of the modified localized graphs of the two modalities (within-modality graphs) are simultaneously diagonalized. As a result, the decoupled data regions of the two modalities are prevented to contribute in the learning of common manifold. Additionally, since the neighborhood graph of each modality is localized by extended coupling/decoupling information, the coupled parts are also emphasized in this learning (section III-B). Fig. 1, explains the steps of the proposed method. Fig. 2 also illustrates the proposed idea.

3.1. Regional initial coupling/decoupling information

We assume a semisupervised scenario, in which a little amount of information about correspondence/non-correspondence between each two modalities i and j (supervised information) is initially given, by n_c pairs of coupling vectors $(\mathbf{f}_l^{i,j}, \mathbf{f}_l^{j,i})$ with $l = 1, \cdots, n_c$ and n_d pairs of decoupling vectors $(\mathbf{g}_l^{i,j}, \mathbf{g}_l^{j,i})$ with $l = 1, \cdots, n_d$, as follows:

$$\mathbf{F}_{i,j} = \left\{ (\mathbf{f}_1^{i,j}, \mathbf{f}_1^{j,i}), \cdots, (\mathbf{f}_{n_c}^{i,j}, \mathbf{f}_{n_c}^{j,i}) \right\} \tag{1}$$

$$\mathbf{G}_{i,j} = \left\{ (\mathbf{g}_1^{i,j}, \mathbf{g}_1^{j,i}), \cdots, (\mathbf{g}_{n_d}^{i,j}, \mathbf{g}_{n_d}^{j,i}) \right\} \tag{2}$$

where $\mathbf{f}_l^{i,j}, \mathbf{g}_l^{i,j} \in \mathbb{R}^{N_i \times 1}$, $\mathbf{f}_l^{j,i}, \mathbf{g}_l^{j,i} \in \mathbb{R}^{N_j \times 1}$, N_i and N_j are the number of data samples in two modalities, and the s'th elements of coupling and decoupling vectors are defined as follows:

$$\mathbf{f}_l^{i,j}(s) = \begin{cases} \alpha & \text{if an index } u \text{ exists such that the point } \mathbf{x}_u^j \text{ in modality } j \text{ is coupled with } \mathbf{x}_s^i \text{ from modality } i, \text{ i.e. } \mathbf{f}_l^{j,i}(u) > 0 \\ 0 & \text{if there is no information about } \mathbf{x}_s^i \text{ to be coupled with modality } j \text{ in the } l'\text{th coupling pair} \end{cases} \tag{3}$$

$$\mathbf{g}_l^{i,j}(s) = \begin{cases} \beta & \text{if an index } u \text{ exists such that the point } \mathbf{x}_u^j \text{ in modality } j \text{ is decoupled with } \mathbf{x}_s^i \text{ from modality } i, \text{ i.e. } \mathbf{g}_l^{j,i}(u) > 0 \\ 0 & \text{if there is no information about } \mathbf{x}_s^i \text{ to be decoupled with modality } j \text{ in the } l'\text{th decoupling pair} \end{cases} \tag{4}$$

where $\alpha, \beta > 0$ indicate the strengths of the above coupling and decoupling relationships, respectively. If there are no prior knowledge about these strengths, we set $\alpha = \beta = 1$.

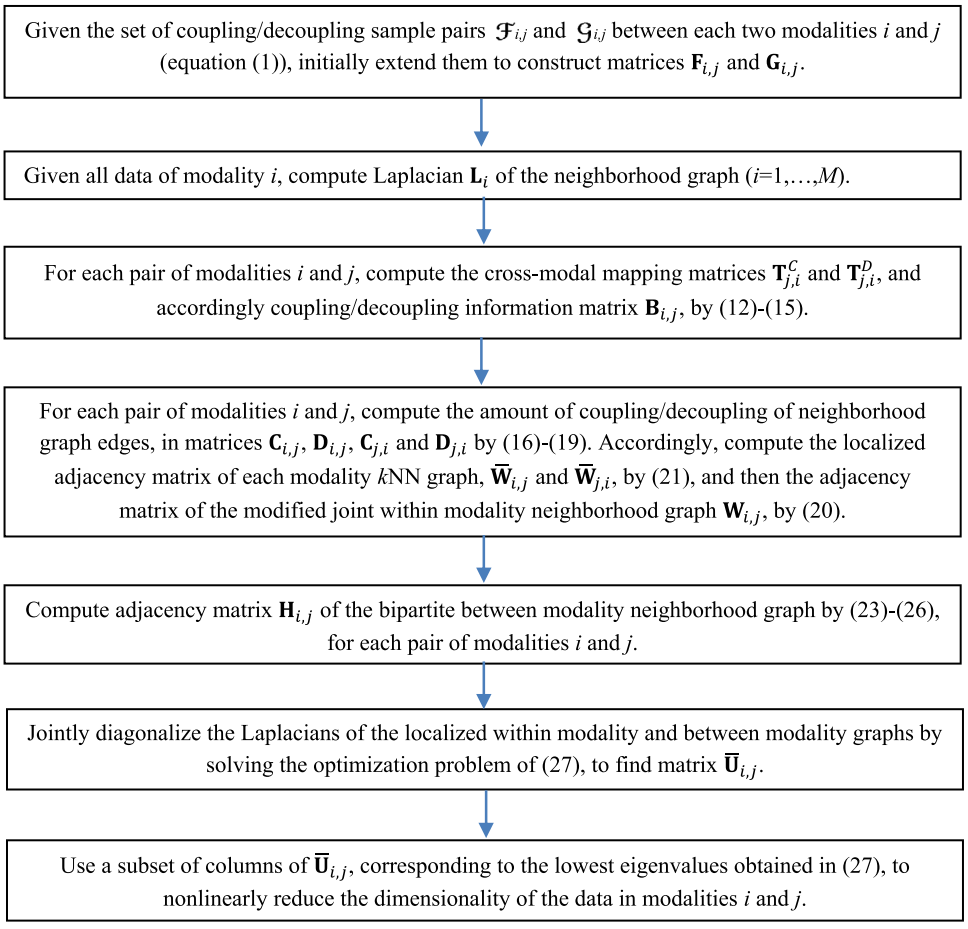

Fig. 1. The flowchart of the proposed method.

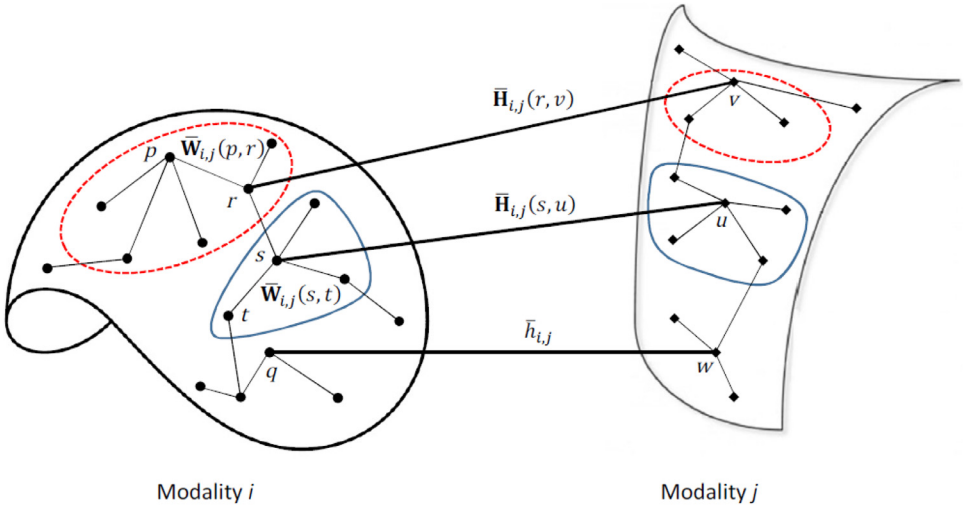

Fig. 2. The within-modality and between-modality graphs for two modalities i and j represented by a number of their edges and nodes (filled circles and diamonds show nodes of the two modalities i and j, respectively, and thin and thick lines show within-modality and between-modality edges, respectively). The manifolds of the two modalities, in addition to one coupling region (solid line) and one decoupling region (dashed line) on each manifold, are also shown. From Eq. (24), we have $\tilde{\mathbf{H}}_{i,j}(s,u) > \tilde{\mathbf{H}}_{i,j}(r,v)$ since (s, u) is an edge connecting coupling parts and (r, v) is connecting decoupling parts. The weight of the third shown between-modality edge is $\tilde{h}_{i,j}$ since $\mathbf{B}_{i,j}(q, w)$ is assumed to be 0 which means no information about coupling or decoupling between nodes q and w. Also, from Eqs. (21) and (10), we have $\bar{\mathbf{W}}_{i,j}(s,t) > \boldsymbol{W}_i^{Spec}(s,t)$ because both ends of edge (s, t) are in a coupling region, while $\bar{\mathbf{W}}_{i,j}(p,r) < \boldsymbol{W}_i^{Spec}(p,r)$ since nodes p and r are both inside a decoupling part.

As will be seen in Eq. (12), in order to the matrix of the functional mapping [16] between two modalities i and j can be obtained, a number of q_c coupling pairs and q_d decoupling pairs are required which should not be less than d_i^x, that is the intrinsic dimensionality of the data manifold of modality i (i.e. the condition $q_c, q_d \geq d_i^x$ is required). If the number of initially given coupling or decoupling pairs defined in Eqs. (1) and (2), namely n_c and n_d, are not enough (i.e. less than d_i^x), we propose to make new coupling pairs from existing ones. Since the nearest neighbors of a data point are generally assumed to inherit the local properties of that point, the additional pairs are generated based on these nearest neighbors. However, the number of nearest neighbors used for this purpose should be sufficiently small, to respect the manifold structure in each modality. The following procedure is proposed to generate the additional pairs.

At first, an additional coupling (or decoupling) pair is produced using the first nearest points of the existing ones, shown by $(\mathbf{f}_l'^{i,j}, \mathbf{f}_l'^{j,i})$'s (or $(\mathbf{g}_l'^{i,j}, \mathbf{g}_l'^{j,i})$'s). If those are not enough, another additional pair is made with the second nearest points, denoted by $(\mathbf{f}_l''^{i,j}, \mathbf{f}_l''^{j,i})$'s (or $(\mathbf{g}_l''^{i,j}, \mathbf{g}_l''^{j,i})$'s). This procedure is continued until enough coupling (or decoupling) pairs are identified. With this type of locally extending the limited given information, we also improve the point-wise coupling/decoupling representation to a regional one, which is more likely to represent the correspondence/non-correspondence, especially in real-world data spaces.

