List of References

Connor H. McCurley

Contents

1 Manifold/ Representation Learning

1.1 Classic Methods

van der Maaten et al. (2007) - *Dimensionality Reduction: A Comparative Review*
Summary:

Jindal and Kumar (2017) - *A Review on Dimensionality Reduction Techniques*
Summary:

Bengio et al. (2012) - *Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives*
Summary:

Tenenbaum et al. (2000) - *A Global Geometric Framework for Nonlinear Dimensionality Reduction*
Summary:

Roweis and Saul (2000) - *Nonlinear Dimensionality Reduction by Locally Linear Embedding*
Summary:

Saul and Roweis (2001) - *An introduction to locally linear embedding*
Summary:

Belkin and Niyogi (2003) - *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation*
Summary:

Bishop et al. (1998) - *GTM: The Generative Topographic Mapping*
Summary:

Delaporte et al. (2008) - *An introduction to diffusion maps*
Summary:

Theodoridis and Koutroumbas (2008b) - *The Karhunen-Loeve Transform*
Summary:

Theodoridis and Koutroumbas (2008a) - *Kernel PCA*
Summary:

Tipping and Bishop (1999) - *Probabilistic Principal Component Analysis*
Summary:

Lawrence (2003) - *Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data*
Summary:

Lawrence (2005) - *Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models*
Summary:

Gorban and Zinovyev (2008) - *Elastic Maps and Nets for Approximating Principal Manifolds and Their Application to Microarray Data Visualization*
Summary:

Lee et al. (2016) - *Learning Representations from Multiple Manifolds*
Summary:

Kokiopoulou and Saad (2007) - *Orthogonal Neighborhood Preserving Projections: A Projection-Based Dimensionality Reduction Technique*
Summary:

Talmon et al. (2015) - *Manifold Learning for Latent Variable Inference in Dynamical Systems*

Summary:

Nickel and Kiela (2017) - *Poincaré Embeddings for Learning Hierarchical Representations*
Summary:

1.2 Supervised and Semi-Supervised

Belkin and Niyogi (2004) - *Semi-Supervised Learning on Riemannian Manifolds*
*Summary:*The authors developed a semi-supervised approach for classification on Riemannian manifolds. The algorithms uses unsupervised data to model the manifold using the adjacency graph and graph Laplacian, then uses labels to learn a discriminant function in the embedding space. **Collecting labeled examples can be costly and time-consuming.** This paper also provides a detailed explanation of the **Graph Laplacian**.

X. Geng et al. (2005) - *Supervised nonlinear dimensionality reduction for visualization and classification*
*Summary:*The author's developed a supervised Isomap, called S-Isomap. Experiments compare S-Isomap to Isomap, LLE, and Weighted Isomap. *Dimensionality reduction can be performed by keeping only the most important dimensions, i.e., the ones that hold the most useful information for the task at hand, or by projecting the original data into a lower dimensionali space that is most expressive for the task.* Most real-world datasets are nonlinear. Both Isomap and LLE attmept to preseerve as well as possible the local neighborhood of each object while trying to obtain highly nonlinear embeddings, and are thus categorized as local embedding methods. The central idea of local embeddings is using the locally linear fitting to solve the globally nonlinear problems, which is based on the assumption that data lying on a nonlinear manifold can be viewed as linear in local areas. The goal of this application is a preprocessing proceedure to increase discriminability and be robust to noise. **In visualization, the goal is to faithfully preserve the intrinsic structure as well as possible, while in classification, the goal is to transform the original data into a feature space that can make classification easier, by stretching or constricting the original metric if necessary.** Weighted IsoMap changes the Euclidean distance matrix between two data points by a constant factor if their class lables are the same. This makes similar points closer in the feature space if they belong to the same class. However, this is not suitable for visualization because if distorts the original structure of the input data. S-Isomap also applies contraints on the first step of Isomap by changing the dissimilarity matrix to make points with the same label more similar and points with opposing labels more dissimilar. When data from the same class are scattered (multimodal) the neighborhood graph may be disconnected and neighter Isomap or S-Isomap can handle this.

Sugiyama (2006) - *Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction*
*Summary:*Dimensionality reduction is one of the important preprocessing steps in high-dimensional data analysis. The authors considered the supervised dimensionality reduction problem, where each sample has an associated label. The author notes that Fisher's discriminant often fails when classes are multi-modal. This lead the author to develop a local FDA which allows multimodal data to be embedded appropriately. This method combines traditional FDA and locality preserving projections (LLP) where between-class separability is maximized while within-class local structure is preserved. This method is linear, but can be kernelized.

Rish et al. (2008) - *Closed-form supervised dimensionality reduction with generalized linear models*

*Summary:*The authors propose a family of supervised dimensionality reduction (SDR) algorithms that combine feature extraction with learning a predictive model in a unified optimization framework, using data- and class-appropriate generalized linear models (GLMs), and handling both classification and regression problems. **Dimensionality reduction helps prevent overfitting when the number of dimensions greatly exceeds the number of samples, thus working as a form of regularization. When the goal is prediction, SDR that combines DR with simultaneously learning a predictor can significantly outperform a simple combination of unsupervised DR with a subsequent learning of a predictor on the resulting low-dimensional representation.** The problem of supervised dimensionality reduction can be viewed as finding a predictive structure, such as a low-dimensional representation, which captures the information about the class label contained in the high-dimensional feature vector while ignoring the "noise". SDR approaches are often restricted to a specific setting, and can be viewed as jointly learning a *particular mapping* from the feature space to the low-dimensional hidden-variable space, together with a *particular predictor* that maps the hidden variables to the class label. The proposed method views both features and labels as exponential-family random variables. It can be viewed as a discriminative learning based on minimization of conditional probability of class given the hidden variables, while using as a regularizer the conditional probability of the features given the low-dimensional hidden variable "predictive" representation. They also use a closed form expectation-maximization-like update by using auxiliary functions.

Z. Zhang et al. (2008) - *Spectral methods for semi-supervised manifold learning*
Summary:

Li et al. (2009) - *A supervised manifold learning method*
Summary:

Gao et al. (2011) - *Supervised Gaussian Process Latent Variable Model for Dimensionality Reduction*
Summary:

Raducanu and Dornaika (2012) - *A supervised non-linear dimensionality reduction approach for manifold learning*
Summary:

Gnen (2013) - *Bayesian Supervised Dimensionality Reduction*
Summary:

Liu et al. (2014) - *Hybrid Manifold Embedding*
Summary:

Bouzas et al. (2015) - *Graph Embedded Nonparametric Mutual Information for Supervised Dimensionality Reduction*
Summary:

Vural and Guillemot (2016) - *Out-of-Sample Generalizations for Supervised Manifold Learning for Classification*
Summary:

Vepakomma et al. (2016) - *Supervised Dimensionality Reduction via Distance Correlation Maximization*
Summary:

Xu et al. (2017) - *Active manifold learning via a unified framework for manifold landmarking*
Summary:

Vural and Guillemot (2018) - *A study of the classification of low-dimensional data with supervised manifold learning*

Summary: **Supervised manifold learning methods learn data representations by preserving the geometric structure of data while enhancing the separation between data samples from different classes.** This paper provides a good lit review of supervised dimensionality reduction methods and explains the differences between linear and nonlinear approaches. The authors studied satisfying conditions on the interpolation functions for nonlinear methods for test. It was found that the misclassification error probability decays at an exponential rate with the number of samples, provided that the interpolation function is sufficiently regular with respect to the separation margin of the embedding.
Supervised Laplacian eigenmaps methods embed the data with the eigenvectors of the linear combination of two graph Laplacian matrices that encode the links between neighboring samples from the same class and different classes. In such a data representation, the coordinates of neighboring data samples change slowly within the same class and rapidly across different classes.
This paper provides good mathematical definitions of data separability. There is also toy data in their experiments section which might be usable for MIL discriminative manifold learning. **Compare against their toy data experiments with MIL.**

Kang et al. (2018) - *Robust Graph Learning from Noisy Data*
Summary:

Chen et al. (2018) - *Robust Semi-Supervised Manifold Learning Algorithm for Classification*
Summary:

Chao et al. (2019) - *Recent Advances in Supervised Dimension Reduction: A Survey*
Summary:

1.3 Manifold Alignment/Matching

Bengoetxea (2002) - *Inexact Graph Matching Using Estimation of Distribution Algorithms*
Summary:

This thesis chapter provides a good general overview of graphs in general as well as the definitions of graph matching (graph isomorphism) and SOA approaches (in 2002). The amount of this type of work shows that the interest on the field is increasing with the years.

Basic notation and terminology: A graph $G = (V, E)$ in its basic form is composed of vertices and edges. V is the set of vertices (also called nodes or points) and $E \subset V \times V$ (also defined as $E \subset [V]^2$ in the literature) is the set of edges (also known as arcs or lines) of graph G. The *order* (or *size*) of a graph G is defined as the number of vertices of G and it is represented as $|V|$ and the number of edges as $|E|$. If two vertices in G, say $u, v \in V$, are connected by and edge $e \in E$, this is denoted by $e = (u, v)$ and the two vertices are said to be *adjacent* or *neighbors*. Edges are said to be undirected when they have no direction, a nd a graph G containing only such types of connections is called *undirected*. When all edges have directions and therefore (u, v) and (v, u) can be distinguished, the graph is said to be *directed*. Usually, the term *arc* is used when the graph is directed, and the term *edge* is used when it is undirected. In addition, a directed graph $G = (V, E)$ is called *complete* when there is always and edge $(u, u') \in E = V \times V$ between any two vertices u, u' in the graph. Graph vertices and edges can also contain information. When this information is a simple label, the graph is called a *labeled graph*. Other times vertices and edges contain more information. These are called vertex and edge *attributes*, and the graph is called an *attributed graph*. More commonly, this concept is further specified by distinguishing between *vertex-attributed* (or *weighted graphs*) and *edge-attributed* graphs. A *path* between any two vertices $u, u' \in V$ is a non-empty sequence of k different vertices $< v_0, v_1, \ldots, v_k >$ where $u = v_0, u' = v_k$ and $(v_{i-1}, v_i) \in E$, $i = 1, 2, \ldots, k$. Finally, a graph G is said to be *acyclic* when there are no cycles between its edges, independently, of whether the graph G is directed or not.

Definition and classification of graph matching problems: Graphs have been proven as an effective way of representing objects. When using graphs to represent objects or images, vertices usually represent regions (or features) of the object or images, and edges between them represent the relations between regions. Similar graphs can be used for representing objects or general knowledge, and they can be either directed or undirected. When edges are undirected, they simply indicate the existence of a relation between two vertices. On the other hand, directed edges are used when relations between vertices are considered in an asymmetric way. Generally speaking, the graph matching problem can be stated as follows: Given two graphs $G_M = (V_M, E_M)$ and $G_D = (V_D, E_D)$, with $|V_M| = |V_D|$, the problem is to find a one-to-one mapping $f : V_D \to V_M$ such that $(u, v) \in E_D$ iff $(f(u), f(v)) \in E_M$. When such a mapping f exists, this is called an *isomorphism*, and G_D is said to be isomorphic to G_M. This type of problem is said to be *exact graph matching*. The term *inexact* applied to some graph matching problems means that it is not possible to find an isomorphism between the two graphs to be matched. This is the case when the number of vertices is different in both the model and data graphs. When dealing with an *inexact graph matching* problem, the goal is not to find the exact match, but to find the *best* match between the two graphs, and we aim to find a non-bijective correspondence between the data and model graphs. This often corresponds to a search for small graphs within big ones. An important sub-type of inexact graph matching problems are the sub-graph matching (also called *subgraph isomorphism*). In some inexact graph matching problems, the aim is still to find a one-to-one correspondence, but with the exception that some of the vertices in the data graph have no correspondence at all. Graph matching in general is sometimes referred to in the literature as *isomorphic* and *homomorphic*. **It is considered to be one of the most complex problems in object recognition and computer vision.** Its complexity is due to its combinatorial nature. The best correspondence of a graph matching problem is defined as the optimum of some objective function which measures the similarity between matched vertices and edges. This objective function is also called a *fitness function*.

Graph matching has been entertained in a variety of domains, including: subgraph transformations, fuzzy set theory, elastic and morphological graph matching, multiple graph matching, Bayesian methods, error-correction, genetic algorithms (GA), other forms of probability theory, probabilistic relaxation, Expectation-Maximization (EM), decision trees, clustering techniques, and neural networks. A popular dissimilarity measure on graphs is called the *graph edit distance* which is defined as the number of modification that one graph has to undertake to become the other. While many approaches have been taken toward graph matching, the complexity of the problem itself makes it very difficult to consider all types of dependencies between the vertices and edges in the model and data graphs.

Livi and Rizzi (2013) - *The Graph Matching Problem*

*Summary:***Refer to this for a good reference on all things graphs.** (I didn't want to type it all out.)

Navaratnam et al. (2007) - *The Joint Manifold Model for Semi-supervised Multi-valued Regression*

Summary: The authors present a method for *Joint Manifold Modeling (JMM)* which makes use of a *Gaussian Processes Latent Variable Model (GPLVM)*. Gaussian Processes (GP) are distributions over functions which, when conditioned on training data, produce radial basis function approximations of the data, along with associated covariance estimates. The GP is one way to model regression. Now, provided training examples $\boldsymbol{X} \in \mathbb{R}^d$ which are assumed to live in a low-dimensional submanifold of $\mathbb{R}^d$. The goal of the GPLVM is to discover this submanifold, that is to associate with each $\boldsymbol{x}_i$ a latent variable $\boldsymbol{t}_i$ representing its coordinates in the manifold $\mathcal{M}^k$ where $k < d$. An advantage of the GPLVM is that it is relatively straightforward to incorporate unlabelled samples.

This method was run on simulated data for the task of pose estimation. Once the GPLVM is learned, inference is straightforward.

Wang and Mahadevan (2010) - *Multiscale Manifold Alignment*

Summary:

One limitation to existing domain adaptation problems is the assumption that source and target domains are defined by the same features and the differences come from their distributions. To address this problem, manifold alignment was proposed. Manifold alignment builds mappings between disparate data sets by aligning their underlying manifolds, thus providing a geometric framework for knowledge transfer across data sets.