More exactly, having $(\mathbf{f}_l^{i,j}, \mathbf{f}_l^{j,i})$ the new coupling pairs $(\mathbf{f}_l'^{i,j}, \mathbf{f}_l'^{j,i})$ and $(\mathbf{f}_l''^{i,j}, \mathbf{f}_l''^{j,i})$ are defined as follows:

$$\mathbf{f}_l'^{i,j}(s) = \begin{cases} \alpha' & \text{if } \exists\, t \text{ such that } \mathbf{x}_s^i \text{ being the nearest point of} \\ & \text{modality } i \text{ to } \mathbf{x}_t^i \text{ with conditions } \mathbf{f}_l^{i,j}(s) = 0 \\ & \& \, \mathbf{f}_l^{i,j}(t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$\mathbf{f}_l'^{j,i}(u) = \begin{cases} \alpha' & \text{if } \exists\, v \text{ such that } \mathbf{x}_u^j \text{ being the nearest point of} \\ & \text{modality } j \text{ to } \mathbf{x}_v^j \text{ with conditions } \mathbf{f}_l^{j,i}(u) = 0 \\ & \& \, \mathbf{f}_l^{j,i}(v) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\mathbf{f}_l''^{i,j}(s) = \begin{cases} \alpha'' & \text{if } \exists\, t \text{ such that } \mathbf{x}_s^i \text{ being the nearest point of} \\ & \text{modality } i \text{ to } \mathbf{x}_t^i \text{ with conditions } \mathbf{f}_l^{i,j}(s) = 0 \\ & \& \, \mathbf{f}_l'^{i,j}(s) = 0 \, \& \, \mathbf{f}_l'^{i,j}(t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$\mathbf{f}_l''^{j,i}(u) = \begin{cases} \alpha'' & \text{if } \exists\, v \text{ such that } \mathbf{x}_v^j \text{ being the nearest point of} \\ & \text{modality } j \text{ to } \mathbf{x}_u^j \text{ with conditions } \mathbf{f}_l^{j,i}(u) = 0 \\ & \& \, \mathbf{f}_l'^{j,i}(u) = 0 \, \& \, \mathbf{f}_l'^{j,i}(v) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $0 < \alpha' < \alpha$ and $0 < \alpha'' < \alpha'$ indicate lower coupling strengths than that of the original given coupling pairs. Given a decoupling pair $(\mathbf{g}_l^{i,j}, \mathbf{g}_l^{j,i})$, new ones $(\mathbf{g}_l'^{i,j}, \mathbf{g}_l'^{j,i})$ and $(\mathbf{g}_l''^{i,j}, \mathbf{g}_l''^{j,i})$ can also be made in a similar manner. Then, the coupling/decoupling matrices $\mathbf{F}_{i,j} = [\mathbf{f}_1^{i,j} \ \cdots \ \mathbf{f}_{q_c}^{i,j}]$, $\mathbf{F}_{j,i} = [\mathbf{f}_1^{j,i} \ \cdots \ \mathbf{f}_{q_c}^{j,i}]$, $\mathbf{G}_{i,j} = [\mathbf{g}_1^{i,j} \ \cdots \ \mathbf{g}_{q_d}^{i,j}]$, and $\mathbf{G}_{j,i} = [\mathbf{g}_1^{j,i} \ \cdots \ \mathbf{g}_{q_d}^{j,i}]$, are built after continuous indexing of the produced vectors (i.e. $\mathbf{f}_{n_c+l}^{i,j} = \mathbf{f}_l'^{i,j}$, $\mathbf{f}_{2n_c+l}^{i,j} = \mathbf{f}_l''^{i,j}$, etc.).

In many applications, including the image processing ones, a reasonable idea to estimate the closeness of data samples is to use spatial and spectral distances, simultaneously. For example, if each data point corresponds to an image patch in an application, usually closeness of the locations of two patches (spatial similarity) in the image as well as their similarity in terms of some visual properties, such as color, texture, etc., or closeness of their projected versions on any embedding space (spectral similarity) are effective to find general similarity between those patches.

If we consider the neighborhood graph of the data in each modality as a weighted undirected graph whose nodes are data points and each edge weight reflects similarity between the corresponding two nodes, simply, we can use the weight matrix of this graph as a linear combination of the weights constituted from the above mentioned spatial and spectral metrics, in proper applications. Thus, the weight matrix of neighborhood graph in modality i in these cases is given as follows:

$$\mathbf{W_i} = \ c_1 \mathbf{W_i^{Spec}} + c_2 \mathbf{W_i^{Spat}} \tag{9}$$

where c_1 and c_2 are proper constants, and $\mathbf{W_i^{Spec}}$ and $\mathbf{W_i^{Spat}}$ are the spectral and spatial adjacency matrices of data neighborhood graph in modality i, whose entries in row s and column t can be defined based on kernel functions like standard radial basis function as follows:

$$\mathbf{W_i^{Spec}}(s,t) = \exp\left(-D_{norm-spec}^2(i,s,t)/2\sigma^2\right) \tag{10}$$

$$\mathbf{W_i^{Spat}}(s,t) = \exp\left(-D_{norm-spat}^2(i,s,t)/2\tau^2\right) \tag{11}$$

with $\sigma > 0$ and $\tau > 0$ being the kernel widths, and the normalized spectral and spatial distances used are defined as $D_{norm-spec}(i,s,t) = (\mathbf{x}_s^i - \mathbf{x}_t^i - D_{min-spec}(i))/(D_{max-spec}(i) - D_{min-spec}(i))$ and $D_{norm-spat}(i,s,t) = (spt(\mathbf{x}_s^i) - spt(\mathbf{x}_t^i) - D_{min-spat}(i))/(D_{max-spat}(i) - D_{min-spat}(i))$, respectively, where $\mathbf{x}_s^i$ is the s'th input sample in modality i, which is assumed to be a proper feature vector, $spt(\mathbf{x}_s^i)$ is the place of this sample in the spatial domain, and functions $D_{min-spec}$, $D_{max-spec}$, $D_{min-spat}$, and $D_{max-spat}$ are minimum and maximum spectral and spatial distances in the modality specified by their input argument, respectively.

Here, the symmetric normalized Laplacian matrix [54] of the data neighborhood graph in modality i, defined as $\mathbf{L_i} = \mathbf{I} - \mathbf{D_i}^{-1/2}\mathbf{W_i}\mathbf{D_i}^{-1/2}$ with $\mathbf{W}_i$ of Eq. (9) and $\mathbf{D_i}$ as the diagonal matrix of node degrees, is partially used to model the intrinsic underlying data manifold, because its first eigenvectors (corresponding to the least eigenvalues) define a nonlinear low-dimensional embedding which introduces this manifold [35].

3.2. Joint diagonalization of manifold charting localized within and between modality graph Laplacians

To cross modal extending the limited amount of given supervised information of coupling and decoupling sample pairs, we seek for linear mappings shown by matrices $\mathbf{T}_{j,i}^C$, $\mathbf{T}_{j,i}^D \in \mathbb{R}^{N_j \times N_i}$ to map a function on the manifold of modality i represented by vector $\mathbf{f}^i$ of size $N_i \times 1$, to its corresponding function in modality j represented by $\mathbf{f}^j$ with size $N_j \times 1$. We first try to find matrix $\mathbf{T}_{j,i}^C$ satisfying the relation $\mathbf{f}^j = \mathbf{T}_{j,i}^C \mathbf{f}^i$.

The eigenvectors of data neighborhood graph Laplacian can be considered as Fourier bases functions, which in this interpretation the eigenvalues play the role of frequencies [48]. Thus, the arbitrary functions $\mathbf{f}^i$ and $\mathbf{f}^j$ can be approximated using a sufficient number, denoted by d_i^x and d_j^x, of these bases as $\mathbf{f}^i \approx \sum_{l=1}^{d_i^x} a_l^i \mathbf{u}_l^i = \tilde{\mathbf{U}}_i \mathbf{a}^i$ and $\mathbf{f}^j \approx \sum_{l=1}^{d_j^x} a_l^j \mathbf{u}_l^j = \tilde{\mathbf{U}}_j \mathbf{a}^j$, where $\mathbf{a}^i = [a_1^i \ \cdots \ a_{d_i^x}^i]^T$ and $\mathbf{a}^j = [a_1^j \ \cdots \ a_{d_j^x}^j]^T$ are Fourier coefficients, and matrices $\tilde{\mathbf{U}}_i$ and $\tilde{\mathbf{U}}_j$ contain as their columns the most important eigenvectors $\mathbf{u}_l^i$'s and $\mathbf{u}_l^j$'s (corresponding to the least eigenvalues) of symmetric normalized Laplacians in two modalities i and j, respectively. Using

the orthonormal property of eigenvectors, we have $\mathbf{a}^i = \tilde{\mathbf{U}}_i^T \mathbf{f}^i$ and $\mathbf{a}^j = \tilde{\mathbf{U}}_j^T \mathbf{f}^j$.

As [16] and [48], assume that there is a linear mapping $\mathbf{C}_{i,j}^C$ of size $d_i^x \times d_j^x$ between Fourier coefficients in two modalities as $\mathbf{a}^j = \mathbf{C}_{i,j}^{C\,T} \mathbf{a}^i$, which can be rewritten as $\tilde{\mathbf{U}}_i^T \mathbf{f}^i = \mathbf{C}_{i,j}^{C\,T} \tilde{\mathbf{U}}_j^T \mathbf{f}^j$. Considering $\mathbf{f}^i$ and $\mathbf{f}^j$ play the role of the corresponding columns of coupling matrices $\mathbf{F}_{i,j}$ and $\mathbf{F}_{j,i}$, the above relation can be generalized to $\tilde{\mathbf{U}}_i^T \mathbf{F}_{i,j} = \mathbf{C}_{i,j}^{C\,T} \tilde{\mathbf{U}}_j^T \mathbf{F}_{j,i}$. By transposing both sides of this relation, and following a similar procedure for decoupling pairs with matrices $\mathbf{G}_{i,j}$ and $\mathbf{G}_{j,i}$ and the mapping matrix $\mathbf{C}_{i,j}^D$, we end up to the following equations, from which matrices $\mathbf{C}_{i,j}^C$ and $\mathbf{C}_{i,j}^D$ are obtained:

$$\mathbf{F}_{j,i}^{\ T} \tilde{\mathbf{U}}_j = \mathbf{F}_{i,j}^{\ T} \tilde{\mathbf{U}}_i \mathbf{C}_{i,j}^C \text{ and } \mathbf{G}_{j,i}^{\ T} \tilde{\mathbf{U}}_j = \mathbf{G}_{i,j}^{\ T} \tilde{\mathbf{U}}_i \mathbf{C}_{i,j}^D \tag{12}$$

As mentioned before, we need q_c and q_d to be greater than d_i^x and d_j^x so that the above systems of equations be determined and $\mathbf{C}_{i,j}^C$ and $\mathbf{C}_{i,j}^D$ can be found from them.

Now, the mapping $\mathbf{f}^j = \mathbf{T}_{j,i}^C \mathbf{f}^i$ is rewritten as $\tilde{\mathbf{U}}_j \mathbf{a}^j = \mathbf{T}_{j,i}^C \tilde{\mathbf{U}}_i \mathbf{a}^i = \tilde{\mathbf{U}}_j \mathbf{C}_{i,j}^{C\,T} \mathbf{a}^i$. From this equation and a similar procedure for decoupling, the cross modal mapping matrices are found as:

$$\mathbf{T}_{j,i}^C = \tilde{\mathbf{U}}_j \mathbf{C}_{i,j}^{C\,T} \tilde{\mathbf{U}}_i^T \text{ and } \mathbf{T}_{j,i}^D = \tilde{\mathbf{U}}_j \mathbf{C}_{i,j}^{D\,T} \tilde{\mathbf{U}}_i^T \tag{13}$$

Similarly, a decoupling pair $(\mathbf{g}^i, \mathbf{g}^j)$ can be completed as $\mathbf{g}^j = \mathbf{T}_{j,i}^D \mathbf{g}^i$. Here, the proposed semi-supervised approach learns with a little coupling/decoupling pairs given in Eqs. (1) and (2), and many other data points in each modality without any additional information. For example, in a classification problem, these pairs can be defined by using a limited number of input labeled samples (the samples with the same class labels in different modalities introduce coupling pairs and the samples with different class labels can be used to define decoupling pairs). This limited coupling/decoupling information are propagated all over the manifolds in a cross-modal manner by finding functional mappings $\mathbf{T}_{j,i}^C$ and $\mathbf{T}_{j,i}^D$ in Eq. (13). This way, the limited supervised information are extended to all other data points in a semi-supervised manner.

We first define matrix $\mathbf{B}_{i,j} \in \mathbb{R}^{N_i \times N_j}$ to contain the mapped coupling and decoupling information from all data points of modality i to j. Having the mapping matrices $\mathbf{T}_{j,i}^C$ and $\mathbf{T}_{j,i}^D$ from modality i to j, we can find the elements of $\mathbf{B}_{i,j}$ as follows. We constitute vectors $\mathbf{f}_s^i = \mathbf{g}_s^i$ containing 1 in their s'th element, and 0 elsewhere, for $s = 1, \cdots, N_i$. Then, these functions are mapped into modality j as $\mathbf{f}_s^j = \mathbf{T}_{j,i}^C \mathbf{f}_s^i$ and $\mathbf{g}_s^j = \mathbf{T}_{j,i}^D \mathbf{g}_s^i$. We expect the elements of $\mathbf{f}_s^j$ with higher magnitudes to introduce the regions in modality j which are corresponded with s'th data point in modality i, and the elements of $\mathbf{g}_s^j$ with high values to show noncorrespondent regions. Then, for u'th data point in modality j which $u = 1, \cdots, N_j$, we define the element in row s and column u of $\mathbf{B}_{i,j}$ as:

$$\mathbf{B}_{i,j}(s,u) = \mathbf{f}_s^j(u) - \mathbf{g}_s^j(u) \tag{14}$$

where the dominance of the propagated coupling and decoupling information in each point is computed. These information will finally be used to localize the graph neighborhood weights in Eq. (21). Thus, each element of $\mathbf{B}_{i,j}$ show that how much the coupling information between two particular points in modalities i and j inferred by the functional mapping is stronger than decoupling information between them. As a result, the sign of the elements of $\mathbf{B}_{i,j}$ show that there is more coupling than decoupling between the two points in different modalities or vice versa. Thus, we can conclude that positive values indicate coupling and negative values indicate decoupling, as follows:

$$\mathbf{B}_{i,j}(s,u) \begin{cases} > 0 & \text{if there is coupling between } \mathbf{x}_s^i \text{ and } \mathbf{x}_u^j \\ < 0 & \text{if there is decoupling between } \mathbf{x}_s^i \text{ and } \mathbf{x}_u^j \\ = 0 & \text{otherwise} \end{cases} \tag{15}$$

To determine how much each graph edge of modality i is involved in coupling or decoupling relations with modality j, the two matrices $\mathbf{C}_{i,j}, \mathbf{D}_{i,j} \in \mathbb{R}^{N_i \times N_i}$ are defined as follows:

$$\mathbf{C}_{i,j}(s,t) = \frac{1}{2} \left\{ \underset{\substack{u \\ \mathbf{B}_{i,j}(s,u) > 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,u)\right) + \underset{\substack{u \\ \mathbf{B}_{i,j}(t,u) > 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(t,u)\right) \right\} \tag{16}$$

$$\mathbf{D}_{i,j}(s,t) = \frac{1}{2} \left\{ \underset{\substack{u \\ \mathbf{B}_{i,j}(s,u) < 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,u)\right) + \underset{\substack{u \\ \mathbf{B}_{i,j}(t,u) < 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(t,u)\right) \right\} \tag{17}$$

Large absolute value of an entry of $\mathbf{C}_{i,j}$ or $\mathbf{D}_{i,j}$ indicates that both end points of the corresponding graph edge in modality i should be coupled or decoupled with some data points in modality j, respectively, and candidates that edge as a part of a coupling or decoupling region between two modalities. Similarly, the two matrices $\mathbf{C}_{j,i}, \mathbf{D}_{j,i} \in \mathbb{R}^{N_j \times N_j}$ are defined as follows:

$$\mathbf{C}_{j,i}(u,v) = \frac{1}{2} \left\{ \underset{\substack{s \\ \mathbf{B}_{i,j}(s,u) > 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,u)\right) + \underset{\substack{s \\ \mathbf{B}_{i,j}(s,v) > 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,v)\right) \right\} \tag{18}$$

$$\mathbf{D}_{j,i}(u,v) = \frac{1}{2} \left\{ \underset{\substack{s \\ \mathbf{B}_{i,j}(s,u) < 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,u)\right) + \underset{\substack{s \\ \mathbf{B}_{i,j}(s,v) < 0}}{\text{avg}} \left(\mathbf{B}_{i,j}(s,v)\right) \right\} \tag{19}$$

The coupling and decoupling regions on data manifolds in two modalities revealed by computing matrices in Eqs. (16) to (19), will be used to improve the learning of the joint embedding of the two modalities. An example of these regions of data manifolds of two modalities and their effect in the proposed idea are shown in Fig. 2.

A weight matrix $\mathbf{W}_{i,j} \in \mathbb{R}^{N_{ij} \times N_{ij}}$ for the joint within modality neighborhood graph on total set of data in modalities i and j, is defined as follows:

$$\mathbf{W}_{i,j} = \begin{bmatrix} \bar{\mathbf{W}}_{i,j} & 0 \\ 0 & \bar{\mathbf{W}}_{j,i} \end{bmatrix} \tag{20}$$

where zero non-diagonal blocks indicate that in this graph the neighbors of a data point in each modality can only be among the other data points in that modality, and its diagonal blocks $\bar{\mathbf{W}}_{i,j} \in \mathbb{R}^{N_i \times N_i}$ and $\bar{\mathbf{W}}_{j,i} \in \mathbb{R}^{N_j \times N_j}$ contain the weights of the data neighborhood graphs in modalities i and j, modified based on modalities j and i, respectively. We propose to modify (localize) these weights based on the coupling and decoupling information computed in Eqs. (16) to (19), to emphasize on coupling regions and deemphasize on the decoupling ones for joint embedding learning in next steps. More exactly, the entry in row s and column t of $\bar{\mathbf{W}}_{i,j}$ is de-

fined as (entries of $\bar{\mathbf{W}}_{j,i}$ are obtained similarly):

$$\bar{\mathbf{W}}_{i,j}(s,t) = \exp\left(-\left[\frac{\mathbf{x}_s^i - \mathbf{x}_t^{i\,2}}{2\sigma^2} - \beta_c \frac{\mathbf{C}_{i,j}(s,t)^2}{2\delta_c^2} - \beta_d \frac{\mathbf{D}_{i,j}(s,t)^2}{2\delta_d^2}\right]_+\right) \tag{21}$$

where parameters β_c, $\beta_d \geq 0$ indicate the importance of coupling and decoupling information obtained for the graph edges with respect to the usual similarities[1] computed for the edges in the first term of the exponent, parameters σ^2, δ_c^2, and δ_d^2 scale the distance, coupling, and decoupling information, respectively, and symbol $[\cdot]_+$ prevents meaningless nonnegative exponent, defined as follows:

$$[z]_+ = \begin{cases} z & z > 0 \\ 0 & o.w. \end{cases} \tag{22}$$

where z is an arbitrary real-valued variable.

If $\beta_c = \beta_d = 0$ the edge weight defined in Eq. (21) is reduced to a usual Gaussian similarity measure used in basic manifold learning methods [35]. The novelty of this equation when positive values are assigned to these parameters, is explained as follows. If both ends of the edge (s, t) in modality i are involved in coupling relation with modality j, its weight is modified to become larger to emphasize the role of this edge in learning the common manifold between these two modalities (common latent space), through the second term in the exponent of Eq. (21). On the other hand, if both ends of an edge in modality i are involved in decoupling relation with modality j, its weight should be decreased in order to de-emphasize (or even remove) the role of this edge in learning the common latent space, through the third term in the exponent of Eq. (21), since the entries of matrix $\mathbf{D}_{i,j}$ are non-positive based on Eq. (17).