Existing manifold alignment approaches can be classified into two types. In two-step alignment, the first step maps the data to low-dimensional spaces, then a subsequent step eliminates some components from one set to the other so alignment can be achieved. One-step alignment combines the embedding projection and alignment steps into one. Two approaches for one-step alignment are *semi-supervised alignment* and *manifold projections*. Semi-supervised alignment learns *instance-level alignment* by constructing non-linear embeddings, while manifold projections learn *feature-level alignment* by constructing linear embedding functions, thus allowing for out-of-sample embeddings.

However, many real-world data sets exhibit non-trivial regularities at multiple levels, which correspond to their underlying intrinsic structure. With this in mind, the authors investigated the previously un-studied notion of multiscale manifold alignment by using multi-resolution wavelet analysis (which discovers multiscale intrinsic structure in a matrix). Multiscale manifold alignment automatically generates alignment results at different scales by discovering the shared intrinsic multilevel structures exhibited by the datasets. This provides a solution to the key problem in manifold alignment where users must define the latent dimensionality.

Wang et al. (2011) - *Manifold Alignment*

Summary:

***Manifold alignment* is a simultaneous solution to the problem of alignment and a framework for discovering a unifying representation of multiple datasets. The goal is to ultimately map disparate datasets to a joint latent space, while preserving the qualities of each dataset and also highlighting the similarities between the datasets. The fundamental idea of manifold alignment is that all datasets included in the fusion lie on the same manifold.**

The problem of alignment is to identify a transformation of one dataset that "matches it up" with a transformation of another dataset. That is, given two datasets X and Y, whose instances lie on the same manifold, Z, but who may be represented by different features, the problem of alignment is to find two functions f and g, such that $f(x_i)$ is close to $g(y_j)$ in terms of Euclidean distance if x_i and y_j are close with respect to geodesic distance along Z. Here, X is an $n \times p$ matrix containing n data instances in p-dimensional space, Y is an n$m \times q$ matrix containing m data instances in q-dimensional space, $f : \mathbb{R}^p \to \mathbb{R}^k$, and $g : \mathbb{R}^q \to \mathbb{R}^k$ for some k called the latent dimensionality. The instances x_i and y_j are in exact correspondence if and only if $f(x_i) = g(y_j)$. On the other hand, prior correspondence information includes any information about the similarity of the instances in X and Y, not just exact correspondence information. The union of the range of f and g is the joint latent space, and the concatenation of the new coordinates

$$\begin{pmatrix} f(X) \\ g(Y) \end{pmatrix}$$

is the unified representation of X and Y, where $f(X)$ is an $n \times k$ matrix containing the result f applied to each row of X, and $g(Y)$ and $m \times k$ matrix containing the result of g applied to each row of Y. $f(X)$ and $g(Y)$ are the new coordinates of X and Y in the joint latent space.

This paper captures this idea mathematically by concatenating the graph Laplacians of each dataset, thus forming a joint Laplacian. A within-dataset similarity function gives the edge weights of this joint Laplacian between the instances within each dataset, and correspondence information fills in the edge weights between the instances in separate datasets. The manifold alignment algorithm then embeds this joint Laplacian in a new latent space. **The primary challenges of manifold alignment are identifying whether**

the datasets are actually sampled from a single underlying manifold, defining the similarity function that captures the appropriate structures of the datasets, inferring any reliable correspondence information, and finding the true dimensionality of this underlying manifold. Experiments were conducted on protein alignment, parallel corpora and aligning topic models.

Manifold alignment is essentially ta graph-based algorithm, but there is also a vast literature on graph-based methods for alignment that are unrelated to manifold learning, typically called *graph matching* or *graph isomorphism*. Researchers study the general problem of alignment under the names of *information fusion* or *data fusion*.

Davenport et al. (2010) - *Joint Manifolds for Data Fusion*

Summary: In sensor networks, multiple observations of the same event acquired simultaneously from interdependent signals often share a common parameterization. As examples, a camera network might observe a single event from a variety of angles where the underlying event is described by a set of common global parameters (i.e. location, orientation of the object of interest). Similarly, when sensing a single event using multiple modalities, the underlying phenomenon may again be described by a single parameterization (which spans all modalities). The authors contend that a joint manifold can be obtained which encompasses all the component manifolds sharing the same parameterization.

The authors compare properties such as geodesic distances, curvature, branch separation, and condition number of the joint manifold in relation to their component manifolds. They observe that the joint manifold leads to improved performance and noise tolerance for a variety of signal processing algorithms (including ATR). The authors also show how the joint manifold structure can be exploited via a data fusion algorithm based on *random projections*.

There is an argument that *linear* dimensionality reduction should be performed to allow for out-of-sample embedding of the individual modalities, thus allowing for fusion based on compressive sensing and random projections.

The problem of manifold-based classification is, given a set of manifolds, to determine which generated a particular sample. This problem has been explored for application to ATR. The authors describe three distances to be used for generalized maximum likelihood. These distances, which define separation in metric spaces are: *minimum separation*, *maximum separation*, and *Hausdorff distance*.

After reading this paper, it is not quite clear how this method would be implemented.

Wang and Mahadevan (2011) - *Heterogeneous Domain Adaptation Using Manifold Alignment*

Summary:

This paper proposed a manifold alignment based approach for heterogeneous domain adaptation. A key aspect of this approach is to construct mappings to link different feature spaces in order to transfer knowledge across domains. The new approach can reuse labeled data from multiple source domains in a target domain even in the case that the input domains do not share any common features or instances. This paper extended existing manifold alignment approaches by making use of labels, rather than correspondences, to align the manifolds. This was a significant extension to the scope of manifold alignment, as correspondence relationships required by previous methods is often difficult to obtain.

One limitation of domain adaptation is that most existing approaches assume that the sources and target domains are defined by the same features, and the differences between them primarily stem from their distributions. However, this assumption is not valid in most cross-domain applications such as multimodal

datasets.

A key difficulty in applying manifold alignment to domain adaptation is that alignment methods require specifying a small amount of cross-domain correspondence relationships in order to learn mapping functions. However, this correspondence information is often difficult to obtain.

The contributions of this paper are two-fold. From the perspective of domain adaptation, the new approach addresses the problem of transfer even when the source and target domains do not share any common features or instances. From the perspective of manifold alignment, the new approach uses labels rather than correspondences to learn alignment.

Here the authors treat each input dataset as a manifold. The goal is to construct K mapping functions to project the input domains to a new latent space while preserving the topology of each domain, matching instances with the same labels and separating instances with different labels. Each domain is represented by its graph Laplacian (which is built by instance similarity, label similarity, and label dissimilarity). A cost function is constructed such that trade-offs between the three objectives can be met. This function is optimized to learn the set of mapping functions. By representing each domain by it's graph Laplacian, the issue of varying feature sizes is nullified.

Choo et al. (2012) - *Heterogeneous Data Fusion via Space Alignment Using Nonmetric Multidimensional Scaling*

Summary:

Heterogeneous data sets are typically represented in feature spaces, making it difficult to analyze relationships spanning different data sets even when they are semantically related. Data fusion via space alignment can remedy this task by integrating multiple data sets lying in different spaces into one common space. Given a set of reference correspondence data that share the same semantic meaning across different spaces, space alignment attempts to place the corresponding reference data as close together as possible, and accordingly, the entire data are aligned in a common space. Space alignment involves optimizing two potentially conflicting criteria: minimum deformation of the original relationships and maximum alignment between the different spaces. To solve this problem, we provide a novel graph embedding framework for space alignment, which converts each data set into a graph and assigns zero distance between reference correspondence pairs resulting in a single graph. We propose a graph embedding method for fusion based on nonmetric multidimensional scaling (MDS).

The goal of data fusion via space alignment is, given data sets and their reference correspondence pairs (some pairs of data items that share the same semantic meaning between different feature spaces), to provide the new representation of the data in a common space, which enables distance-based comparison across different data sets. The idea is to attain the best alignment between different spaces while allowing a minimum deformation within each space. The authors introduce a novel graph embedding based off of nonmetric MDS, which unlike metric MDS which aims to preserve given pairwise similarities, preserves only the rank ordering of the dissimilarities.

Graph embedding takes as input a graph in which the vertices are data items and the edge weights are their pairwise dissimilarities. As output, graph embedding produces the coordinates of the data in a given dimensional space, which best preserves the relationships described in the graph. It is typically performed in four steps: 1.) forming distance graphs from feature vectors, normalizing the graphs to account for distance scales individual to each dataset, assigning edges between the graphs, and running a graph embedding algorithm to produce the optimal coordinates in the new feature space.

Dimensionality reduction methods such as MDS, kernel PCA, manifold learning methods such as Laplacian Eigenmaps, Isomap, maximum variance unfolding and Sammon's mapping have all been applied as graph embedding procedures for space alignment. However, not all methods are directly applicable since they usually assume a complete graph as input, which is not always available. Additionally, there is no

guarantee that the datasets have a similar intrinsic manifold structure. Therefore, the primary purpose of this paper is neither dimensionality reduction nor intrinsic manifold discover, but finding a common space that reveals the relationships between different data sets by filling out the "unknown" parts.

Another approach to space alignment is to seek linear transformations that map each space into a single space. The advantage of linear methods is that out-of-sample extension is straightforward. In other words, unseen data that are not involved in the alignment process can be directly mapped into the aligned space. Additionally, linear methods generally give an idea about the feature-level relationships between different spaces from an analysis of the linear transformation coefficients.

A few approaches to space alignment (which were compared in this paper) are Constrained Laplacian Eigenmaps (which preserves locality), Orthogonal Procrustes Analysis (which gives an optimal linear transformation represented as an orthogonal matrix which best maps one data set to another that has been used widely in image registration) and PARAFAC2 which is similar to SVD for tensors. One potential problem of Procustes Analysis is that eh dimensionality of the datasets must be equal. **It should also be noted that the perfect alignment between the reference correspondence pairs does not always lead to good alignment for the rest of the data, which is analogous to overfitting in the context of classification.**

Two evaluation metrics were proposed in conjunction with the standard method of mean average precision (mAP). One metric was proposed to measure the amount of deformation by determining the preservation of local neighborhoods through the transformation. The second metric assesses alignment quality by determining how close correspondence pairs are matched in the new space.

Hertzberg et al. (2013) - *Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds*

*Summary:*Sensor fusion is the process of combining information obtained from a variety of different sensors into a joint belief over the system state. In the design of a sensor fusion system, a keyengineering task lies in finding a state representation that 1.) adequately describes the relevant aspects of reality and is 2.) compatible with the sensor fusion algorithm in the sense that the latter yields meaningful or even optimal results when operating over the state representation.
Satisfying both of these goals at the same time has been a long-standing challenge. Standard sensor fusion algorithms typically operate on real valued vector state representations while mathematically sound representations often form more complex, non-Euclidean topological spaces. To address these issues, the authors' approach is based on the observation that sensor fusion algorithms employ operations whcih are inherently local i.e., they compare and modify state variables in a local neighborhood around some reference. Thus, they arrive at a generic solution that bridges the gap between the two goals by **viewing that state space as a manifold.**

Cui et al. (2014) - *Generalized Unsupervised Manifold Alignment*

Summary:

The authors proposed a *Generalized Unsupervised Manifold Alignment* (GUMA) method to build connections between differnt but correlated datasets without any known correspondences.

Manifold alignment tries to build or strengthen the relationships of different datasets and ultimately project samples into a mutual embedding space, where the embedded features can be compared directly. Since samples from different (even heterogeneous) datasets are usually located in different high dimensional spaces, direct alignment in the original spaces is very difficult. On contrast, it is easier to align manifolds of lower intrinsic dimensions.

Supervised and semi-supervised methods require some known between-set counterparts as prerequisite for

the transformation learning (aka labels or handcrafted correspondences). In contrast, *unsupervised manifold alignment* learns from manifold structures and avoids the need for hand-labeled correspondence.

In order to perform this matching, three assumptions are made: 1.) manifolds under the same theme (same action sequences of different persons usually imply a certain similarity in geometric structure), 2.) the embeddings of corresponding points from different manifolds should be as close as possible, and 3.) the geometric structures of both manifolds should be preserved in the mutual embedding space. To obtain these goals, a three part objective function is proposed. Alternating optimization is performed between the correspondence matrix $\boldsymbol{F}$ and embedded points $\boldsymbol{P}_x$ and $\boldsymbol{P}_z$ to learn the correspondence points for matching and to corresponding embedding functions.

A few questions I have... Do the same actions imply similar manifold structure even for different modalities? Also, how difficult would it be to extend this problem to multiple datasets? What does this mean in terms of computational complexity? **Additionally, the authors make a note that there might be partial alignment cases, in which some points on one manifold might not correspond to any points on the other manifold (i.e. different resolutions) and that these points should be detected and not considered in the computation of the matching. However, are these points still considered in the final embedding? i.e. are we losing information?**

Zhang et al. (2014) - *Semi-Supervised Manifold Learning Based Multigraph Fusion for High-Resolution Remote Sensing Image Classification*

Summary:

A semi-supervised manifold learning approach based on Multigraph Fusion (SSM-MF) is proposed. Results demonstrate the benefit of the method toward classification of land-use remote sensing imagery. **Semi-supervised manifold alignment has the advantage that discriminative information and geometric information in labeled data and structural information from unlabeled data can be jointly utilized to enhance the manifold learning. Feature concatenation is a common solution for combining multiple feature vectors. However, concatenation has four apparent drawbacks: 1.) the concatenated vector is physically meaningless since each feature vector has a specific statistical property, 2.) it ignores the diversity of multiple patterns and thus cannot efficiently explore the complementary nature of different patterns, 3.) it may result in Curse of Dimensionality issues, and 4.) one to one correspondence is required between feature vectors.** Their method extends the *Patch Alignment Framework (PAF)* proposed by Zhang et al. which consists of two parts: objective function part optimization and whole alignment. Six methods were compared with the proposed SSM-MF: Feature Concatenation (FC), Best Single Feature (BSF), Best Single Feature Embedding (BSFE), MFC, and two kernel fusion methods, Average Kernel Fusion (AKF) and Product Kernel Fusion (PKF). All classification predictions were tested using an SVM. Naturally, their method outperformed the alternatives in terms of classification accuracy, with AKF as a close second.

Tuia and Camps-Valls (2015) - *Kernel Manifold Alignment*

Summary:

The authors present a method for *manifold alignment* (also denoted as *feature representation transfer* or *feature transformation learning*) which addresses a few key deficiencies in manifold alignment approaches. Specifically, their *kernel manifold alignment* (KEMA) (which can perform alignment or domain adaptation without corresponding pairs, only a few labeled examples in each of the domains) 1.) generalizes other manifold alignment methods, 2.) can align manifolds of different complexities by performing manifold unfolding along with matching, 3.) **can define a domain-specific metric to cope with multimodal specificities**, 4.) **can align data spaces of different dimensionalities**, 5.) is robust to strong nonlinear feature deformations and 6.) is closed-form invertible, which allows transfer across domains and data synthesis.

Roughly speaking, manifold alignment reduces to finding projections to a common latent space where all datasets show similar statistical characteristics. There are three types of manifold alignment/ domain adaptation. *Unsupervised adaptation, semi-supervised adaptation, and supervised adaptation.* Manifold alignment aims at concurrently matching the corresponding instances while preserving the topology of each input data

domain, generally using a *Graph Laplacian*. While appealing, **these methods often require specifying a small amount of cross-domain sample correspondences.**

The method proposed in this paper is a kernel extension to Semi-supervised Manifold Alignment proposed by Wang et. al. A key benefit to that method is that it can easily project data between domains by first mapping the data to the latent domain, then from there inverting back to the target domain. Therefore, the method can be used for both domain adaptation and data synthesis. Additionally, that method does not require data correspondence, only a few labeled samples from each domain. Experiments show that the kernelized extension provided good performance, especially on data with strong nonlinear deformations.

Liao et al. (2016) - *A Manifold Alignment Approach for Hyperspectral Image Visualization With Natural Color*

Summary:

In this paper, the authors apply manifold alignment to the transfer of HSI imagery to RGB for visualization. The manifold embedding made bridged the gap between the high-dimensional spectral space of the HSI and the RGB space of the color imagery. Once the embedding functions were learned, they were easily transferable to alternative images not included in the training set.

Manifold alignment considers the mutual relationships of several data sets at the same time to generate a shared embedding space representing the common low-dimensional manifold shared among the datasets. Of course, there is an assumption that seemingly disparate data sets are produced by similar generating processes and will share a similar underlying manifold structure. Generally, there are two levels of manifold alignment. *Instance-level alignment* builds connections between instances from different datasets, but the alignment result is limited only to known instances and is difficult to generalize to new instances. *Feature-level alignment*, on the other hand, transforms features of different data sets to a common embedding space, which makes direct knowledge transfer possible. The alignment result provides direct connections between features in different spaces, so it easily generalizes to new instances (although you have to assume your features in each space are accurate, discriminative representatives).

In this paper, feature-level manifold alignment was used to find direct mappings between HSI and RGB space. First, a few matching pairs were denoted between the datasets (by hand). Next, the manifolds were aligned in a shared embedding space (using *Locality Preserving Projections*, a linear approximation of the nonlinear *Laplacian eigenmaps*). Next, the inverse mapping could be used to transform the HSI imagery into the RGB colorspace.

Experiments were conducted on HSI data over the Washington D.C. Mall and California Bay. Results demonstrated good performance over four quantitative evaluation criteria.

Manifold alignment based visualization only requires that a matching pair represents the same or similar class of objects/materials rather than being from the same geospatial location. This is different from the traditional matching pair searching in image fusion/ image registration. THIS SUGGESTS THAT EXACT IMAGE REGISTRATION IS NOT A NECESSARY CONDITION TO EXTRACT THE MATCHING PAIRS. *Connor note:* This signifies that weak labels could be used to discover the correspondence pairs.

Yang and Crawford (2016) - *Spectral and Spatial Proximity-Based Manifold Alignment for Multitemporal Hyperspectral Image Classification*

Summary:

The authors explore two methods of manifold alignment on HSI imagery. The methods are compared to the SOA method of LapSVM which is a semi-supervised, manifold-regularized support vector machine.

To fully exploit multi-temporal hyperspectral data, two issues must be addressed, 1.) high dimensionality and 2.) nonstationarity within temporal sequences. To deal with the high dimensionality, **a dimensionality**

reduction process is required prior to analysis, with the goal of extracting the most relevant information to characterize data in a lower dimensional space (if the goal is classification, this suggests incorporating label information!!). For classification, one of the benefitsof exploring the low-dimensional space is that fewer training samples are required to obtain a reliable classifier.

The impact of nonstationarity phenomena is particularly significant due to the narrow spectral bands. Spectral signatures are subject to change over time due to natural (e.g. seasonal phenology of vegetation or environment conditions) and disruptive (e.g. fire, anthropogenic) impacts. Temporal nonstationarity can result in differences in the optimally reduces dimension space. For this reason, additional ground reference information data are needed for classification.

Recently, nonlinear manifold learning was proposed as a successful dimensionality-reduction method to explore intrinsic information concealed in high-dimensional data and to capture nonlinearity Keith hyperspectral data. **As data-driven methodologies, manifold learning methods often yield distance manifolds whose correspondence is difficult to model, particularly for environments that vary over space and time.** One may attempt to develop a joint manifold representation by exploiting spectrally or spatially "nearby" samples across two data sets. However, spectral drift resulting from nonstationarity within a temporal sequence can lead to changes within a cluster of spectral data, and may not provide a reliable means for deriving joint data manifolds. Assuming that the class types are the same and the intrinsic local structures of the classes contained in multiple images are similar, a proper cluster describing the same class may be obtained by aligning the local geometric structures of the spectral data. Global spectral discrepancies associated with the class-dependent signature drift across a series of images may be reduced by grouping spectral neighbors within each image and mapping these clusters to an adequate common latent space supported by these local structures.

In this paper, the authors propose two manifold alignment (MA) techniques that involve aligning underlying local manifolds of temporally sequential data sets. The first approach exploits a local-based manifold of a source image, considered to be the optial "prior" manifold that cannot be intuitively reused by another image. The second method links local manifolds of two images using bridging pairs (correspondence pairs) by considering spectral and spatial relationships between two temporal images. The proposed methods extend graph-based semi-supervised learning and explore MA for the multitemporal hyperspectral image classification task, while providing a domain adaptation framework for HSI analysis from the geometric learning point of view. This paper explores the data geometry problem where only one labeled image is assumed to be available, and the goal is to classify another image.

A data manifold is expected to be geometrically supported by data points. The grate the number of data points, the better the data structure is represented, as long as the points are not redundant and are well distributed across the data manifold. Popular manifold learning methods can generally be formulated in a graph embedding framework. When exploring data manifolds that are developed separately, **it may be reasonable to assume that the intrinsic structures of classes contained in the sequential images are similar.** Manifold alignment has been investigated on a variety of tasks, however, **TO DATE, FEW STUDIES HAVE INVESTIGATED MA IN THE REMOTE SENSING COMMUNITY.**

The authors propose two methods for MA, one of which considers both spectral and spatial similarity by incorporating a spatial proximity term in the construction of the weight adjacency matrix. Their methods are compared to the SOA LapSVM as well as several variations of KNN. Their methods obviously outperforms the alternatives.

A possible fruitful direction for extending these results would be to consider the use of submanifolds for classification tasks (this is great for my proposal!!!). At a local scale, the regions with different desities can be viewed as submanifolds (within-class variatino in each cluster, similar to hierarchical methods. This could be a good place for the hierarchical growing neural gas) which are combined at the global level. Such manifolds could be learned at the local scale and prove beneficial to our local MA concepts. Local and global-based approaches concentrate

on different characteristics of data manifolds. Local approaches are more favored in discriminating difficult classes, whereas global methods are advantageous for data representation over sequences of images.

Damianou et al. (2017) - *Manifold Alignment Determination: finding correspondences across different data views*

Summary:

Learning alignments is an important (and very challenging) task in multiview learning that we believe needs to be highlighted in order to stimulate further research. The author's present a methods for learning alignments between data points from multiple views of modalities, called *Manifold Alignment Determination (MAD)*. MAD requires only a few aligned examples from which it is capable of recovering global alignment through probabilistic models.

Multiview learning is a strand of machine learning which aims at consolidating corresponding views into a single model. The underlying idea is to exploit correspondence between the views to learn a shared representation. A very common scenario is when the views are coming from different modalities. An example is when observing a conservation where both audio and video are separately being captures. However, we can exploit the fact that the signals are aligned temporally ro obtain correspondence and thus to learn a joint representation for both signals. This scenario, however, is based on the assumption that the data-points in each view are aligned, i.e. that for each video frame there is a single corresponding sound snippet. (Connor note: What if the sampling/ polling rates are different??) **In many scenarios we do not know the correspondence between the data-points and this uncertainty should be included in the multiview model. (Connor note: this would be the perfect scenario for cross-modality MIL to select the appropriate corresponding instances.)**

In this paper, the authors exploit the regularization of a probabilistic factorized mutliview latent variable model which allows the search to be formulated as a bipartite-matching problem. Two different approaches are proposed: one myopic, which aligns data-poins in a sequential (iterative) fashion and a nonmyopic algorithm whcih produces the optimal alignment for a batch of data-points. Both implementations require a small number of initially aligned instances to act as a "prior" to dictate what an alignment means in the particular scenario. Experiments were conducted on simulated data and digits datasets. The key insight from the paper is that the combination of a factorized latent space model and Bayesian learning naturally reflects alignments in multiview data.

Shen et al. (2018) - *Manifold learning algorithms for sensor fusion of image and radio-frequency data*

Summary:

In many remote sensing applications using multiple sensor modalities, the data streams not only have high dimensionality, but also belong to different phenomena. For example, a moving vehicle may have an emitter that transmits radio-frequency (RF) signals, its exhaust system sends acoustic signals, and its perspective observed which may be collected by passive RD sensor, acoustic sensors, and video cameras. Therefore, a fusing these three sensors could increase the tracking accuracy of a moving object.

Sensor fusion includes low-level information fusion (LLIF) in which raw data is processed upstream near the sensors for object and situation assessments such as extraction of color features from pixel imagery. High-level information fusion (HLIF) includes downstream methods in which context is used for sensor, user, and mission refinement. Sensor fusion is typically performed by combining the outputs (decisions) of several signature modalities through decision-level fusion. While this data fusion approach improves performance by incorporating decisions from different modalities, it requires the consideration of the correlation/dependence between data of different modalities. **All the data in the measurement domain factually reflects the same objects of interest, which indicates that the measurements of different modalities have strong mutual information between them. The transformation from sensor data to a decision introduces information loss. How to efficiently fuse all the data of different modalities in the measurement domain with a tolerable**

cost is investigated in this paper through *feature level fusion* using the joint manifold learning (JML) approach in Davenport's paper. The numerical results show that the proposed heterogeneous data fusion approach can discover the low intrinsic dimensionalities from the high dimensional sensor data. Promising results are obtained to demonstrate the effectiveness of the manifold learning algorithms in sensor fusion, specifically on the fusion of mid-wave infrared (MWIR) and distributed radio-frequency (RF) from the DIRSIG dataset.

Hong et al. (2018) - *CoSpace: Common Subspace Learning from Hyperspectral-Multispectral Correspondences*

Summary:

A cross-modality feature learning framework called *CoSpace* is proposed which jointly considers subspace learning and supervised classification. CoSpace **linearly** learns a shared latent subspace from HSI/ multispectral correspondences. The multispectral out-of-samples can be projected into the subspace since the function is linear. Experiments on Houston and Chikusei datasets show improved classification performance.

A variety of multi-modal feature learning techniques have been proposed and can roughly be categorized into two classes: fusion-based joint feature learning (FJFL) and alignment-based shared feature learning (ASFL). FJFL aims to learn discriminative features by absorbing the different properties from multi-modal data. FJFL fuses the different sources at the data level to diversify the information and then to further learn the higher-level feature representation. **One intuitive way for FJFL is to directly learn a joint data representation at the feature level. At present, this is the mainstream approach for multimodal data analysis. However, FJFL requires *complete data correspondence*. Unlike FJFL, ASFL is more apt for cross-modal feature learning, since ASFL can adaptively shuttle back and forth between the different modalities or domains by means of the learned common subspace. Manifold alignment (MA) is also a powerful tool for ASFL which can align class distributions for robust HSI classification. It should be noted that these methods mentioned above only consider the differences of a unimodality between the source and target domains at the level of original features, but they fail to investigate the transferability of multi-modality since the different modalities usually hold the different feature dimensions.Although these approaches can build connections between features or instances, a poorly connected relationship between the learned common subspace and label information is still hindering the low-dimensional feature representation from being more discriminative.** We propose a cross-modality feature learning framework, called common subspace learning (CoSpace), that learns the shared feature representation (common subspace) from partial HS-MS correspondences.

The problem was formulated by creating a constrained optimization problem where the projection functions and weights are learned through ADMM. The goal is to minimize correspondence error while regularizing both the weight size and manifold geometry. **(This would be a good one to show in the lit review along with LeMA. Because this is involved in LeMA.)** Experiments for classification of land-use were performed with SVMs after the alignment. Results were good.

Although CoSpace's ability in handling heterogeneous data sources remains limited due to its linearized modeling, it provided an effective joint strategy for simultaneously considering subspace learning and classification by bridging the gap between the learned subspace and label information, leading to a more discriminative feature representation.

Stanley III et al. (2018) - *Manifold Alignment with Feature Correspondence*

Summary:

The authors propose a novel framework form combining datasets via alignment of their intrinsic geometry. This alignment can be used to fuse data originating from disparate modalities. **Importantly, their method does not assume any pointwise correspondence between datasets, but instead relies on correspondence between a (possibly unknown) subset of data features**.