The symmetric normalized Laplacian of the graph with adjacency matrix $\mathbf{W}_{i,j}$ is called $\mathbf{L}_{i,j}$, and the diagonal matrix containing its d_{ij} smallest eigenvalues is called $\mathbf{\Lambda}_{i,j}$. Finding a matrix $\mathbf{U}_{i,j} \in \mathbb{R}^{N_{ij} \times d_{ij}}$ that minimizes the objective function $\mathbf{U}_{i,j}^T \mathbf{L}_{i,j} \mathbf{U}_{i,j} - \mathbf{\Lambda}_{i,j}{}_F^2$ is called diagonalization of Laplacian $\mathbf{L}_{i,j}$, where $\|\cdot\|_F$ indicates Frobenius norm of a matrix. It is straight-forward to show that, diagonalization of $\mathbf{L}_{i,j}$ is equivalent to simultaneously diagonalizing the Laplacians of graphs with adjacency matrices $\bar{\mathbf{W}}_{i,j}$ and $\bar{\mathbf{W}}_{j,i}$. Thus, we can introduce a generalized version of the models previously presented in joint Laplacian diagonalization literature [46–48,50,55], since here the corresponding parts of different modalities are highlighted in joint diagonalization, and the role of their non-corresponding regions are diminished.

In addition, the coupling regions should be mapped into close areas of the embedding space, and decoupled ones to faraway places. This is performed throughout a further joint diagonalization of the previously defined Laplacian and the Laplacian of a new between-modality neighborhood graph. We propose to define a between-modality bipartite graph, with edges connecting a data point of modality i to a data point of modality j (see Fig. 2). We define the weight matrix $\mathbf{H}_{i,j} \in \mathbb{R}^{N_{ij} \times N_{ij}}$ of this bipartite graph, as follows:

$$\mathbf{H}_{i,j} = \begin{bmatrix} 0 & \bar{\mathbf{H}}_{i,j} \\ \bar{\mathbf{H}}_{i,j}^T & 0 \end{bmatrix} \tag{23}$$

where each element of matrix $\bar{\mathbf{H}}_{i,j} \in \mathbb{R}^{N_i \times N_j}$ is computed as follows:

$$\bar{\mathbf{H}}_{i,j}(s,u) = \begin{cases} \exp\left(-\mu_c \frac{\left(B_{i,j}^x - \mathbf{B}_{i,j}(s,u)\right)^2}{2\sigma_c^2}\right) & \text{if } \mathbf{B}_{i,j}(s,u) > 0 \\ \exp\left(-\mu_d \frac{\mathbf{B}_{i,j}(s,u)^2}{2\sigma_d^2}\right) & \text{if } \mathbf{B}_{i,j}(s,u) < 0 \\ \bar{h}_{i,j} & \text{if } \mathbf{B}_{i,j}(s,u) = 0 \end{cases} \tag{24}$$

where:

$$B_{i,j}^x = \max_{s,u}\left(\mathbf{B}_{i,j}(s,u)\right) \tag{25}$$

$$\bar{h}_{i,j} = \frac{1}{2}\left\{ \underset{\mathbf{B}_{i,j}(s,u) > 0}{\underset{s,u}{\text{avg}}}\left(\bar{\mathbf{H}}_{i,j}(s,u)\right) + \underset{\mathbf{B}_{i,j}(s,u) < 0}{\underset{s,u}{\text{avg}}}\left(\bar{\mathbf{H}}_{i,j}(s,u)\right) \right\} \tag{26}$$

In the bipartite between-modality graph, the edges connecting coupling points have higher weights than the edges connecting decoupled ones, since $0 \leq \bar{\mathbf{H}}_{i,j}(s,u) \leq 1$, and $\bar{\mathbf{H}}_{i,j}(s,u)$ is closer to 1 if $\mathbf{B}_{i,j}(s,u) > 0$ and it is closer to 0 if $\mathbf{B}_{i,j}(s,u) < 0$. The symmetric normalized Laplacian of this bipartite graph with adjacency matrix $\mathbf{H}_{i,j}$ is called $\mathbf{P}_{i,j}$ and the diagonal matrix containing its d_{ij} smallest eigenvalues is called $\bar{\mathbf{\Gamma}}_{i,j}$.

Now, we solve an optimization problem between two modalities i and j for learning their common underlying manifold (latent space):

$$\min_{\bar{\mathbf{U}}_{i,j}} \left\{ \left\| \bar{\mathbf{U}}_{i,j}^T \mathbf{L}_{i,j} \bar{\mathbf{U}}_{i,j} - \bar{\mathbf{\Lambda}}_{i,j} \right\|_F^2 + \left\| \bar{\mathbf{U}}_{i,j}^T \mathbf{P}_{i,j} \bar{\mathbf{U}}_{i,j} - \bar{\mathbf{\Gamma}}_{i,j} \right\|_F^2 \right\} s.t. \quad \bar{\mathbf{U}}_{i,j}^T \bar{\mathbf{U}}_{i,j} = \mathbf{I} \tag{27}$$

where $\bar{\mathbf{\Lambda}}_{i,j}$ and $\bar{\mathbf{\Gamma}}_{i,j}$ are diagonal matrices containing d_{ij} smallest eigenvalues of $\mathbf{L}_{i,j}$ and $\mathbf{P}_{i,j}$, respectively. Thus, the columns of matrix $\bar{\mathbf{U}}_{i,j}$ are the approximations of d_{ij} first eigenvectors of $\mathbf{L}_{i,j}$ and $\mathbf{P}_{i,j}$, simultaneously.

Note that in the optimization problem of (27), diagonalization of Laplacian $\mathbf{L}_{i,j}$ through the minimization of the first term, emphasizes on the local parts of the two modalities which are coupled, in learning the common manifold. Conversely, it decreases the importance of the decoupled parts in this learning. This is because matrices $\mathbf{C}_{i,j}$ and $\mathbf{D}_{i,j}$ are used to compute the weights of the graph constructed from the data points of the two modalities in Eq. (21). In addition, by diagonalization of Laplacian $\mathbf{P}_{i,j}$, a spectral clustering effect is included in learning of the low-dimensional common embedding of the two modalities, which tries to embed the coupling parts closer to each other than the decoupling ones.

Using the idea of subspace parametrization [48], which gives a number of computational advantages, the above optimization problem can be rewritten as follows:

$$\min_{\mathbf{A}_{i,j}} \left\{ \left\| \mathbf{A}_{i,j}^T \tilde{\mathbf{\Lambda}}_{i,j} \mathbf{A}_{i,j} - \bar{\mathbf{\Lambda}}_{i,j} \right\|_F^2 + \left\| \mathbf{A}_{i,j}^T \tilde{\mathbf{\Gamma}}_{i,j} \mathbf{A}_{i,j} - \bar{\mathbf{\Gamma}}_{i,j} \right\|_F^2 \right\} s.t. \quad \mathbf{A}_{i,j}^T \mathbf{A}_{i,j} = \mathbf{I} \tag{28}$$

where $\mathbf{A}_{i,j} \in \mathbb{R}^{\tilde{d}_{ij} \times d_{ij}}$ is the parameter to be optimized, with properties $\bar{\mathbf{U}}_{i,j} = \tilde{\mathbf{U}}_{i,j} \mathbf{A}_{i,j}$, with the columns of matrix $\tilde{\mathbf{U}}_{i,j}$ being approximations of the first $\tilde{d}_{ij}$ eigenvectors of $\mathbf{L}_{i,j}$ and $\mathbf{P}_{i,j}$. Also $\tilde{\mathbf{\Lambda}}_{i,j}$ and $\tilde{\mathbf{\Gamma}}_{i,j}$ are diagonal matrices containing $\tilde{d}_{ij}$ smallest eigenvalues of $\mathbf{L}_{i,j}$ and $\mathbf{P}_{i,j}$, respectively. The standard constrained optimization procedure included in MATLAB *fmincon* function is used to solve the problem given in (28). The gradient of the objective function required for this procedure, is equal to the following term:

$$4(\tilde{\mathbf{\Lambda}}_{i,j} \mathbf{A}_{i,j} \mathbf{A}_{i,j}^T \tilde{\mathbf{\Lambda}}_{i,j} \mathbf{A}_{i,j} + \tilde{\mathbf{\Gamma}}_{i,j} \mathbf{A}_{i,j} \mathbf{A}_{i,j}^T \tilde{\mathbf{\Gamma}}_{i,j} \mathbf{A}_{i,j}) - 4\left(\tilde{\mathbf{\Lambda}}_{i,j} \mathbf{A}_{i,j} \bar{\mathbf{\Lambda}}_{i,j} + \tilde{\mathbf{\Gamma}}_{i,j} \mathbf{A}_{i,j} \bar{\mathbf{\Gamma}}_{i,j}\right)$$

[1] Other similarity measures like the one shown in equations (9) to (11) can also be used in place of the Euclidean one in the first term of the exponent in equation (21).

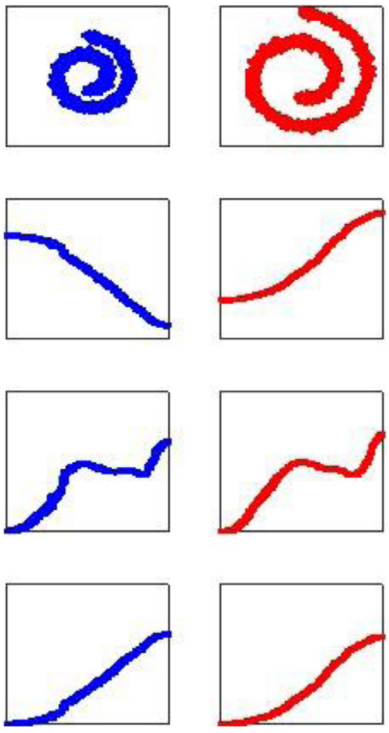

Fig. 3. First row shows the 2-D spirals dataset in two modalities. Second row shows the first related eigenvectors of the Laplacians of the graphs for two modalities (the horizontal axis is the point index which is consistent with their positions on the manifold). The third and fourth rows show diffusion and eigenmap distances of the points from the first point in two modalities.

The optimization problem of (28) can be solved alternatively using other efficient optimization techniques, like the one presented in [56].

The time complexity analysis of the overall algorithm summarized in Fig. 1, is as follows. Initial supervised information propagation step for all pairs of modalities needs $O(NnM^2)$ computations, where $N = max_i\{N_i\}$ and $n = max\{q_c - n_c, q_d - n_d\}$. Computing all Laplacians is $O(N^2M)$. Finding all matrices in Eqs. (12) to (23) is $O(N^2M^2)$. All these computations can be done offline with complexity of $O(N^2M^2)$. Finally, each iteration of the optimization process shown in Eq. (28) which constitutes the online part of the method needs $O(d\tilde{d}qM^2)$ [48], where $d = max_{i,j}\{d_{ij}\}$, $\tilde{d} = max_{i,j}\{\tilde{d}_{ij}\}$, and $q = max\{q_c, q_d\}$.

4. Experimental results

In this section one synthetic and four real-world problems are used to evaluate the proposed method. The datasets and the criteria used in the experiments are first introduced (section IV-A). Then, the details of the conducted experiments and the comparative study are given (section IV-B).

4.1. Datasets and evaluation criteria

Spirals: This dataset contains two spirals with different scales in two modalities (first row of Fig. 3), which are one-dimensional manifolds in two-dimensional input spaces, each containing 2000 points ($M = 2$, $d_1 = d_2 = 1$, and $N_1 = N_2 = 2000$).

Multimodal data of 2018 IEEE GRSS data fusion contest: A real-world semantic segmentation of multimodal remote sensing images, from the 2018 IEEE GRSS data fusion contest [57] is used. This publicly available data [58] has three modalities, containing hyperspectral images (HSI), aerial RGB images, and LiDAR images. There are seven classes of land cover in this problem. The images of different modalities have different spatial and spectral resolutions. Therefore, to make the data in different modalities compatible with each other, we use the following policy. Each data point in HSI modality corresponds to a pixel which is a 50-dimensional vector, since there are 50 frequency bands in this modality. For RGB modality, a data point is defined on a square of size 20×20, which corresponds to one pixel in HSI modality. This area of pixels given in three frequency bands (R, G, and B), by using resampling and applying a histogram, gives a 60-dimensional vector as a data point in RGB modality. In LiDAR modality, the corresponding area is a square with size 2×2 that contains grayscale pixels and makes a 4-dimensional input vector.

NUS multimodal data: A subset of NUS-WIDE dataset [48] is used for object classification. This dataset contains images represented by histograms, and their corresponding annotations represented by bag of words features. In this way, two visual and textual modalities are used in 64-dimensional and 1000-dimensional input spaces, respectively.

Office-Caltech multimodal data: This is a visual object recognition domain adaptation dataset originally introduced in [42]. There are four visual domains, from them the images of 10 common classes are selected. In this way, we have 958, 1123, 157, and 295 images from these four domains called amazon, caltech, DSLR, and webcam, respectively. Each image is represented by a 4096-dimensional feature vector extracted from a convolutional network [41].

Multimodal data of 2013 IEEE GRSS data fusion contest: The image in this multimodal datasets are captured over the University of Houston, in hyperspectral (HS) and multispectral (MS) modalities. HS data in this datasets has a spatial size of 349×1905 with 144 spectral bands covering the wavelength range from 364 nm to 1046 nm. From that, the MS data is generated with dimensions of $349 \times 1905 \times 10$. The labeled data in a region involved all kinds of classes are selected as the training set and the rest of the MS labeled data are considered as the test set. More details about this dataset are given in [59].

Evaluation criteria: Mean and standard deviation of classification accuracy are generally applied for evaluations, especially in object recognition problems. In the last set of experiments, overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) criteria used in [44] are applied to be able to compare our results with those reported in this work.

4.2. Details of the experiments and comparative study

Experimental details on spirals dataset: The synthesized spirals dataset is first used to intuitively investigate several proposed ideas like functional mapping and charting. Second row of Fig. 3 shows the eigenvectors corresponding to the first nonzero smallest eigenvalues of the Laplacians of the k-nearest neighbor (kNN) graphs in two modalities, with $k = 30$. In a kNN graph, each data point as a node is connected to its k nearest data points by edges with weights showing their similarities. For this dataset, the edge weights are computed according to Eq. (10). Since the points indices (horizontal axis in this figure) are consistent with their locations on the manifold, the monotonic increasing/decreasing behavior observed, indicates good nonlinear dimensionality reduction. Diffusion [60] and Laplacian eigenmap distances of the points from the first point of each manifold are also plotted in third and fourth rows of Fig. 3, where obviously, the Laplacian eigenmap shows better results in finding on-manifold distances than diffusion map.

To evaluate the performance of the multimodal manifold charting and functional mapping ideas, only coupling supervised infor-

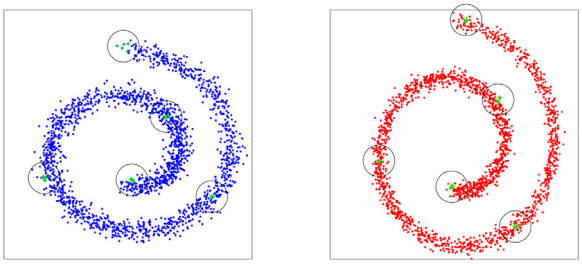

Fig. 4. The points shown by '+' are initially defined $n_c = 25$ coupling points between two manifolds in 5 almost equally distant regions shown by circles.

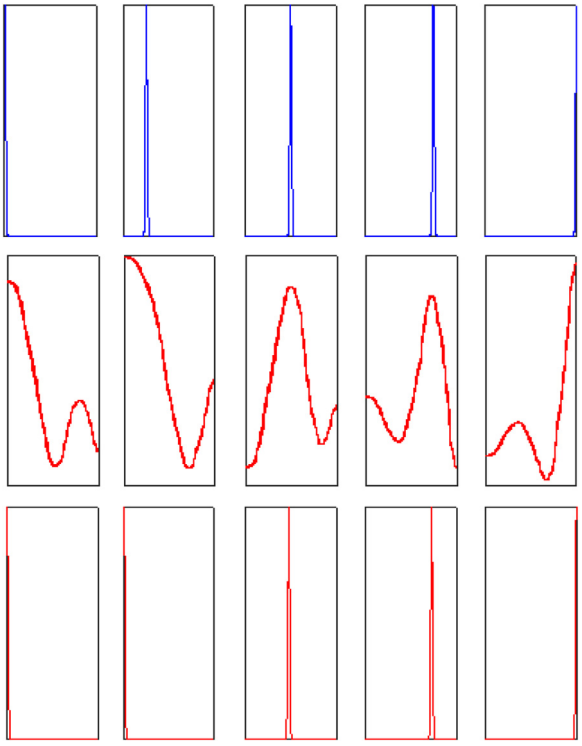

Fig. 5. First row shows several Gaussian functions centered at different data points as $\mathbf{f}^i$'s. Second row shows the corresponding mapped functions $\mathbf{f}^j$'s found by a mapping $\mathbf{T}^C_{j,i}$ computed having the coupling points of Fig 4. Third row is the post processed version of the second row (the horizontal axis is the point index which is consistent with their positions on the manifold).

mation are used. We defined 5 almost equally distant corresponding regions on the two manifolds, and 5 coupling points on each region, as shown in Fig 4, to make matrices $\mathbf{F}_{i,j}$ and $\mathbf{F}_{j,i}$. We set $i = 1$, $j = 2$, and $q_c = n_c$ which means no initial coupling propagation is required. Then, matrices $\tilde{\mathbf{U}}_i$, $\tilde{\mathbf{U}}_j$ (with $d_i^x = d_j^x = 3$), $\mathbf{C}^C_{i,j}$, and $\mathbf{T}^C_{j,i}$ are computed using Eqs. (12) and (13). Each function defined on points in modality i shown by a vector $\mathbf{f}^i$ can then be mapped to its corresponding function in modality j as $\mathbf{f}^j = \mathbf{T}^C_{j,i}\mathbf{f}^i$.

The first row of Fig 5 shows several narrow Gaussian functions as $\mathbf{f}^i$'s and the second row shows their mapped vectors $\mathbf{f}^j$'s. Exact Gaussian form cannot be reconstructed through this mapping, but the peaks of the mapped functions in modality j are more or less at the same positions as those of their corresponding functions in modality i. The third row of Fig 5 shows the results of a post processing step on the mapped functions in which all mapped functions are replaced with Gaussian functions centered at their peaks. Also, the left part of Fig 6 shows matrix $\mathbf{B}$ of Eq. (14), and the right part is obtained by replacing $\mathbf{f}^j$'s with the post processed Gaussian

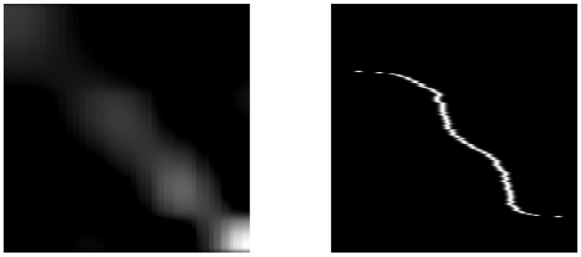

Fig. 6. Matrix $\mathbf{B}_{i,j}$ shown as image, before (left) and after (right) post processing, having the coupling points of Fig 4.

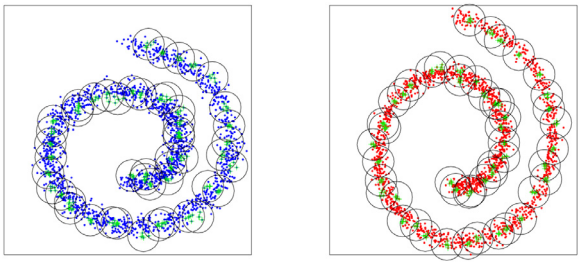

Fig. 7. The points shown by '+' are initially defined 250 coupling points between two manifolds in 50 almost equally distant regions shown by circles.

functions as mentioned above for third row of Fig 5. The results make sense but are not desirable, though.

To study the effect of the number of Fourier coefficients used in Eq. (12) to constitute matrices $\tilde{\mathbf{U}}_i$ and $\tilde{\mathbf{U}}_j$, and the number of used coupling points, the experiments are repeated with $d_i^x = d_j^x = 29$, and with 50 almost equally distant corresponding regions on the two manifolds, each containing 5 coupling points, as shown in Fig 6. Figs 8 and 9 show the results of the experiments explained about Figs 5 and 6, with this new setting. Now the results are considerably better, which underlines the requirement to have more Fourier coefficients and many corresponding pairs to ensure good results. The first requirement is easily fulfilled, while the limitation of initial supervised information should be relaxed to meet the second one.

In this purpose, the charting idea is used to propagate initial information to neighboring points in an incremental manner. We start with the same amount of supervised information shown in Fig 4. Then, these information are extended using a radially symmetric patch around each coupling point. The radius of the patch is increased step by step and the results are observed in each step. The first two columns of Fig 10 show the coupling points after several steps, and the other two columns show the resulting matrix $\mathbf{B}_{i,j}$ in original and post processed formats, respectively, corresponding to this situation. We observe good results in this experiment (the last row of Fig 10) with the supposed minimal amount of the supervised information.