Manifold alignment aims to map disparate datasets into a common representation, under the assumption that the datasets all originate from noisy sampling of a common manifold determined by the data generation process. Under this assumption, the intrinsic geometry of the data should be similar across datasets, and global differences between the geometry of different datasets are considered as noise or data collection artifacts. Therefore, a common representation, which aligns the datasets from their original measurements onto a shared data manifold, both eliminates such artifacts and recovers clean intrinsic relations across measured datasets. Furthermore, such alignment enables data fusion and transfer of knowledge between separate domains when datasets are taken from different data collection environments (e.g. different sensors, technologies, or subjects). We note that often we can expect such registration to be feasible, even for different technologies, since their measurements reflect related properties capturing the underlying state of equivalent scenes or subjects. **Even when using the same measurement technology, data is often systematically different based on machine calibration, day-to-day temperature variation and underlying background differences which are not always known a priori. Therefore, generalizing results across different scenes, settings, or subjects is challenging, if not impossible, without proper alignment to make varied collection comparable.**

The proposed alignment approach explicitly takes advantage of the typical correspondence between the underlying features quantified by measurement and data collection systems that can be aligned. The approach uses graph signal processing tools (specifically the *graph Fourier Transform (GFT)* which treats the eigenvectors of the graph Laplacian as generalized Fourier harmonics, i.e. intrinsic sines and cosines over a graph) to relate measured data features (observed as graph or manifold signals) to intrinsic coordinates over the intrinsic geometry of each dataset, which are revealed via *diffusion maps*. The method leverages feature correspondence to capture pairwise relations between the intrinsic diffusion map coordinates of the separate data manifolds. Finally, it uses those relations to compute an isometric transformation that aligns the data manifolds on top of each other without distorting their internal structure. **Essentially, their method is based on the principal that corresponding features across samples or datasets should have similar "frequency" components on these intrinsic data geometries, represented as manifolds. Their *harmonic alignment* leverages this understanding to compute cross-dataset similarity between manifold harmonics, which is then used to construct an isometric transformation that aligns the data manifolds. Additionally, their method aligns the data geometry, rather than density, and is thus insensitive to sampling differences.** Experiments were demonstrated on artificial manifolds created from corrupted MNIST digits and single-cell biological data for both the batch effect removal and multimodal data fusion. They were compared to the SOA approaches of ***mutual nearest neighbors (MNN)*** and ***Manifold Aligning GAN* (MAGAN)**. MAGAN is a particular cycle GAN that adds a supervised partial feature correspondence to enforce alignment of two manifolds over the mapping provided by the trained network. Mutual nearest neighbors provides locally linear alignment by calculating a correction vector for each point in the data as defined by the distances from the point to all points for which it is a mutual k-nearest neighbor. In both cases, the method successfully aligned the data manifolds to recover appropriate data neighborhoods both within and across the two datasets.

This paper gives a good summary of **graph Fourier transform and diffusion maps**.

Hong et al. (2019) - *Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification*

Summary:

In this paper, the authors aimed at tackling the cross-modality feature learning question of whether aligning a small amount of highly discriminative HSI data with copious amounts of poorly-discriminative multi-spectral data would aide in landcover classification. Therefore, they propose a novel semi-supervised cross-modality learning framework called *learnable manifold alignment (LeMA)*. LeMA learns a joint graph structure directly from the data instead of using a given fixed graph defined by a RBF kernel function. With the learned graph, we can further capture the data distribution by graph-based label propagation, which enables finding more accurate decision boundaries.

In general, a reliable classifier needs to be trained on large amounts of labeled, discriminative, and high-quality samples. Unfortunately, labeling data, in particular, large scale data, is very grueling and time consuming. Incorporating only a small amount of labeled data can lead to the poor generalization problem of over-fitting, where the few samples cannot fully represent the entirety of class variability. A natural way to alleviate this issue is to consider tons of unlabeled samples, yielding a semi-supervised learning problem.

To address the issue of limited data capture, considerable work has been done in the areas of transfer learning and domain adaptation. One such tool aiding in domain adaptation is manifold alignment and is one of the most popular semi-supervised learning frameworks that facilitates transfer learning. **The key idea of MA can be generalized as learning a common (or shared) subspace where different data can be aligned to learn a joint feature representation. Generally, unsupervised MA fails to align multimodal data sufficiently well, as their corresponding low-dimensional embeddings may be quite divers. In the supervised case, only aligning the limited number of correspondence samples often leads to weak transferability (over-fitting). Although the joint manifold structure used in conventional semi-supervised MA approaches can relate features or instances, poor connections between the common subspace and label information still hinder the low-dimensional feature representation from being more discriminative (Connor note: this is a good argument for supervised manifold learning!). More importantly, in most graph-based semi-supervised learning algorithms (e.g. graph-based label propagation (GLP)), semi-supervised manifold alignment (S-SMA), the topology of unlabeled samples is merely given by a fixed Gaussian kernel function, which is computed in the original space, rather than the common space. This makes it difficult to adaptively transfer unlabeled samples into the learned common subspace, particularly when applied to multimodal data due to different number os dimension.**

To address the aforementioned issues, the authors proposed a learn-able manifold alignment by a data-driven graph learning directly from a common subspace so as to make the multimodal data comparable as well as improve the explain-ability of the learned common subspace. This is done by extending the commonly used semi-supervised manifold alignment model where the common subspace learning problem is formulated by alternating optimization of an objective designed to both learn the mapping functions and the embedded coordinates of the training set (with label information) (This is based off the graph Laplacian and was formulated in a paper I already read. From Wang maybe?). The authors refer to this framework for joint common subspace learning as *CoSpace*. The CoSpace framework is then extended to further exploit information of unlabled samples by learning a joint Laplacian matrix. The whole process boils down to optimizing a linear objective function (so this method cannot handle non-linearity well) using the ADMM optimization approach.

Experiments were conducted on both simulated and real heterogeneous modality datasets (Houston and Chikusei multispectral-lidar and hyperspectral) and classification was compared using two high-performing classifiers after alignment, the linear SVM (LSVM) and canonical classification forests (CCF). The proposed method greatly outperformed the alternatives.

1.4 Competitive Hebbian Learning

Rumelhart and Zipser (1985) - *Feature Discovery by Competitive Learning*
Summary:

Kohonen (1990) - *The self-organizing map*
Summary:

The self-organizing map (SOM) creates spatially organized intrinsic representations of features. It belongs to the category of neural networks which use “competitive learning”, or “self-organization”. It is a sheet-like artificial neural network in which the cells become tuned to various input patterns through an unsupervised learning process. Only a neighborhood of cells give an active response to the current input sample. The

spatial location or coordinates of cells in the network correspond to different modes of the input distribution. The self-organizing map is also a form of vector quantization (VQ). The purpose of VQ is to approximate a continuous probability density function $p(\boldsymbol{x})$ of input vectors $\boldsymbol{x}$ using a finite number of codebook vectors, $\boldsymbol{m}_i$, $i = 1, 2, \ldots, k$. After the "codebook" is chosen, the approximation of $\boldsymbol{x}$ involves finding the reference vector, $\boldsymbol{m}_c$ closest to $\boldsymbol{x}$. The "winning" codebook vector for sample $\boldsymbol{x}$ satisfies the following:

$$||\boldsymbol{x} - \boldsymbol{m}_c|| = \min_i ||\boldsymbol{x} - \boldsymbol{m}_i||$$

The algorithm operates by first initializing a spatial lattice of codebook elements (also called "units"), where each unit's representative is in $\boldsymbol{m}_i \in \mathbb{R}^D$ where D is the dimensionality of the input samples $\boldsymbol{x}$. The training process proceeds as follows. A random sample is selected and presented to the network and each unit determines its activation by computing dissimilarity. The unit who's codebook vector provides the smallest dissimilarity is referred to as the *winner*.

$$c(t) = \arg\min_i d(\boldsymbol{x}(t), \boldsymbol{m}_i(t))$$

Both the winning vector and all vectors within a neighborhood of the winner are updated toward the sample by

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [\boldsymbol{x}(t) - \boldsymbol{m}_i(t)]$$

where $\alpha(t)$ is a learning rate which decreases over time and $h_{ci}(t)$ is a neighborhood function which is typically unimodal and symmetric around the location of the winner which monotonically decreases with increasing distance from the winner. A radial basis kernel is typically chosen for the neighborhood function as

$$h_{ci}(t) = \exp\left(-\frac{||\boldsymbol{r}_c - \boldsymbol{r}_i||^2}{2\sigma^2(t)}\right)$$

where the top expression represents the Euclidean distance between units c and i with $\boldsymbol{r}_i$ representing the 2-D location of unit i in the lattice. The neighborhood kernel's bandwidth is typically initialized to a value which covers a majority of the input space and decreases over time such that solely the winner is adapted toward the end of the training procedure.

The SOM essentially performs density estimation of high-dimensional data and represents it in a 2 or 3-D representation. At test time, the dissimilarity between each unit in the map and an input sample are computed. This dissimilarity can be used to effectively detect outliers, thus making the SOM a robust method which can provide confidence values for it's representation abilities.

In this paper, the SOM was applied to speech recognition, but made note of previous uses in robotics, control of diffusion processes, optimization problems, adaptive telecommunications, image compression, sentence understanding, and radar classification of sea-ice.

Rauber et al. (2002) - *The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data*

*Summary:*The Growing Hierarchical Self-organizing Map (GHSOM) is an extension of the classical SOM. It is an artificial neural network with a hierarchical architecture, composed of individually growing SOMs. Layer 0 is composed of a single neuron representing the mean of the training data. A global stopping criteria is developed as a fraction of the mean quantization error. This means that all units must represent their respective subsets of data an a MQE smaller than a fraction of the 0 layer mean quantization error. For all units not satisfying this criteria, more representation is required for that area of the feature space and additional units are added. After a particular number of training iterations, the quantization errors are computed and the unit with the highest error is selected as the *error unit*. The most dissimilar neighbor of the error unit is chosen is and a row/ column of nodes is injected between them. The growth process continues until a second stopping criteria is met. Any units still not satisfying the global criteria are deemed to need extra representation. Child map are initialized below these units and trained with the subset of data mapped to its parent node.

In conclusion, the GHSOM is a growing self-organizing map architecture which has the ability to grow itself until the feature space is adequately represented. For areas of the space needing a more specific level of granularity, a hierarchical structure is imposed to "fill-in" areas of high density.

The GHSOM has been applied to the areas of finance, computer network traffic analysis, manufacturing and image analysis (Palomo 2017).

Chiang and Gader (1997) - *Hybrid fuzzy-neural systems in handwritten word recognition*
Summary:

Frigui and Gader (2009) - *Detection and Discrimination of Land Mines in Ground-Penetrating Radar Based on Edge Histogram Descriptors and a Possibilistic K-Nearest Neighbor Classifier*
Summary:

Fritzke (1994) - *A Growing Neural Gas Network Learns Topologies*
*Summary:*Abstract: An incremental network model is introduced which is able to learn the important topological relations in a given set of input vectors by means of a simple Hebb-like learning rule. In contrast to previous approaches like the "neural gas" method of Martinetz and Schulten (1991, 1994), this model has no parameters which change over time and is able to continue learning, adding units and connections, until a performance criterion has been met. Applications of the model include vector quantization, clustering, and interpolation.

In contrast to SOMs and "growing cell structures", which can project data onto non-linear subspaces which are chosen *a priori*, the GNG is able to adapt its topology to match that of the input data distribution. The growing process continues until a pre-defined level of quantization error has been reached.

The base algorithm is outlined in Palomo (2017), *Growing Hierarchical Neural Gas Self-Organizing Network*.

Palomo and Lopez-Rubio (2017) - *The Growing Hierarchical Neural Gas Self-Organizing Neural Network*
Summary:

Abstract: The growing neural gas (GNG) self-organizing neural network stands as one of the most successful examples of unsupervised learning of a graph of processing units. Despite its success, little attention has been devoted to its extension to a hierarchical model, unlike other models such as the self-organizing map, which has many hierarchical versions. Here, a hierarchical GNG is presented, which is designed to learn a tree of graphs. Moreover, the original GNG algorithm is improved by a distinction between a growth phase

where more units are added until no significant improvement in the quantization error is obtained, and a convergence phase where no unit creation is allowed. This means that a principled mechanism is established to control the growth of the structure. Experiments are reported, which demonstrate the self-organization and hierarchy learning abilities of our approach and its performance for vector quantization applications. Experiments were performed in structure learning, color quantization, and video sequence clustering.

The aim of this method was to improve the adaptation ability of the Growing Hierarchical Self-Organizing Map proposed by Rauber (2002). This was to be done through the extension of the Growing Neural Gas, which disposes of the fixed lattice topology enforced by the SOM. Addtionally, the GNG learns a dynamic graph with variable numbers of neurons and connections. The graph represents the input data in a more plastic and flexible way than the fixed-topology map.

All clustering methods that learn a hierarchical structure have advantages even when used for non-hierarchical data. The learned hierarchical structure can be pruned at several levels, which yields alternative representations of the input data set at different levels of detail. This can be used to visualize a data set in coarser or more detailed way. For vector quantization applications, the different pruning levels correspond to smaller or larger codebooks, so that a balance can be attained between the size of the codebook and the quantization error within the same hierarchical structure.

The growing hierarchical neural gas (GHNG) model is defined as a tree of self-organizing graphs. Each graph is made of a variable number of neurons or processing units, so that its size can grow or shrink during learning. In addition, each graph is the child of a unit in the upper level, except for the top level (root) graph. The training procedure is described by the following:

Each graph begins with $H \geq 2$ units and one or more undirected connections between them. Both the units and connections can be created and destroyed during the learning process. It is also not necessary that the graph is connected. Let the training set be denoted as $\mathcal{S}$ with $\mathcal{S} \subset \mathbb{R}^D$, where D is the dimensionality of the input space. Each unit $i \in \{1, \ldots, H\}$ has an associated prototype $\boldsymbol{w}_i \in \mathbb{R}^D$ and an error variable $e_i \in \mathbb{R}$, $e_i \geq 0$. Each connection has an associated age, which is a nonnegative integer. The set of connections will be notetd as $A \subseteq \{1, \ldots, H\} \times \{1, \ldots, H\}$. The learning mechanism for the GHNG is based on the original GNG, but includes a novel procedure to control the growth of the graph. First, a growth phase is performed where the graph is allowed to enlarge until a condition is met, which indicates that further growing would provide no significant improvement in the quantization error. After that, a convergence phase is executed where no unit creation is allowed in order to carry out a fine tuning of the graph. the leraning algorithm is provided in the following steps.

1. Start with two units ($H = 2$) joined by a connection. Each prototype is initialized to a sample drawn at random from $\mathcal{S}$. The error variables are initialized to zero. The age of the connection is initialized to zero.
2. Draw a training sample $\boldsymbol{x}_t \in \mathbb{R}^D$ at random from $\mathcal{S}$.
3. Find the nearest unit q and second nearest unit s in terms of Euclidean distance

$$q = \underset{i \in \{1,\ldots,H\}}{\arg\min} \; ||\boldsymbol{w}_i(t) - \boldsymbol{x}(t)||$$
$$s = \underset{i \in \{1,\ldots,H\} - \{q\}}{\arg\min} \; ||\boldsymbol{w}_i(t) - \boldsymbol{x}(t)||$$

4. Increment the age of all edges departing from q
5. Update the winning unit's error variable, e_q

$$e_q(t+1) = e_q(t) + ||\boldsymbol{w}_q(t) - \boldsymbol{x}_t||$$

I believe the author's experimental approach did not take advantage of the method's strengths. The author's only demonstrated experiments in vector quantization, and used corresponding metrics. This method could

be used to represent manifold topology of differing dimensionality. This could be useful in HSI imagery, for example where different environment patches require manifold representations of various dimensionality. Additionally, this could potentially be used to handle the sensor fusion problem with sensor loss/ drop-out.

Sun et al. (2017) - *Online growing neural gas for anomaly detection in changing surveillance scenes*
Summary:

Lopez-Rubio and Palomo (2011) - *Growing Hierarchical Probabilistic Self-Organizing Graphs*
Summary:

Palomo and Lopez-Rubio (2016) - *Learning Topologies with the Growing Neural Forest*
Summary:

1.5 Deep Learning

Goodfellow et al. (2016) - *Deep Learning*
Summary:

Haykin (2009) - *Neural networks and learning machines*
Summary:

Dai et al. (2017) - *dden Talents of the Variational Autoencoder*
Summary:

Rojas (1996) - *Associative Networks*
Summary:

2 Information Measures

Arandjelovic et al. (2005) - *Face recognition with image sets using manifold density divergence*
Summary:

Wang et al. (2012) - *ManifoldManifold Distance and its Application to Face Recognition With Image Sets*
Summary:

3 Manifold Regularization

Tsang and Kwok (2007) - *Large-Scale Sparsified Manifold Regularization*
Summary:

Ren et al. (2017) - *Unsupervised Classification of Polarimetirc SAR Image Via Improved Manifold Regularized Low-Rank Representation With Multiple Features*
Summary:

Belkin et al. (2006) - *Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples*
Summary:

Ratle et al. (2010) - *Semisupervised Neural Networks for Efficient Hyperspectral Image Classification*
Summary:

Li et al. (2015) - *Approximate Policy Iteration with Unsupervised Feature Learning based on Manifold Regularization*
Summary:

Meng and Zhan (2018) - *Zero-Shot Learning via Low-Rank-Representation Based Manifold Regularization*

Summary:

4 Multiple Instance Learning

4.1 Multiple Instance Concept Learning

Bocinsky (2019) - *Learning Multiple Target Concepts from Uncertain, Ambiguous Data Using the Adaptive Cosine Estimator and Spectral Match Filter*
Summary:

Jiao (2017) - *Target Concept Learning From Ambiguously Labeled Data*
Summary:

McCurley et al. (2019) - *Comparison of hand-held WEMI target detection algorithms*
Summary:

Bocinsky et al. (2019) - *Investigation of initialization strategies for the Multiple Instance Adaptive Cosine Estimator*
Summary:

Zare et al. (2015) - *Multiple instance dictionary learning for subsurface object detection using handheld EMI*
Summary:

Cook (2015) - *Task driven extended functions of multiple instances (TD-eFUMI)*
Summary:

Cook et al. (2016) - *Buried object detection using handheld WEMI with task-driven extended functions of multiple instances*
Summary:

Zare et al. (2016) - *Multiple Instance Hyperspectral Target Characterization*
Summary:

Jiao and Zare (2017) - *Multiple instance hybrid estimator for learning target signatures*
Summary:

Xiao et al. (2017) - *A Sphere-Description-Based Approach for Multiple-Instance Learning*
Summary:

Cheplygina et al. (2019) - *Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis*
Summary:

Li et al. (2017) - *Cross-validated smooth multi-instance learning*
Summary:

Cheplygina et al. (2016) - *Dissimilarity-Based Ensembles for Multiple Instance Learning*
Summary:

Wang et al. (2017) - *Incorporating Diversity and Informativeness in Multiple-Instance Active Learning*
Summary:

Hajimirsadeghi and Mori (2017) - *Multi-Instance Classification by Max-Margin Training of Cardinality-Based Markov Networks*
Summary:

Du et al. (2016) - *Multiple Instance Choquet integral for classifier fusion*
Summary:

Ilse et al. (2018) - *Attention-based Deep Multiple Instance Learning*
Summary:

Karem and Frigui (2016) - *Multiple Instance Learning with multiple positive and negative target concepts*
Summary:

Xiao et al. (2017) - *Multiple-Instance Ordinal Regression*
Summary:

Gao et al. (2017) - *C-WSL: Count-guided Weakly Supervised Localization*
Summary:

Li et al. (2017) - *Multi-View Multi-Instance Learning Based on Joint Sparse Representation and Multi-View Dictionary Learning*
Summary:

Cao et al. (2016) - *Weakly Supervised Vehicle Detection in Satellite Images via Multi-Instance Discriminative Learning*
Summary:

Dietterich et al. (1997) - *Solving the multiple instance problem with axis-parallel rectangles*
Summary:

Maron and Lozano-Pérez (1998) - *A Framework for Multiple-instance Learning*
Summary:

Maron and Ratan (1998) - *Multiple-Instance Learning for Natural Scene Classification*
Summary:

Carbonneau et al. (2016) - *Multiple Instance Learning: A Survey of Problem Characteristics and Applications*
Summary:

Zhang and Goldman (2002) - *EM-DD: An Improved Multiple-Instance Learning Technique*
Summary:

Zare and Jiao (2014) - *Extended Functions of Multiple Instances for target characterization*
Summary:

Jiao et al. (2018) - *Multiple instance hybrid estimator for hyperspectral target characterization and sub-pixel target detection*
Summary:

4.2 Multiple Instance Classification

Cao et al. (2016) - *Weakly Supervised Vehicle Detection in Satellite Images via Multi-Instance Discriminative Learning*
Summary:

4.3 Multiple Instance Regression

Trabelsi and Frigui (2018) - *Fuzzy and Possibilistic Clustering for Multiple Instance Linear Regression*
Summary:

Ruiz et al. (2018) - *Multi-Instance Dynamic Ordinal Random Fields for Weakly Supervised Facial Behavior Analysis*
Summary:

4.4 Applications

5 Fusion

5.1 Classical Approaches

5.1.1 General Approach

Mohandes et al. (2018) - *Classifiers Combination Techniques: A Comprehensive Review*
Summary:

Should reference this paper for hierarchical representations of classifier combination methods. Many good diagrams. Combining expert opinions before making decisions can substantially increase the reliability of critical application systems such as medical diagnosis, security, and so on. Evidence from multiple classifiers can be combined on the data, feature, or decision level. Classifier ensemble combination methods are known under many different names: multi-classifier combination, multi-classifier fusion, mixture of experts, and ensemble based classification, to name a few. Since the most recent review, 8 years earlier, a few more methods for classifier combination were introduced, including: a signal strength based combination approach, a novel Bayes voting strategy, a modified weighted averaging technique using graph-theoretic clustering, a neural network based approach for training combination rules, weighted feature combination, and hierarchical fuzzy stack generalization.
Typical classifier combination algorithms begin with a set of scores from individual classifiers and produce a combined score for each class along with a final class label. The problem then generalizes to finding a combination function with accepts a K dimensional score vector from each of the M classifiers, then produces a single, final classification score representing the selected class. The M classifiers could be identical but use different feature sets as inputs, or use different parameter sets. Alternatively, the classifiers could be different by nature but use the same set of input features. The important distinction is that individual classifiers should not make identical erroneous decisions on the same observation set, i.e. they should provide complementary information.
Most classifier combination techniques assume independence between features.
Adaptive techniques for classifier fusion are mainly based on evolution or artificial intelligence algorithms. They include neural network combination strategies and genetic algorithms as well as fuzzy set theory. Fusion using ANNs allows for non-linear combination of classifier outputs. Adaptive methods also include adaptive weighting, associative switching, adaptive fuzzy integrals, mixture of local experts and hierarchical MLE. Adaptive classifiers tend to do better than the non-adaptive type.
Describes the differences between bagging, boosting, AdaBoost, and HME. *Bagging* is creating different datasets by bootstrapped versions of the original dataset (sampling with replacement). In *boosting*, individual classifiers are trained hierarchically to discriminate more complex regions of the feature space. *AdaBoost* is a variation of boosting which combines the outputs of weak classifiers into a weighted sum representing the final decision. However, it is sensitive to noisy data and outliers. Additionally, AdaBoost on the feature level falls victim to the curse of dimensionality.
Classifier fusion methods have been used on HSI data, to improve accuracy when sensor data is subjected to drift, handwritten word recognition, sequential data with HMM classifiers only,etc.
The literature still lacks a comprehensive performance analysis of techniques for a given application. Important research questions still include: classifier post-processing before combination, using meta-heuristic algorithms to improve performance, such as using optimization algorithms with majority voting, showing the advantages/ disadvantages of using different strategies such as probabilistic, learning, decision based, or evidence based, additionally, finding the optimal number/ type of classifiers to fuse is an open question.

Ruta and Gabrys (2000) - *An Overview of Classifier Fusion Methods*

Summary:

"The objective of all decision support systems (DSS) is to create a model, which given a minimum amount of input data/information, is able to produce correct decisions." "the solution might be just to combine existing, well performing methods, hoping that better results will be achieved. Such fusion of information seems to be worth applying in terms of uncertainty reduction. Each of individual methods produces some errors, not mentioning that the input information might be corrupted and incomplete. However, different methods performing on different data should produce different errors, and assuming that all individual methods perform well, combination of such multiple experts should reduce overall classification error and as a consequence emphasize correct outputs." "Fusion of data/information can be carried out on three levels of abstraction closely connected with the flow of the classification process: data level fusion, feature level fusion, and classifier fusion" This paper focused on the later method of classifier fusion. This process can essentially be categorized into two eruditions. The first methods put emphasis on the classifier structure and do not do anything with the outputs until the combination process finds the best classifier or a selected group of classifiers. Then their outputs are taken as a final decision or used for further processing. The second category operates primarily on classifier outputs and can be further divided.

There are three possibles types of output labels generated by individual classifiers. Crisp labels provide the lowest amount of information for fusion, as no information about potential alternatives is available. Some additional information can be gleaned from labels in the form of class rankings. However, fusion methods operating on classifiers with soft/fuzzy outputs can be expected to produce the greatest improvement in classification performance. (Connor Note: This is valuable in terms of outlier rejection as well!). The following explains an overview of classifier fusion methods operating on single class labels, class rankings, and fuzzy measures, respectively.

Methods operating on classifiers:

Dynamic Classifier Selection (DCS) methods replect the tendency to extract a single best classifier instead of mixing many different classifiers, by attempting to determing the single classifier which is most likely to produce the correct classification label for an input sample. Only the output of the selected classifier is taken as a final decision. The classifier selection process includes a partitioning of the input samples. A classifier is is for each partition is selected locally. All DCS methods rely on strong training data and by choosing only locally best classifier. **They potentially lose some useful information from other well-performing classifiers.** Classifiers and their combination functions are typically organized in parallel and simultaneously and separately get their outputs as in input for a combination function. A more reasonable approach, however, is **to organize all classifiers into groups and to apply different fusion methods for each group.** A very important factor for the success of this method is the diversity of classifier types, training data, and methods involved. **Any classification improvement may only be achieved if the total information uncertainty is reduced.** This in turn depends on the diversity of information supporting different classification methods. **The same goal can be achieved by reduction of errors produced by individual classifiers.** *Hierarchical Mixture of Experts* (HME) is an example of a fusion method whose strength comes from classifier's structure. It is a supervised learning method based on the *divide-and-conquer* principle. It is organized as a tree-like structure of leaves. Each leaf represents an individual expert in the network, each of which tries to solve a local supervised learning problem. The outputs of the elements of the same node are partitioned and combined by the gating network and the total output of the node is given as a convex combination. The expert networks are trained to increase the posterior probability according to Bayes rule. A number of learning algorithms can be applied to tune the mixture model. *Expectation-Maximization* (EM) is often used to learn the model parameters. *The HME technique does not seem to be applicable to large-dimensional datasets.*

Fusing Single Class Labels: Classifiers producing crisp, single-class labels (SCL) provide the least amount of useful information for the combination process. The two most common techniques for fusing SCL classifiers are *Generalized Voting* and *Knowledge-Behavior Space* methods.

Voting Methods:

Voting strategies can be applied to a multiple classifier system assuming that each classifier gives a single class label as as output and no training data are available. While there are many methods for combining

these labels, they all lead to the following generalized voting definition. Let the output of the classifiers form the decision vector $\boldsymbol{d} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^T$ where $\boldsymbol{d}_i \in \{c_1, c_2, \ldots, c_m, r\}$, c_i denotes the class label of the i-th class and r the rejection of assigning the input sample to any classes. The binary characteristic function is defined as follows: (Have not input math)

Class Ranking Based Techniques:

There are two primary methods for fusion of class rankings. *Class set reduction* (CSR) attempts to reduce the number of eligible classes by compromising between minimizing the class set size and maximizing the likelihood of inclusion in the true class. This is typically performed through the *intersection or union of neighborhoods*. The second popular CSR method is *Class Set Reordering* (CSRR) which tries to improve the overall rank of the true class through techniques such as the *Highest Rank Method*, *Borda Count*, or *Logistic Regression*.