Experimental details of multimodal object recognition on Office-Caltech and NUS datasets: To evaluate the proposed method on the previously introduced Office-Caltech dataset, several pairs of four existing domains are selected (the first is considered as the source and the second is used as the target domain). The recognition accuracy on the target domain of the proposed method is compared with that of the following domain adaptation techniques: MultiNMF [61], Unsupervised GFK [42], OT-lab [62], JDA [63], and KEMA [41]. Each method is trained on the corresponding source domain in each experiment. Results reported in Table 2, show good performances and assess the superiority of the proposed method in most of the tested problems, such that in 37% of the experiments it has the best results and in 63% of the exper-

Table 2
Mean and Stand Deviation of Classification Accuracies in 50 runs on Office-Caltech Domain Adaptation Visual Object Recognition Study.

Modalities	Methods							
	CD pos [48]	CD pos & neg [48]	Proposed method	MultiNMF [61]	Unsup. GFK [42]	OT-lab [62]	JDA [54]	KEMA [41]
A - C	92.9 ± 2.6	**93.0 ± 2.5**	89.5 ± 1.7	22.8 ± 0.1	80.2 ± 1.9	87.2 ± 1.2	82.6 ± 2.9	80.3 ± 3.4
C - A	90.2 ± 2.5	92.8 ± 0.2	**93.5 ± 1.8**	22.8 ± 0.1	87.8 ± 2.1	92.1 ± 1.3	89.6 ± 2.0	91.5 ± 1.5
C - D	75.0 ± 5.3	72.1 ± 2.8	89.3 ± 2.7	32.8 ± 0.0	83.5 ± 3.6	85.4 ± 6.0	85.0 ± 4.9	**93.6 ± 3.1**
A - W	89.8 ± 3.6	85.9 ± 3.6	91.5 ± 2.5	32.1 ± 0.0	78.0 ± 4.8	84.5 ± 2.4	83.0 ± 4.6	**92.7 ± 2.5**
W - C	75.5 ± 3.2	71.1 ± 3.1	**84.7 ± 2.3**	28.3 ± 5.8	75.1 ± 2.5	83.7 ± 1.5	79.8 ± 2.0	82.1 ± 2.0
W - A	90.6 ± 3.1	86.8 ± 3.1	**91.9 ± 2.4**	28.4 ± 5.7	81.2 ± 2.2	91.9 ± 1.4	90.9 ± 1.2	91.6 ± 1.3
D - A	87.5 ± 2.8	82.2 ± 3.2	89.9 ± 1.8	26.5 ± 6.7	85.4 ± 2.1	**92.9 ± 1.1**	91.9 ± 0.8	90.3 ± 1.1
D - W	70.6 ± 4.0	93.3 ± 3.6	91.1 ± 1.9	27.7 ± 6.6	96.7 ± 1.9	94.1 ± 3.4	**97.0 ± 1.5**	91.0 ± 3.5

A: Amazon, C: Caltech, D: DSLR, W: Webcam.

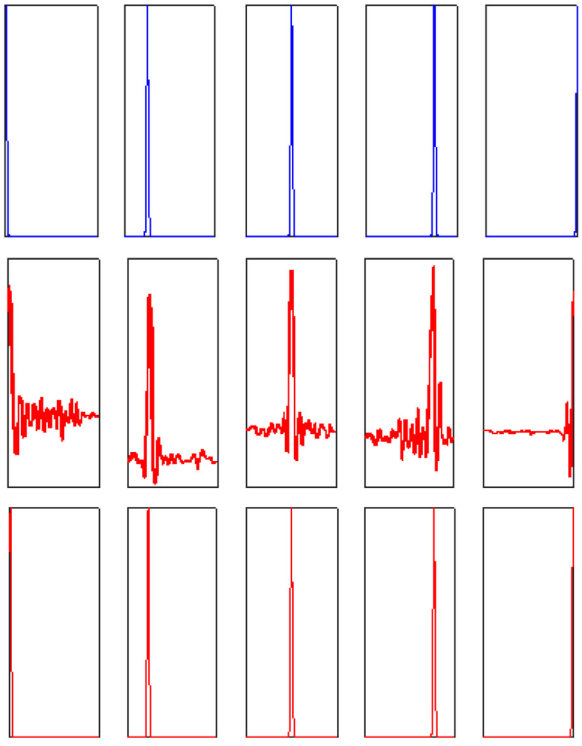

Fig. 8. First row shows several Gaussian functions centered at different data points as $\mathbf{f}^i$'s. Second row shows the corresponding mapped functions $\mathbf{f}^i$'s found by a mapping $\mathbf{T}_i^C$ computed having the coupling points of Fig 7 Third row is the post processed version of the second row (the horizontal axis is the point index which is consistent with their positions on the manifold).

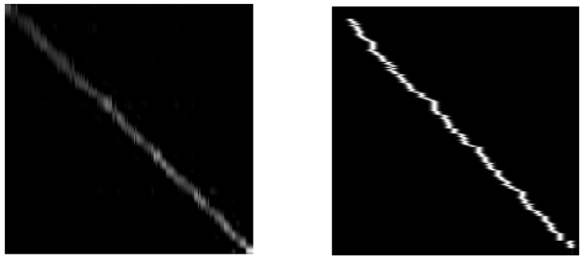

Fig. 9. Matrix **B** shown as image, before (left) and after (right) post processing, having the coupling points of Fig 7.

iments, it is among the 2 best methods. No other method achieves such statistics over the whole set of experiments. The clear advantage of using manifold charting idea for extending the limited amount of supervised information is obvious from the results. The proposed method is also evaluated on the introduced object classification NUS dataset, and compared with the related methods: [48] in two cases (given only positive (coupling) pairs and having both positive and negative (decoupling) pairs), and MultiNMF [61]. The reported results in the first row of Table 3, show the best results of the proposed method with considerable improvements against [61] and moderately better performance than [48]. A reason of this difference may be related to considering multimodal data manifolds in [48] despite [61].

Multimodal remote sensing semantic image segmentation: From the dataset of 2018 IEEE GRSS contest, introduced before, the considered segmentation problem is to classify the pixels of the sensed image shown in Fig 11, in three bi-modal cases: HSI-RGB, HSI-LiDAR, and HSI-HSI. In HSI-HSI experiment, the first 5 frequency bands of the hyperspectral image are considered as the first modality and the last 5 frequency bands are considered as the second modality. The proposed technique is compared with several related multimodal classification methods, and the results are reported in the last three rows of Table 3, in which the best performance of the proposed method, is observed.

Finally, the results of the proposed method are compared with those of LeMA [44] as a recent related state-of-the-art method, as well as the other models with good classification performances tested there, on the dataset of 2013 IEEE GRSS contest, introduced before. All methods are evaluated on multimodal datasets, with hyperspectral (HS) and multispectral (MS) modalities. Two commonly used and high-performance classifiers, namely linear support vector machines (LSVM) and canonical correlation forest (CCF) [64] are used to work on the feature vectors obtained by the proposed and the other methods. The results are reported in Table 4. It is observed that the proposed method is the best in terms of overall accuracy and kappa coefficient, and the second best in terms of average accuracy. Also, the accuracy of each class has been separately reported in this table. It is observed from the results that classes 7 to 13 are the most challenging ones. For most of them the proposed method still has very good performances. It has in 4 classes out of 7 (classes 7, 8, 9, and 11, respectively) the best accuracy and in 1 other (class 12) the second best accuracy. Moreover, the best improvements made by the proposed method are especially observed for the classes with obvious manifold structures, like road and railway (classes 9 and 11) which appear as elongated regions. This confirms that the proposed method is quite successful in its main goal that is geometric modeling of multimodal problems by joint data manifold learning and alignment.

Table 3
Classification accuracy of the proposed method in comparison with other related methods, on three remote sensing and nus multimodal problems. Average and standard deviations in 50 runs are reported.

Modalities	Methods			
	CD pos [48]	CD pos & neg [48]	Proposed method	MultiNMF [61]
NUS	80.69±2.89	81.79±3.17	**83.95±2.42**	54.48±3.23
HSI - RGB	60.68±3.14	64.22±1.11	**79.79±1.13**	49.79±0.02
HSI - Lidar	55.68±1.46	53.17±0.41	**90.68±1.23**	49.80±0.02
HSI – HSI	25.54±1.22	25.85±0.15	**71.51±1.68**	30.00±1.50

Table 4
Classification performance comparison of the proposed method (the last two columns) with those of the related algorithms considered in [44].

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LEMA (%)		Our Soluton (%)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	62.1	68.2	64.7	70	68	69.6	69.3	70.1	69.4	72.2	70.4	73.8	73.4	76.4	72.6	**77.8**
AA	66	70.5	68.2	72.2	70.5	71	72	72.9	71.7	73.6	73.1	75.6	74.8	**77.2**	73.8	76.7
κ	0.59	0.65	0.62	0.67	0.65	0.67	0.67	0.68	0.67	0.7	0.68	0.71	0.71	0.74	0.7	**0.76**
Class1	76.4	68	77.8	78	75.3	68.5	74.3	73.5	75.5	70	**91.9**	88	89.6	85.8	100	80
Class2	80.6	78.1	93.9	**98**	97.6	77.9	97.6	93.7	73.7	78	90.1	91.6	93.7	93.9	70.2	93
Class3	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Class4	85.5	92.3	89.7	96.6	94.8	98.7	95.9	98.6	98.7	98.3	92.8	97.3	97.5	99.6	**100**	**100**
Class5	99.1	99.4	99.5	**99.7**	99	99.1	99.3	99.4	99.4	99.4	99.4	99.7	99.5	99.6	95	**100**
Class6	86.1	86.1	96.4	99	86.5	71	**99.7**	**99.7**	85.5	85.2	**99.7**	96.7	86.5	86.5	91.1	89.4
Class7	50.6	63.8	48.6	64	72.3	77.1	72.2	69.7	74	80.1	75.1	81	83.2	88	80	**89.9**
Class8	56.5	56.1	56.6	59.9	62	62.2	64.6	63.9	63.5	62	55.8	60.4	62.8	62	**72.8**	67.4
Class9	56.2	70.6	69.6	69	50	91.3	50.6	45	59.8	64.9	65.8	71.5	64.5	61.9	59.9	**74.2**
Class10	45.4	45.3	45.5	49.9	58.1	52.3	58.3	63.6	**64.1**	57.7	59	51.8	61	53.6	62.1	50.6
Class11	27.4	43.9	22.5	38.7	28.9	36.5	36.5	34.8	36.5	47.3	35.8	38.7	41.3	50	35	**57.7**
Class12	31.6	56.1	31.8	37.8	35.8	62.5	34.2	55.2	46.8	62.7	34.3	58.5	45	**76.9**	68.9	75
Class13	0	0.7	0	1.1	0	0	0	0.5	0	0.5	0	0.9	0	**1.8**	0.2	0.5
Class14	97.5	98.8	94.5	92.6	**100**	**100**	99.4	98.2	**100**	99.4	99.4	**100**	99.4	**100**	75.3	75.3
Class15	96.6	98.2	96.6	**98.4**	97.4	98.2	97.6	97.6	97.6	98.2	97.9	98.2	97.6	98.2	96.6	97.1