Soft-Label Classifier Fusion:

Soft labels are outputs in the range $[0, 1]$ and are typically referred to as *fuzzy measures*, which cover all known measures of evidence: probability, possibility, necessity, belief, and plausibility. Each of these measures are used to describe different dimensions of information uncertainty. This class of fusion attempts to reduce the level of uncertainty by maximizing suitable measures of evidence. Common methods for this type of fusion include: Bayesian, Fuzzy Integrals, Dempster-Shaffer Combination, Fuzzy Templates, Product of Experts, and Artificial Neural Networks. *Bayesian* methods can be applied under the condition that the outputs of the classifier are expressed as posterior probabilities. Typical methods of Bayesian fusion include Bayes Average and Bayes Belief Integration. *Fuzzy Integrals* aim at searching for the maximal agreement between the real possibilities relating to objective evidence and the expectation, g, which defines the level of importance of a subset of sources. The concept of fuzzy integrals arises from the λ-fuzzy measure, g, developed by Sugeno. Common methods for Fuzzy Integration include the Sugeno Fuzzy Integral, Choquet Fuzzy Integral, and Webster Fuzzy Integral. *Product of Experts* combines different probabilistic models of the same data by performing a weighted average of individual probability distributions.

Tulyakov et al. (2008) - *Review of Classifier Combination Methods*

Summary:

This paper provides different categorizations of classifier combination methods, including ensemble and non-ensemble based techniques. Simple combination functions include sum, wighted sum, max etc. More complex functions include ANN, k-NN and the like. Classifier fusion methods either assume a small number of classifier inputs/ sensors which obtain their performance benefit from information diversity, or they rely on large sets of ensembles which operate using different features/ areas of the input space, modalities, etc.

Fusion methods can operating on the *feature level* aim to form a joint feature vector, and subsequently perform classification in the new feature space. This approach can potentially provide additional information about the classes, but often requires a large training set due to the increased dimensionality of the joint feature space. It should be noted that if the features used in the different classifiers are not related, then there is no real benefit for fusion on the feature level. Additionally, this method is not conceptually different from attempting to incorporate information from alternative sources into a single feature vector.

Methods operating on the *score level* fuse the outputs of individual classifiers. It is possible, in this approach, that information is lost during combination. This, however, is usually compensated by the lower computational cost of combination and superior training of the final system.

Classifiers can be organized based on their outputs. The three main types are *abstract, rank, and measurement*. Abstract provides the lowest amount of information and is simply a single class label or unordered set of candidate classes. Rank level outputs provide an ordered sequence of candidate classes, also called the *n-best list*. The candidate class at the first position is the most likely class while the class at the end is the most unlikely. There are no confidence values attached to the class labels. Measurement level outputs provide an ordered n-best list along with corresponding confidence levels. Measurement level outputs provide the most information of the three output types. While fusion on the measurement level is desired, it might be difficult to achieve since confidence values from different types of classifiers may not align easily (i.e. different ranges, scales, means, etc).

Additionally, fusion methods can be classified based on their *complexity*. Simple combination rules such as sum, weighted sum, product, etc exhibit low complexity, while rank based methods such as the Borda

count represent medium complexity. An example of a high complexity combination type includes Behavior-knowledge spaces (BKS).
While there is substantial research on classifier ensembles, there are very few theoretical results explaining why they work. Most explanations use bias and variance. However, such approaches can only provide asymptotic explanations of observed performance improvements. **Ideally, the theoretical foundation for classifier ensembles should use statistical learning theory.** The following few sections analyze performance improvements due to ensemble methods.
A good description of bagging a nd boosting is given.
Compared to ensemble-based classifiers, non-ensemble based methods attempt to combine heterogeneous classifiers which complement each other. **The advantage of complementary classifiers is that each classifier can concentrate on its own small subproblem instead of trying to cope with the classification problem as a whole, which may be too difficult for a single classifier. Ideally, the expertise of the specialized classifiers do not overlap.**
Normalization methods are described and corresponding advantages/ dis-advantages are elucidated.
A classic approach to classifier combination is the ***Dempster-Shafer theory of evidence*** **(DS)**. It was originally adopted by researchers in AI in order to process probabilities in expert systems, but has recently been adopted to sensor fusion and classifier combination. DS theory is a generalization of the Bayesian theory of probability and differs in several aspects. First, DS theory introduces *degrees of belief* that do not necessarily meet the mathematical properties of probabilities. Second, it assigns probabilities to sets of possible outcomes rather than single events only. Third, it considers probability intervals that contains the precise probability for sets of possible outcomes. The two main ideas of DS theory are to obtain degrees of belief for one question from subjective probabilities for a related question, and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence. Dempster's rule of combination is a generalization of Baye's rule. Dempter's rule defines the joint mass $m_{1,2}(X)$ for an outcome set X as follows:

$$m_{1,2}(X) = \begin{cases} 0 & \text{if} \quad X = \emptyset \\ \frac{1}{1-K} \sum_{A_i \cap B_j = X} m_1(A_i) m_2(B_j) & \text{if} \quad X \neq \emptyset \end{cases}$$

where

$$K = \sum_{A_i \cap B_j = X} m_1(A_i) m_2(B_j)$$

DS theory has produced good results in document processing and is still used today.
Another complex, and popular, approach is the ***Behavior-Knowledge Space (BKS)*** **method**. BKS is a trainable combination scheme on the abstract level which requires neither measurements not ordered sets of candidate classes. It tries to estimate the a posteriori probabilities by computing the frequency of each class for every possible set of classifier decisions, based on a given training set. The result is a lookup table that associates the final classification result with each combination of classifier outputs, i.e. each combination of outputs in the lookup table is represented by its most often encountered class label. Given a specific classifier decision $S_1, \ldots, \S_M$ from M individual classifiers, the a posteriori probability $\hat{P}(c_i|S_1, \ldots, \S_M)$ of class c_i is estimated as follows:

$$\hat{P}(c_i|S_1, \ldots, S_M) = \frac{N(c_i|S_1, \ldots, S_M)}{\sum_j N(c_j|S_1, \ldots, S_M)}$$

where $N(c_i|S_1, \ldots, S_M)$ counts the frequency of class c_i for each possible combination of crisp classifier outputs.

Hackett and Shah (1990) - *Multi-sensor fusion: a perspective*
Summary:

Multi-Sensor fusion deals with the combination of complementary and sometimes competing sensor data into a reliable estimate of the environment to achieve an output which is better than the modalities, individually. Multi-sensor fusion has been used in target recognition, autonomous robot navigation, automatic manufacturing, scene segmentation, sensor modeling, and object recognition. *Sensor fusion combines the outputs from two or more devices that retrieve a particular property of the environment.* Each sensor's measurements are, in general, imprecise and contain errors and uncertainties, so the consensus of multiple sensors measuring the same property can reduce uncertainty and reduce measurement ambiguity. Every sensor modality is sensitive to a different property of the environment; it is necessary to use multiple sensors in order to address these sensitivities. *Sensor fusion deals with the selection of a proper model for each sensor, and identification of an appropriate fusion method.* There are several methods for combining multiple data sources. A few are: deciding, guiding, averaging, Bayesian statistics, and integration. Deciding is the use of a particular data source during a certain time of the fusion process, usually based on some confidence measure. Averaging is the weighted combination of several data sources. This type of fusion ensures all sensors contribute to the fusion process, but not all to the same degree. Guiding is the use of one or more sensors to focus the attention of another sensor on some part of the scene. Integration is the delegation of carious sensors to particular tasks, thus eliminating redundancy in sensor measurements. The most simple method of fusion uses raw data of the same property obtained by multiple sensors of the same type. Multi-sensor integration is the use of several sensors in a sequential manner.
Data from different sensors must be put into equivalent forms to allow for fusion. In order for data from multiple sources to be fused, there must be some method to relate data points from one sensor with corresponding data points from the other sensors. The *registered* data points allow for easy gathering of sensor information about one particular point in the scene.
Fusion methods can be broadly classified into two categories, *direct* and *indirect*. Direct fusion combines raw sensor measurements while indirect methods transform the sensor data to be fused.
Before sensor measurements can be combined, we must ensure that the measurements represent the same physical entity. Therefore, we need to check the consistency of sensor measurements. One such method for checking measurement consistency is the *Mahalanobis* distance.
Since each sensor is sensitive to a different modality, multiple sensors not only can provide multiple views of objects, but they can also impose more constraints to reduce the search space during matching.

Zhang (2010) - *Multi-source remote sensing data fusion: Status and trends*
Summary:

Remote sensing data fusion, as one of the most commonly used techniques for fusion, aims to integrate the information acquired with different spatial and spectral resolutions from sensors mounted on satellites, aircraft and ground platforms to produce fused data that contains more detailed information than each of the sources, individually. Fusing remotely sensed data, especially multi-source data, remains challenging due to reasons such as landscape complexity, temporal and spectral variations, and accurate data co-registration. *Pixel level* fusion is the combination of raw data from multiple sources into single resolution data, which are expected to be more informative and synthetic than either of the input data or reveal the changes between data sets acquired at different times. *Feature level* fusion extracts various features, e.g. edges, corners, lines, texture parameters, etc., from different data sources and then combines them into one or more feature maps that may be used instead of the original data for further processing. This is particularly important when the number of available spectral bands becomes so large that it is impossible to analyze each band separately. Methods applied to extract features usually depend on the characteristics of the individual source data, and therefore may be different if the data sets used are heterogeneous. Typically, in image processing, such fusion requires a precise (pixel-level) registration of the available images. Feature maps thus obtained are then used as input to pre-processing for image segmentation or change detection. *Decision level* fusion combines the results from multiple algorithms to yield a final fused decision. When the results from different algorithms are expressed as confidences (or scores) rather than decisions, it is called soft fusion; otherwise, it is called hard fusion.

Methods of decision fusion include voting methods, statistical methods and fuzzy logic based methods. PROVIDES A GREAT DESCRIPTION OF LiDAR USE DESCRIPTION FROM XIAOXIAO'S DISSERTATION.
An undesirable property when applying pixel-level fusion techniques to the fusion of SAR and optical images is that either spectral features of the optical imageery or the microwave backscattering information is destroyed, or both simultaneously.
Applications: satellite Earth observations, computer vision, medical image processing, defense security, land use classification, Digital Surface Modeling (DSM), Digital Elevation Modeling (DEM), environmental monitoring, road mapping. archeology, building detection and reconstruction, etc.
For specific purposes, ancillary and terrestrial meta-data such as laser-scanners, GIS data, web-sensors, field survey data, economic consensus data, and meteorological data me be combined with remote sensing data to improve the performance of data fusion.

5.1.2 Hierarchical Mixture of Experts

Jordan and Jacobs (1993) - *Hierarchical mixtures of experts and the EM algorithm*
Summary:

Yuksel et al. (2012) - *Twenty Years of Mixture of Experts*
Summary:

Beyer et al. (2009) - *Heterogeneous mixture-of-experts for fusion of locally valid knowledge-based submodels*
Summary:

Shazeer et al. (2017) - *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*
Summary:

5.1.3 Choquet Integral

Du (2017) - *Multiple Instance Choquet Integral For MultiResolution Sensor Fusion*
Summary:

Gader et al. (1996) - *Fusion of handwritten word classifiers*
*Summary:*This paper compares fusion methods for handwritten word classifiers. They found that a novel approach to applying a Choquet fuzzy integral outperformed other methods such as neural networks, Borda count, weighted Borda count and Sugeno fuzzy integral.

Ryan E. Smith (2017) - *Aggregation of Choquet integrals in GPR and EMI for handheld platform-based explosive hazard detection*
Summary:

Smith et al. (2017) - *Genetic programming based Choquet integral for multi-source fusion*
Summary:

Du and Zare (2019) - *Multiple Instance Choquet Integral Classifier Fusion and Regression for Remote Sensing Applications*
Summary:

Anderson et al. (2017) - *Binary fuzzy measures and Choquet integration for multi-source fusion*
Summary:

Du and Zare (2018) - *Multi-Resolution Multi-Modal Sensor Fusion For Remote Sensing Data With Label Uncertainty*
Summary:

Gader et al. (2004) - *Multi-sensor and algorithm fusion with the Choquet integral: applications to landmine detection*
Summary:

Keller et al. (2001) - *Experiemnts in predictive sensor fusion*
*Summary:*They use a Choquet fuzzy integral to predict the value of fusing sensors for landmine detction.