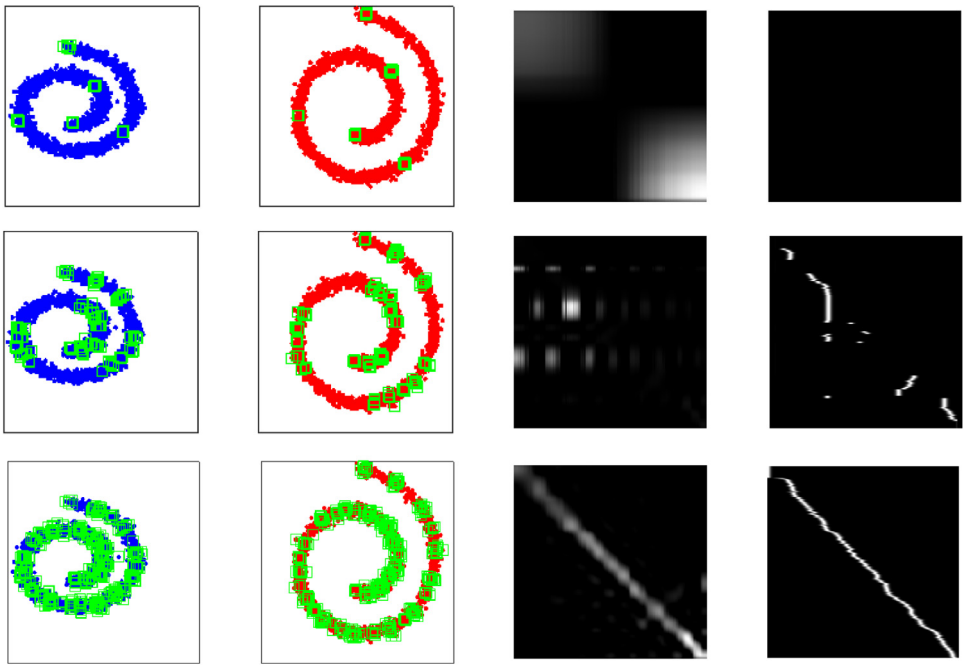

Fig. 10. The points shown by '□' in the first two columns are coupling points. The number of these points are increasing from the first to the last row. The last two columns show matrix $\mathbf{B}_{i,j}$ corresponding to each set of coupling points shown in their rows before and after post processing.

We set the parameters of the proposed method as follows. To construct the functional mapping, we used $d_i^x = d_j^x = 13$. In Eqs. (3) and (4), we set $\alpha = 1$ and $\beta = 1$. In equations (9) we set $c_1 = c_2 = 1$, because after the normalization applied in (10) and (11), the spectral and spatial weights are expected to have the same importance for image segmentation. In Eq. (21), the values of σ^2 for two modalities are computed by self-tuning [48], and in Eq. (11) we set $\tau = 1$. For HS modality we set $\beta_c = 0.6$, $\beta_d = 0.7$, $\delta_d = 0.5$ and $\delta_c = 1$, and for MS modality we set $\beta_c = 0.8$, $\beta_d = 0.9$, $\delta_d = 0.5$ and $\delta_c = 0.7$. In Eq. (22), we set $\mu_c = \mu_d = 1$ and $\sigma_d^2 = \sigma_d^2 = 0.5$. The dimensions of two modalities are reduced to 8.

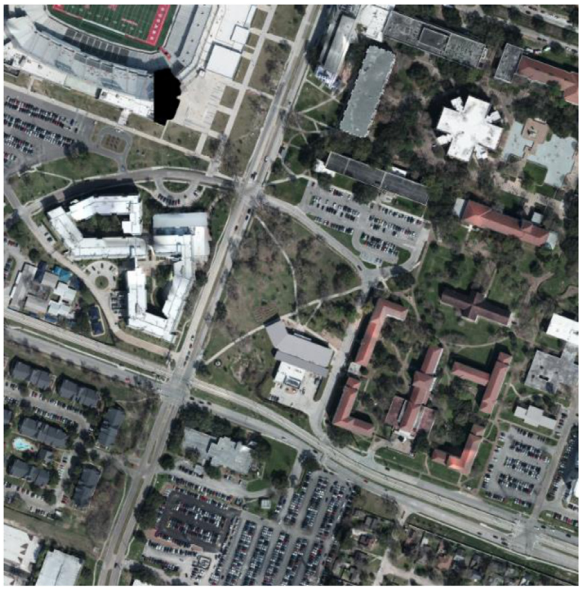

Fig. 11. The RGB modality of the used remote sensing multimodal problem of 2018 IEEE GRSS dataset.

5. Conclusion and future work

In this paper, a new idea for multimodal manifold learning based on semi-supervised charting and joint analysis of manifolds in different modalities of a problem was proposed. The local nature of the low dimensional underlying data manifolds in many real-world datasets motivates to include local properties of the data in the learning and alignment of globally nonlinear data manifolds in different modalities. The proposed charting idea obtains local patches around the given coupling and decoupling regions, and propagates this minimal supervised information to neighbor data points, which is followed by extending these information all over the modalities manifold, by functional mapping idea. The obtained mapping matrices are used to respectively modify and define the weights of within-modality and between-modality neighborhood graphs, to highlight the coupling regions and diminish the decoupling ones, in learning the joint embeddings. Finally, joint diagonalization of the Laplacians of these two graphs, gives the final representation of the learned and aligned manifolds in different modalities.

The experimental results on synthetic spirals dataset show promising behavior of charting and functional mapping steps of the proposed algorithm. We then applied successfully our proposed method on real-world applications containing multimodal remote sensing image semantic segmentation and multimodal object classification.

In the future, we will try to use the proposed method on other applications like multimodal image registration, cross-modal image retrieval, etc. Also, from a theoretical point of view, we will try to develop new ideas for better learning and alignment of manifolds in different modalities. Local representation by tangent spaces and linear manifold learning [65-68] can be used to improve our applied spectral analysis. More complicated representations like multi-manifold and multi-geodesic modeling [69] and hierarchical manifold searching and representation [70,71] can be considered in multimodal manifold learning framework proposed here. Also, applying new optimization techniques on Stiefel or Grassman manifolds [47,67,69,72] for designing better optimizer for the model should be considered. Kernel methods for manifold regularization in multi-view learning [52] is another promising direction to extend the model. Out-of-sample generalization technics for manifold learning [73] will also be applicable to improve the proposed model on unseen data. Finally, the proposed ideas for multimodal manifold learning can be considered to define more suitable representations in multimodal deep learning frameworks [17,22].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Davide Eynard and his colleagues and also Danfeng Hong and his colleagues for sharing the codes and/or datasets of their work [48] and [44] with us. We would also like to thank Christian Jutten for his precise comments on the first version of this paper. Additionally, P. Adibi (the first author) would like to thank Campus France organization and French embassy in Iran for financial support to visit GIPSA-lab for 6 months in 2017–2018, where this research has been started.

References

[1] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477.
[2] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, J.A. Benediktsson, Challenges and opportunities of multimodality and data fusion in remote sensing, Proc. IEEE 103 (9) (2015) 1585–1601.
[3] T. Baltrušaitis, C. Ahuja, L.-.P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–433.
[4] M. Angelou, V. Solachidis, N. Vretos, P. Daras, Graph-based multimodal fusion with metric learning for multimodal classification, Pattern Recognit. 95 (2019) 296–307.
[5] Z. Zhang, M. Zhao, T.W.S. Chow, Constrained large margin local projection algorithms and extensions for multimodal dimensionality reduction, Pattern Recognit. 45 (12) (2012) 4466–4493.
[6] Y. Liu, L. Liu, Y. Guo, M.S. Lew, Learning visual and textual representations for multimodal matching and classification, Pattern Recognit. 84 (2018) 51–67.
[7] D. Rafailidis, S. Manolopoulou, P. Daras, A unified framework for multimodal retrieval, Pattern Recognit. 46 (12) (2013) 3358–3370.
[8] G. Tochon, M. Dalla Mura, M. Angel Veganzones, T. Géraud, J. Chanussot, Braids of partitions for the hierarchical representation and segmentation of multimodal images, Pattern Recognit. 95 (2019) 162–172.
[9] P. Song, X. Deng, J.F.C. Mota, N. Deligiannis, P. Luigi Dragotti, M.R.D. Rodrigues, Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries, IEEE Trans. Comput. Imag. 6 (2020) 57–72.
[10] P. Turaga, R. Anirudh, R. Chellappa, Manifold learning, in: K Ikeuchi (Ed.), Computer Vision, Springer Nature Switzerland AG, 2020.
[11] B. Li, Y.-.R. Li, X.-.L. Zhang, A survey on laplacian eigenmaps based manifold learning methods, Neurocomputing 335 (2019) 336–351.
[12] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang "Hyperspectral image super-resolution with optimized RGB guidance," in Proc. CVPR, 2019.
[13] Y. Jia, Y. Zheng, L. Gu, A. Subpa-Asa, A. Lam, Y. Sato, and I. Sato "From RGB to spectrum for natural scenes via manifold-based mapping," in Proc. ICCV, 2017.
[14] J. Hu, D. Hong, X.X. Zhu, MIMA: mAPPER-induced manifold alignment for semi-supervised fusion of optical image and polarimetric SAR data, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9025–9040.
[15] L. Gómez-Chova, D. Tuia, G. Moser, G. Camps-Valls, Multimodal classification of remote sensing images: a review and future directions, Proc. IEEE 103 (9) (2015).
[16] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, L. Guibas, Functional maps: a flexible representation of maps between shapes, ACM Trans. Graph. 31 (4) (2012).
[17] M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond Euclidean data, IEEE Signal Process. Mag. 34 (4) (2017) 18–42.
[18] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on Riemannian manifolds," in Proc. 3DRR, 2015.
[19] D. Boscaini, J. Masci, E. Rodola, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in Proc. NIPS, 2016.
[20] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in Proc. CVPR, 2017.
[21] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, IEEE Trans. Image Process. 27 (5) (2018).
[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in Proc. ICML, 2011.