Gunatilaka and Baertlein (2001) - *Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection*
Summary:

Frigui et al. (2010) - *Context-Dependent Multisensor Fusion and Its Application to Land Mine Detection*
Summary:

5.1.4 Deep Learning

L.Jian et al. (2019) - *A Symmetric Encoder-Decoder with Residual Block for Infrared and Visible Image Fusion*
Summary:

5.1.5 Graph-Based

Vivar et al. (2019) - *Multi-modal Graph Fusion for Inductive Disease Classification in Incomplete Datasets*
Summary:

5.2 Fusion Metrics

5.3 Co-registration

Dawn et al. (2010) - *Remote Sensing Image Registration Techniques: A Survey*
Summary:

Brigot et al. (2016) - *Adaptation and Evaluation of an Optical Flow Method Applied to Coregistration of Forest Remote Sensing Images*
Summary:

Zitov and Flusser (2003) - *Image registration methods: a survey*
Summary:

Image registration is the process of overlaying images (two or more) of the same scene taken at different times, from different viewpoints, and/or by different sensors. Image registration can broadly be broken into two categories, *area-based* and *feature-based* and according to four basic steps of image registration: *feature detection, feature matching, mapping function design, and image transformation and resampling.*
Image acquisition can be divided into four methodologies. *Different viewpoints (multiview analysis)* involves collecting images of the same scene from different viewpoints. *Different times (multitemporal analysis)* collects images of the same scene acquired at different times, often on a regular basis. The aim is to find and evaluate changes in the scene over time. *Differnt sensors (multimodal analysis)* collects images of the same scene through different sensors. The aim is to integrate information obtained from different source streams to gain a more complex and detailed scene representation. *Scene to model registration* involves images of a scene and a corresponding model (such as digital elevation models (DEM)). The objective is to localize the acquired image in the scene/ model and/ or compare them.
Feature detection involves finding salient and distinctive points of interest in an image (closed-boundary regions, edges, contours, corners, line intersections, etc.) These features can be represented by their point

representatives (centers of gravity, line endings, distinctive points) called *control points* (CP). Physically corresponding features can be dissimilar due to different imaging conditions and/or due to differing spectral sensitivities among sensors. Features should be sufficiently robust and stable as to not be influenced by unexpected variations a nd noise. *Feature matching* involves pairing corresponding features between the query and reference image. Various features and (dis)similarity measures along with spatial relationships are employed during this process. Classical area-based feature matching involves methods such as cross-correlation (CC) (which does not incorporate any structural analysis) and mutual information (MI) which measures the statistical dependency between two datasets. Other feature-based matching methods include graph matching, clustering, Iterative Closest Points (ICP) and more. Features should be invariant , unique, stable, and independent. Area-based matching is preferable when images do not have many prominent details/ distinctive information. However, they rely on the images having similar intensity functions. Feature-based matching are typically applied when local structure information is more significant than the information carried by the image intensities. These methods allow images of completely different natures to be registered (i.e. multi-modal). However, the drawback of this class of matching is that the respective featues might be difficult to detect between images. *Transform model estimation* involves selecting the type and parameters of a *mapping function* to align the sensed image with the reference. Mapping models are either Global (which use all CPs) or Local (which decompose the image into patches and the function parameters are locally dependent to a patch.) Global methods preserve shape (curvatures, angles, etc), while local methods allow for local image transformations, thus addressing local deformations. Types of mapping functions include: bivariate polynomials, radial basis functions, elastic, fluid, diffusion based, level sets, and optical flow registration. *Image resampling and transformation* involves transforming an image by means of the mapping function. The main limitation of image resampling is computational complexity, especially for high dimensional images.
Accuracy evaluation is a non-trivial problem, partially because errors can evolve into the registration process in each of its stages and partially because it is difficult to distinguish between registration inaccuracies and actual physical differences. Accuracy error is usually measured by local *local error* (displacement of CP coordinates due to inaccurate detection), *matching error* (measured by the number of false matches when establishing the correspondence between CP candidates), and *alignment error* (the difference between the mapping model used for registration and the actual between-image geometric distortion).
Applications: Image mosaicing, creating super-resolution images, integrating information into geographic information systems (GIS), in medicine, cartography, computer vision, classification, environmental monitoring

Liang et al. (2014) - *Automatic Registration of Multisensor Images Using an Integrated Spatial and Mutual Information (SMI) Metric*
*Summary:*arg4

5.4 Multi-resolution Fusion

5.5 Fusion of Mixed Data Types

Butenuth et al. (2007) - *Integration of heterogeneous geospatial data in a federated database*
Summary:

Guo (2019) - *Latent Variable Algorithms for Multimodal Learning and Sensor Fusion*
Summary:

Zhang et al. (2019) - *Fusion of Heterogeneous Earth Observation Data for the Classification of Local Climate Zones*
Summary:

5.6 Unsorted

Shen et al. (2016) - *An Integrated Framework for the SpatioTemporalSpectral Fusion of Remote Sensing Images*

Summary:

-

Summary:

6 Data Processing on Graphs

Bronstein et al. (2017) - *Geometric Deep Learning: Going beyond Euclidean data*

Summary:

Nicolicioiu et al. (2019) - *Recurrent Space-time Graph Neural Networks*

Summary:

Wu et al. (2019) - *A Comprehensive Survey on Graph Neural Networks*

Summary:

Zhou et al. (2018) - *Graph Neural Networks: A Review of Methods and Applications*

Summary:

7 Outlier/ Adversarial Detection

8 Army

Hall et al. (2018) - *Probabilistic Object Detection: Definition and Evaluation*

*Summary:*A probabilistic object detection metric (PDQ - Probability-based Detection Quality) was proposed, thus defining the new task of defining probabilistic object detection metrics. The ability of deep CNNs to quantify both *epistemic* and *aleatoric uncertainty* is paramount for deployment safety-critical applications. PDQ aims to measure the accuracy of an image object detector in terms of its label uncertainty and spatial quality. This is achieved through two steps. First, a detector must reliably quantify its *semantic uncertainty* by providing full probability distributions over known classes for each detection. Next, the detectors must quantify spatial uncertainty by reporting *probabilistic bounding boxes*, where the box corners are modeled as normally distributed. A loss function was constructed to consider both label and spatial quality when providing a final detection measure. The primary benefit of this method is that it provides a measure for the level of uncertainty in a detection.

Is it possible to replace the probabilistic metric with a possibilistic one? Could this be more effective at handling outlying cases?

Mahalanobis and McIntosh (2019) - *A comparison of target detection algorithms using DSIAC ATR algorithm development data set*

*Summary:*The authors provided an initial characterization of detection performance on the DSIAC dataset using the *Faster R-CNN* algorithm and *Quadratic Correlation Filter (QCF)*. Performance was evaluated on two datasets, "easy" and "difficult", where the difficulty was determined by number of pixels on target and local contrast. Under difficult conditions, the Faster R-CNN algorithm achieved noteworthy performance, detecting as much as 80% of the targets at a low false alarm rate of 0.01 FA/Square degree. The dataset was limited by a lack of background diversity.

Tanner and Mahalanobis (2019) - *Fundamentals of Target Classification Using Deep Learning*

*Summary:*A shallow CNN was utilized for ATR on the DSIAC MWIR dataset. The goal of the study was to determine the range of optimal thresholds which would optimally separate the target and clutter class distributions defined by the CNN predictions (output of softmax), as well as determine an upper bound on the number of training images required for optimizing performance. The shallow CNN (5 layers) and a Difference of Gaussians (DoG), which finds regions of high intensity on dark backgrounds were used to detect and classify targets. The CNN could correctly classify 96% of targets as targets and as few as 4% of clutter as targets. It was found that the DoG detector failed when the targets were small (long range) or if the overall image was bright (infrared taken during the daytime). It was also determined that guessing the bright pixels were at the center of the targets was a bad assumption. (The brightest part of a target is not necessarily at its center.)

Li2 - *Collaborative sparse priors for multi-view ATR*
Summary:

Kokiopoulou and Frossard (2010) - *Graph-based classification of multiple observation sets*
Summary:

9 Segmentation

Caselles et al. (1997) - *Geodesic Active Contours*
Summary:

Álvarez et al. (2010) - *Morphological Snakes*
*Summary:*The authors introduce a morphological approach to curve evolution. Snakes or curves iteratively solve partial differential equations (PDEs). By doing so, the shape of the snake deforms to minimize the internal and external energies along its boundary. The internal component keeps the curve smooth, while the external component attaches the curve to image structures such as edges, lines, etc. Curve evolution is one of the most widely used image segmentation/ object tracking algorithms. The main contribution of the paper is a new morphological approach to the solution of the PDE associated with snake model evolution. They approach the solution using only inf-sup operators which has the main benefit of providing simpler level sets (0 outside the contours and 1 inside).

Márquez-Neila et al. (2014) - *A Morphological Approach to Curvature-Based Evolution of Curves and Surfaces*
Summary:

References

L. Álvarez, L. Baumela, P. Henríquez, and P. Márquez-Neila. Morphological snakes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2197–2202, June 2010. doi: 10.1109/CVPR.2010.5539900.

D. T. Anderson, M. A. Islam, R. King, N. H. Younan, J. R. Fairley, S. Howington, F. Petry, P. Elmore, and A. Zare. Binary fuzzy measures and choquet integration for multi-source fusion. In *2017 International Conference on Military Technologies (ICMT)*, pages 676–681, May 2017. doi: 10.1109/MILTECHS.2017.7988843.

O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 581–588 vol. 1, June 2005. doi: 10.1109/CVPR.2005.151.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

M. Belkin and P. Niyogi. Semi-supervised learning on riemannian manifolds. *Machine Learning*, 56(1): 209–239, Jul 2004. ISSN 1573-0565. doi: 10.1023/B:MACH.0000033120.25363.1e.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, December 2006. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1248547.1248632`.

Y. Bengio, A. C. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`.

E. Bengoetxea. *Inexact Graph Matching Using Estimation of Distribution Algorithms*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, Dec. 2002.

J. Beyer, K. Heesche, W. Hauptmann, and C. Otte. Heterogeneous mixture-of-experts for fusion of locally valid knowledge-based submodels. 01 2009.

C. Bishop, M. Svensn, and C. K. I. Williams. Gtm: The generative topographic mapping. 10:215–234, January 1998. URL `https://www.microsoft.com/en-us/research/publication/gtm-the-generative-topographic-mapping/`.

J. Bocinsky. Learning multiple target concepts from uncertain, ambiguous data using the adaptive cosine estimator and spectral match filter. Master's thesis, Univ. of Florida, Gainesville, FL, May 2019.

J. Bocinsky, C. H. McCurley, D. Shats, and A. Zare. Investigation of initialization strategies for the multiple instance adaptive cosine estimator. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV, 110120N*, volume 11012 of *Proc.SPIE*, May 2019. doi: 10.1117/12.2519463.

D. Bouzas, N. Arvanitopoulos, and A. Tefas. Graph embedded nonparametric mutual information for supervised dimensionality reduction. *IEEE Transactions on Neural Networks and Learning Systems*, 26 (5):951–963, May 2015. ISSN 2162-237X. doi: 10.1109/TNNLS.2014.2329240.

G. Brigot, E. Colin-Koeniguer, A. Plyer, and F. Janez. Adaptation and evaluation of an optical flow method applied to coregistration of forest remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(7):2923–2939, July 2016. ISSN 1939-1404. doi: 10.1109/JSTARS.2016.2578362.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, July 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2693418.

M. Butenuth, G. v. Gsseln, M. Tiedge, C. Heipke, U. Lipeck, and M. Sester. Integration of heterogeneous geospatial data in a federated database. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62: 328–346, 10 2007. doi: 10.1016/j.isprsjprs.2007.04.003.

L. Cao, F. Luo, L. Chen, S. Yihan, H. Wang, C. Wang, and R. Ji. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recognition*, 64, 12 2016. doi: 10.1016/j.patcog.2016.10.033.

M. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *CoRR*, abs/1612.03365, 2016. URL `http://arxiv.org/abs/1612.03365`.

V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, Feb 1997. ISSN 1573-1405. doi: 10.1023/A:1007979827043. URL `https://doi.org/10.1023/A:1007979827043`.

G. Chao, Y. Luo, and W. Ding. Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358, 2019. ISSN 2504-4990. doi: 10.3390/make1010020.

M. Chen, J. Wang, X. Li, and X. Sun. Robust semi-supervised manifold learning algorithm for classification. *Mathematical Problems in Engineering*, 2018:1–8, 02 2018. doi: 10.1155/2018/2382803.

V. Cheplygina, D. M. J. Tax, and M. Loog. Dissimilarity-based ensembles for multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1379–1391, June 2016. ISSN 2162-237X. doi: 10.1109/TNNLS.2015.2424254.

V. Cheplygina, M. Bruijne, and J. P. W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280 – 296, 2019. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2019.03.009.

J. Chiang and P. D. Gader. Hybrid fuzzy-neural systems in handwritten word recognition. *IEEE Transactions on Fuzzy Systems*, 5(4):497–510, Nov 1997. ISSN 1063-6706. doi: 10.1109/91.649901.

J. Choo, S. Bohn, G. Nakamura, A. M. White, and H. Park. Heterogeneous data fusion via space alignment using nonmetric multidimensional scaling. In *SDM*, 2012.

M. Cook. Task driven extended functions of multiple instances (td-efumi). Master's thesis, Univ. of Missouri, Columbia, MO, 2015.

M. Cook, A. Zare, and D. K. C. Ho. Buried object detection using handheld wemi with task-driven extended functions of multiple instances. In *Proc. SPIE 9823, Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXI, 98230A*, volume 9823 of *Proc. SPIE*, pages 9823 – 9823 – 9, Apr. 2016. doi: 10.1117/12.2223349.

Z. Cui, H. Chang, S. Shan, and X. Chen. Generalized unsupervised manifold alignment. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2429–2437. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5620-generalized-unsupervised-manifold-alignment.pdf`.

B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Hidden talents of the variational autoencoder, 2017.

Andreas Damianou, Neil D. Lawrence, and Carl Henrik Ek. Manifold alignment determination: finding correspondences across different data views. 01 2017.

M. A. Davenport, C. Hegde, M. F. Duarte, and R. G. Baraniuk. Joint manifolds for data fusion. *IEEE Transactions on Image Processing*, 19(10):2580–2594, Oct 2010. ISSN 1057-7149. doi: 10.1109/TIP.2010.2052821.

S. Dawn, V. Saxena, and B. Sharma. Remote sensing image registration techniques: A survey. In A. Elmoataz, O. Lezoray, F. Nouboud, D. Mammass, and J. Meunier, editors, *Image and Signal Processing*, pages 103–112, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-13681-8.

J. Delaporte, B. M. Herbst, W. Hereman, and S. Van der Walt. An introduction to diffusion maps. 2008.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Prez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31 – 71, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(96)00034-3.

X. Du. *Multiple Instance Choquet Integral For MultiResolution Sensor Fusion.* PhD thesis, Univ. of Missouri, Columbia, MO, Dec. 2017.

X. Du and A. Zare. Multi-resolution multi-modal sensor fusion for remote sensing data with label uncertainty. *CoRR*, abs/1805.00930, 2018. URL `http://arxiv.org/abs/1805.00930`.

X. Du and A. Zare. Multiple instance choquet integral classifier fusion and regression for remote sensing applications. *IEEE Transactions on Geoscience and Remote Sensing*, 57(5):2741–2753, May 2019. ISSN 0196-2892. doi: 10.1109/TGRS.2018.2876687.

X. Du, A. Zare, J. M. Keller, and D. T. Anderson. Multiple instance choquet integral for classifier fusion. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1054–1061, July 2016. doi: 10.1109/CEC.2016.7743905.

H. Frigui and P. Gader. Detection and discrimination of land mines in ground-penetrating radar based on edge histogram descriptors and a possibilistic k-nearest neighbor classifier. *IEEE Transactions on Fuzzy Systems*, 17(1), Feb 2009. ISSN 1063-6706. doi: 10.1109/TFUZZ.2008.2005249.

H. Frigui, L. Zhang, and P. D. Gader. Context-dependent multisensor fusion and its application to land mine detection. *IEEE Transactions on Geoscience and Remote Sensing*, 48(6):2528–2543, June 2010. ISSN 0196-2892. doi: 10.1109/TGRS.2009.2039936.