[23] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, J. Mach. Learn. Res. 15 (2014) 2949–2980.
[24] J. Gao, P. Li, Z. Chen, J. Zhang, A survey on deep learning for multimodal data fusion, Neural Comput. 32 (5) (2020) 829–864.
[25] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," arXive 2019.
[26] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence., Early access, 2019.
[27] C. Hong, J. Yu, J. Zhang, X. Jin, K.-.H. Lee, Multi-modal face pose estimation with multi-task manifold deep learning, IEEE Trans. Ind. Inf. 15 (7) (2019).
[28] N. Zheng, L. Qi, and L. Guan, "Multiple-manifolds discriminant analysis for facial expression recognition from local patches set," in Proc. MPRSS, 2014.
[29] J. Li, Y. Wu, J. Zhao, K. Lu, Multi-manifold sparse graph embedding for multi–modal image classification, Neurocomputing 173 (P3) (2016) 501–510.
[30] Z. Zhang, T. Chow, M. Zhao, Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization, IEEE Trans. Knowl. Data Eng. 25 (5) (2013) 1148–1161.
[31] M. San-Biagio, M. Crocco, M. Cristani, S. Martelli, and V. Murino, "Low-level multimodal integration on riemannian manifolds for automatic pedestrian detection," in Proc. FUSION, 2012.
[32] C. Hong, J. Yu, D. Tao, M. Wang, Image-based 3D human pose recovery by multi-view locality sensitive sparse retrieval, IEEE Trans. Indus. Electron. 62 (6) (2015) 3742–3751.
[33] J. Zhang, J. Yu, J. You, D. Tao, N. Li, J. Cheng, Data-driven facial animation via semi-supervised local patch alignment, Pattern Recognit. 57 (2016) 1–20.
[34] T. Cao, C. Zach, S. Modla, D. Powell, K. Czymmek, and M. Niethammer, "Multi-modal image registration for correlative microscopy," arXiv: 1411. 3229, 2015.
[35] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[36] C. Wachinger and N. Navab, "Manifold learning for multi-modal image registration," in Proc. BMVC, 2010.
[37] F.S. Bashiri, A. Baghaie, R. Rostami, Z. Yu, R.M. D'Souza, Multi-modal medical image registration with full or partial data: a manifold learning approach, J. Imag. 5 (5) (2019).
[38] A. Kazi, S. Conjeti, A. Katouzian, and N. Navab, "Coupled manifold learning for retrieval across modalities," in Proc. ICCV, 2017.
[39] B. Jie, Daoqiang Zhang, Bo Cheng, D. Shen, Manifold regularized multitask feature learning for multimodality disease classification, Hum. Brain Mapp. 36 (2) (2015) 489–507.
[40] D. Tuia, M. Volpi, M. Trolliet, G. Camps-Valls, Semisupervised manifold alignment of multimodal remote sensing images, IEEE Trans. Geosci. Remote Sens. 52 (12) (2014).
[41] D. Tuia, G. Camps-Valls, Kernel manifold alignment for domain adaptation, PLoS ONE 11 (2) (2016).
[42] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in Proc. CVPR, 2012.
[43] G. Iyer, J. Chanussot, and A. Bertozzi, "A graph-based approach for feature extraction and segmentation of multimodal images," in Proc. ICIP, 2017.
[44] D. Hong, N. Yokoya, N. Ge, J. Chanussot, X.-.X. Zhu, Learnable manifold alignment (LeMA): a semi-supervised cross-modality learning framework for land cover and land use classification, ISPRS J. Photogramm. Remote Sens.. 147 (2019) 193–205.
[45] A. Bunse-Gerstner, R. Byers, V. Mehrmann, Numerical methods for simultaneous diagonalization, SIAM J. Matrix Anal. Appl. 14 (4) (1993) 927–949.
[46] J. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, Radar Signal Process. 140 (6) (1993) 362–370.
[47] J. Cardoso, A. Souloumiac, Jacobi angles for simultaneous diagonalization, SIAM J. Matrix Anal. Appl 17 (1996) 161–164.
[48] D. Eynard, A. Kovnatsky, M.M. Bronstein, K. Glasho, A.M. Bronstein, Multimodal manifold analysis by simultaneous diagonalization of Laplacians, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2505–2517.
[49] K. Nenning, K. Kollndorfer, V. Schopf, D. Prayer, and G. Langs, "Multi-subject manifold alignment of functional network structures via joint diagonalization," in Proc. IPMI, 2015.
[50] M. Pilancı and E. Vural, "Domain adaptation via transferring spectral properties of label functions on graphs," in Proc. IVMSP, 2016.
[51] O. Lindenbaum, A. Yeredor, and M. Salhov, "Learning coupled embedding using MultiView diffusion maps," in Proc. LVA/ICA, 2015.
[52] H. Minh, L. Bazzani, V. Murino, A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning, J. Mach. Learn. Res. 17 (2016) 1–72.
[53] A. Nazarpour, P. Adibi, Two-stage multiple kernel learning for supervised dimensionality reduction, Pattern Recognit 48 (2015) 1854–1862.
[54] U. von Luxburg, A tutorial on spectral clustering, Stat. Comput. 17 (4) (2007) 395–416.
[55] A. Yeredor, Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation, IEEE Trans. Signal Process. 50 (7) (2002) 1545–1553.
[56] Z. Wen, W. Yin, A feasible method for optimization with orthogonality constraints, Math. Prog. 142 (1–2) (2013) 397–434.
[57] IEEE GRSS data fusion contest 2018: http://www.grss-ieee.org/community/technical-committees/data-fusion/2018-ieee-grss-data-fusion-contest, <Date of access: December 2018>
[58] Ticinum aerospace company: http://dase.ticinumaerospace.com, <Date of access: December 2018>.
[59] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: a comparative review," IEEE Geosci. Remote Sens. Mag. 5 (2), 29–56.
[60] R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006) 5–30.
[61] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in Proc. SDM, 2013
[62] N. Courty, R. Flamary, and D. Tuia, "Domain adaptation with regularized optimal transport," In: Proc. ECML, Nancy, France, pp. 274–289, 2014.
[63] M. Long, J. Wang, G. Ding, J. Sun, and P.S. Yu, "Transfer feature learning with joint distribution adaptation," In: Proc. ICCV, pp. 2200–2207, 2013.
[64] T. Rainforth and F. Wood, "Canonical correlation forests," arXiv preprint arXiv: 1507.05444, 2015.
[65] Y. Zhou, S. Sun, Local tangent space discriminant analysis, Neural Process. Lett. 43 (2016) 727–744.
[66] S. Sun, X. Xie, Semisupervised support vector machines with tangent space intrinsic manifold regularization, IEEE Trans Neural Netw Learn Syst 27 (9) (2016).
[67] L. Liu, R. Ge, J. Meng, G. You, Dual subspace learning via geodesic search on Stiefel manifold, Int. J. Mach. Learn. Cybernetic. 5 (2014) 753–759.
[68] P. Adibi and R. Safabakhsh, "Batch linear manifold topographic map with regional dimensionality estimation," in Proc. IJCNN, 2009.
[69] X. Wang, K. Slavakis, and G. Lerman, "Multi-manifold modeling in non-euclidean spaces," in Proc. AISTATS, 2015.
[70] A. Moutzouris, J. Martinez-del-Rincon, J. Nebel, D. Makris, Efficient tracking of human poses using a manifold hierarchy, Comput. Vis. Image Understand. 132 (2015) 75–86.
[71] P. Adibi, A growing hierarchical approach to batch linear manifold topographic map formation, J. Comput. Secur. 1 (1) (2014) 47–59.
[72] A. Kovnatsky, K. Glashoff, and M. Bronstein, "MADMM: a generic algorithm for non-smooth optimization on manifolds," in Proc. ECCV, 2016.
[73] E. Vural, C. Guillemot, Out-of-Sample Generalizations for Supervised Manifold Learning for Classification, IEEE Trans. Image Process. 25 (3) (2016) 1410–1424.

Ali Pournemat received the B.*Sc.* degree from the department of information technology, faculty of computer engineering, University of Isfahan, in 2016 (first rank), and M.*Sc.* degree from the department of artificial intelligence, faculty of computer engineering, University of Isfahan, in 2019. His-current research interests include Machine Learning, Pattern Recognition and Image Processing.

Peyman Adibi received the Ph.D. degree from faculty of computer engineering, Amirkabir University of Technology, Tehran, Iran in 2009. He is currently an assistant professor of the faculty of computer engineering at University of Isfahan, where he is the head of artificial intelligence department. His-current research interests include Machine Learning and Pattern Recognition, Computer Vision and Image Processing, Computational Intelligence and Soft Computing and their applications.

Jocelyn Chanussot received the M.*Sc.* degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998. In 1999, he was with the Geography Imagery Perception Laboratory for the Delegation Generale de l'Armement (DGA—French National Defense Department), Arcueil, France. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He is also conducting his research at the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; KTH Royal Institute of Technology, Stockholm, Sweden; and National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. His-research interests include image analysis, multicomponent image processing, nonlinear filtering, and data fusion in remote sensing. Dr. Chanussot was a member of the IEEE Geoscience and Remote Sensing Society AdCom from 2009 to 2010, in charge of membership development, the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008, and the Institut Universitairede France from 2012 to 2017. He is the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010 which received the 2010 IEEE GRS-S Chapter Excellence Award. He was aco-recipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 and 2015 Symposium Best Paper Award, the IEEE GRSS 2012 Transactions Prize Paper Award, and the IEEE GRSS 2013 Highest Impact Paper Award. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair from 2009 to 2011 and the Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing, (2009). He served as an Associate Editor for IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2005 to2007 and Pattern Recognition from 2006 to 2008. Since 2007, he has an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was a Guest Editor for the Proceedings of the IEEE in 2013 and the IEEE Signal Processing Magazine in 2014. He was the Editor-in-Chief of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing from 2011 to 2015.