B. Fritzke. A growing neural gas network learns topologies. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 625–632, Cambridge, MA, USA, 1994. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2998687.2998765`.

P. Gader, A. Mendez-Vasquez, K. Chamberlin, J. Bolton, and A. Zare. Multi-sensor and algorithm fusion with the choquet integral: applications to landmine detection. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages 1605–1608 vol.3, Sep. 2004. doi: 10.1109/IGARSS.2004.1370635.

P. D. Gader, M. A. Mohamed, and J. M. Keller. Fusion of handwritten word classifiers. *Pattern Recognition Letters*, 17(6):577 – 584, 1996. ISSN 0167-8655.

M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis. C-WSL: count-guided weakly supervised localization. *CoRR*, abs/1711.05282, 2017. URL `http://arxiv.org/abs/1711.05282`.

X. Gao, X. Wang, D. Tao, and X. Li. Supervised gaussian process latent variable model for dimensionality reduction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):425–434, April 2011. ISSN 1083-4419. doi: 10.1109/TSMCB.2010.2057422.

M. Gnen. Bayesian supervised dimensionality reduction. *IEEE Transactions on Cybernetics*, 43(6):2179–2189, Dec 2013. ISSN 2168-2267. doi: 10.1109/TCYB.2013.2245321.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

A. N. Gorban and A. Y. Zinovyev. Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization. In A. N. Gorban, B. Kégl, D. C. Wunsch, and A. Y. Zinovyev, editors, *Principal Manifolds for Data Visualization and Dimension Reduction*, pages 96–130, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-73750-6.

A. H. Gunatilaka and B. A. Baertlein. Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 577–589, June 2001. ISSN 0162-8828. doi: 10.1109/34.927459.

L. Guo. Latent variable algorithms for multimodal learning and sensor fusion. *CoRR*, abs/1904.10450, 2019. URL `http://arxiv.org/abs/1904.10450`.

J. K. Hackett and M. Shah. Multi-sensor fusion: a perspective. In *Proceedings., IEEE International Conference on Robotics and Automation*, pages 1324–1330 vol.2, May 1990. doi: 10.1109/ROBOT.1990.126184.

H. Hajimirsadeghi and G. Mori. Multi-instance classification by max-margin training of cardinality-based markov networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1839–1852, Sep. 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2613865.

D. Hall, F. Dayoub, J. Skinner, P. Corke, G. Carneiro, and N. Sünderhauf. Probability-based detection quality (PDQ): A probabilistic approach to detection evaluation. *CoRR*, abs/1811.10800, 2018. URL `http://arxiv.org/abs/1811.10800`.

S. S. Haykin. *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition, 2009.

C. Hertzberg, R. Wagner, U. Frese, and L. Schrder. Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds. *Information Fusion*, 14(1):57 – 77, 2013. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2011.08.003. URL `http://www.sciencedirect.com/science/article/pii/S1566253511000571`.

D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu. Cospace: Common subspace learning from hyperspectral-multispectral correspondences. *CoRR*, abs/1812.11501, 2018. URL `http://arxiv.org/abs/1812.11501`.

D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu. Learnable manifold alignment (lema): A semi-supervised cross-modality learning framework for land cover and land use classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:193 – 205, 2019. ISSN 0924-2716.

M. Ilse, Jakub M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018. URL `http://arxiv.org/abs/1802.04712`.

C. Jiao. *Target Concept Learning From Ambiguously Labeled Data*. PhD thesis, Univ. of Missouri, Columbia, MO, Dec. 2017.

C. Jiao and A. Zare. Multiple instance hybrid estimator for learning target signatures. In *2017 IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, pages 988–991, July 2017. doi: 10.1109/IGARSS.2017.8127120.

Changzhe Jiao, Chao Chen, Ronald G. McGarvey, Stephanie Bohlman, Licheng Jiao, and Alina Zare. Multiple instance hybrid estimator for hyperspectral target characterization and sub-pixel target detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:235 – 250, 2018. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2018.08.012.

P. Jindal and D. Kumar. A review on dimensionality reduction techniques. *International Journal of Computer Applications*, 173, 09 2017.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pages 1339–1344 vol.2, Oct 1993.

Z. Kang, H. Pan, S. C. H. Hoi, and Z. Xu. Robust graph learning from noisy data. *CoRR*, abs/1812.06673, 2018. URL `http://arxiv.org/abs/1812.06673`.

A. Karem and H. Frigui. Multiple instance learning with multiple positive and negative target concepts. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 474–479, Dec 2016. doi: 10.1109/ICPR.2016.7899679.

J. M. Keller, S. Auephanwiriyakul, and P. D. Gader. Experiments inf predictive sensor fusion. In *Detection and Remediation Technologies for Mines and Minelike Targets VI*, volume 4394 of *Proc.SPIE*, pages 1047 – 1058, Oct. 2001. doi: 10.1117/12.445433.

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, Sep. 1990. ISSN 0018-9219. doi: 10.1109/5.58325.

E. Kokiopoulou and P. Frossard. Graph-based classification of multiple observation sets. *Pattern Recognition*, 43(12):3988 – 3997, 2010. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.07.016.

E. Kokiopoulou and Y. Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12): 2143–2156, Dec 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1131.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, dec 2005. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1046920.1194904`.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 329–336, Cambridge, MA, USA, 2003. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2981345.2981387`.

C. Lee, A. Elgammal, and M. Torki. Learning representations from multiple manifolds. *Pattern Recogn.*, 50(C):74–87, February 2016. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.08.024. URL `http://dx.doi.org/10.1016/j.patcog.2015.08.024`.

B. Li, C. Yuan, W. Xiong, W. Hu, H. Peng, X. Ding, and S. Maybank. Multi-view multi-instance learning based on joint sparse representation and multi-view dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2554–2560, Dec 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2669303.

D. Li, L. Zhu, W. Bao, F. Cheng, Y. Ren, and D. Huang. Cross-validated smooth multi-instance learning. pages 1321–1325, 05 2017. doi: 10.1109/IJCNN.2017.7966005.

H. Li, D. Liu, and D. Wang. Approximate policy iteration with unsupervised feature learning based on manifold regularization. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, July 2015.

Z. Li, W. Shi, X. Shi, and Z. Zhong. A supervised manifold learning method. *Comput. Sci. Inf. Syst.*, 6: 205–215, 12 2009.

J. Liang, X. Liu, K. Huang, X. Li, D. Wang, and X. Wang. Automatic registration of multisensor images using an integrated spatial and mutual information (smi) metric. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):603–615, Jan 2014. ISSN 0196-2892. doi: 10.1109/TGRS.2013.2242895.

D. Liao, Y. Qian, J. Zhou, and Y. Y. Tang. A manifold alignment approach for hyperspectral image visualization with natural color. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3151–3162, June 2016. ISSN 0196-2892. doi: 10.1109/TGRS.2015.2512659.

Y. Liu, Y. Liu, K. C. C. Chan, and K. A. Hua. Hybrid manifold embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 25(12):2295–2302, Dec 2014. ISSN 2162-237X. doi: 10.1109/TNNLS.2014.2305760.

L. Livi and A. Rizzi. The graph matching problem. *Pattern Anal. Appl.*, 16(3):253–283, Aug 2013. ISSN 1433-7541. doi: 10.1007/s10044-012-0284-8.

L.Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm. A symmetric encoder-decoder with residual block for infrared and visible image fusion. *ArXiv*, abs/1905.11447, 2019.

E. Lopez-Rubio and E. J. Palomo. Growing hierarchical probabilistic self-organizing graphs. *IEEE Transactions on Neural Networks*, 22(7):997–1008, July 2011. ISSN 1045-9227. doi: 10.1109/TNN.2011.2138159.

A. Mahalanobis and B. McIntosh. A comparison of target detection algorithms using dsiac atr algorithm development data set. Proc.SPIE, Apr. 2019.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97, pages 570–576, Cambridge, MA, USA, 1998. MIT Press. ISBN 0-262-10076-2.

O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 341–349, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

P. Márquez-Neila, L. Baumela, and L. Álvarez. A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):2–17, Jan 2014. ISSN 0162-8828.

C. H. McCurley, J. Bocinsky, and A. Zare. Comparison of hand-held wemi target detection algorithms. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV, 110120U*, volume 11012 of *Proc.SPIE*, May 2019. doi: 10.1117/12.2519454.

M. Meng and X. Zhan. Zero-shot learning via low-rank-representation based manifold regularization. *IEEE Signal Processing Letters*, 25(9):1379–1383, Sep. 2018. ISSN 1070-9908. doi: 10.1109/LSP.2018.2857201.

M. Mohandes, M. Deriche, and S. O. Aliyu. Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2813079.

R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007. doi: 10.1109/ICCV.2007.4408976.

M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., 2017.

A. L. Nicolicioiu, I. Duta, and M. Leordeanu. Recurrent space-time graph neural networks. *CoRR*, abs/1904.05582, 2019. URL `http://arxiv.org/abs/1904.05582`.

E. J. Palomo and E. Lopez-Rubio. Learning topologies with the growing neural forest. *International Journal of Neural Systems*, 26(04):1650019, 2016. doi: 10.1142/S0129065716500192. URL `https://doi.org/10.1142/S0129065716500192`. PMID: 27121995.

E. J. Palomo and E. Lopez-Rubio. The growing hierarchical neural gas self-organizing neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9):2000–2009, Sep. 2017. ISSN 2162-237X. doi: 10.1109/TNNLS.2016.2570124.

B. Raducanu and F. Dornaika. A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognition*, 45(6):2432 – 2444, 2012. ISSN 0031-3203.

F. Ratle, G. Camps-Valls, and J. Weston. Semisupervised neural networks for efficient hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(5):2271–2282, May 2010. ISSN 0196-2892. doi: 10.1109/TGRS.2009.2037898.

A. Rauber, D. Merkl, and M. Dittenbach. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13(6):1331–1341, Nov 2002. ISSN 1045-9227. doi: 10.1109/TNN.2002.804221.

B. Ren, B. Hou, J. Zhao, and L. Jiao. Unsupervised classification of polarimetirc sar image via improved manifold regularized low-rank representation with multiple features. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2):580–595, Feb 2017. ISSN 1939-1404. doi: 10.1109/JSTARS.2016.2573380.

I. Rish, G. Grabarnik, G. A. Cecchi, F. Pereira, and G. J. Gordon. Closed-form supervised dimensionality reduction with generalized linear models. pages 832–839, 01 2008. doi: 10.1145/1390156.1390261.

R. Rojas. Associative networks. In *Neural Networks - A Systematic Introduction*, chapter 12, pages 311–336. Springer-Verlag, Berlin, New-York, 1st edition, 1996.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2323. URL `https://science.sciencemag.org/content/290/5500/2323`.

A. Ruiz, O. Rudovic, X. Binefa, and M. Pantic. Multi-instance dynamic ordinal random fields for weakly supervised facial behavior analysis. *IEEE Transactions on Image Processing*, 27(8):3969–3982, Aug 2018. ISSN 1057-7149. doi: 10.1109/TIP.2018.2830189.

D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9(1):75–112, 1985. doi: 10.1207/s15516709cog0901_5. URL `https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0901_5`.

D. Ruta and B. Gabrys. An overview of classifier fusion methods. *Computing and Information Systems*, 7: 1–10, 01 2000.

John E. Ball Alina Zare Brendan Alvey Ryan E. Smith, Derek T. Anderson. Aggregation of choquet integrals in gpr and emi for handheld platform-based explosive hazard detection. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXII, 1018217*, volume 10182, May 2017. doi: 10.1117/12.2263005. URL `https://doi.org/10.1117/12.2263005`.

L. K. Saul and S. T. Roweis. An introduction to locally linear embedding. *Journal of Machine Learning Research*, 7, 01 2001.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017. URL `http://arxiv.org/abs/1701.06538`.

D. Shen, P. Zulch, M. Disasio, E. Blasch, G. Chen, Z. Wang, J. Lu, and R. Niu. Manifold learning algorithms for sensor fusion of image and radio-frequency data. In *2018 IEEE Aerospace Conference*, pages 1–9, March 2018. doi: 10.1109/AERO.2018.8396395.

H. Shen, X. Meng, and L. Zhang. An integrated framework for the spatiotemporalspectral fusion of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7135–7148, Dec 2016. ISSN 0196-2892. doi: 10.1109/TGRS.2016.2596290.

R. E. Smith, D. T. Anderson, A. Zare, J. E. Ball, B. Smock, J. R. Fairley, and S. E. Howington. Genetic programming based choquet integral for multi-source fusion. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8, July 2017. doi: 10.1109/FUZZ-IEEE.2017.8015481.

J. S. Stanley III, G. Wolf, and S. Krishnaswamy. Manifold alignment with feature correspondence. *CoRR*, abs/1810.00386, 2018. URL `http://arxiv.org/abs/1810.00386`.

M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 905–912, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143958.

Q. Sun, H. Liu, and T. Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187 – 201, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2016.09.016. URL `http://www.sciencedirect.com/science/article/pii/S0031320316302771`.

R. Talmon, S. Mallat, H. Zaveri, and R. R. Coifman. Manifold learning for latent variable inference in dynamical systems. *IEEE Transactions on Signal Processing*, 63(15):3843–3856, Aug 2015. ISSN 1053-587X. doi: 10.1109/TSP.2015.2432731.

I. L. Tanner and A. Mahalanobis. Fundamentals of target classification using deep learning. Proc.SPIE, Apr. 2019.

J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2319. URL `https://science.sciencemag.org/content/290/5500/2319`.

S. Theodoridis and K. Koutroumbas. Kernel pca. In *Pattern Recognition, Fourth Edition*, chapter 6, pages 351–353. Academic Press, Inc., Orlando, FL, USA, 4th edition, 2008a. ISBN 1597492728, 9781597492720.

S. Theodoridis and K. Koutroumbas. The karhunen-loeve transform. In *Pattern Recognition, Fourth Edition*, chapter 6, pages 326–334. Academic Press, Inc., Orlando, FL, USA, 4th edition, 2008b. ISBN 1597492728, 9781597492720.

M. E. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21(3):611–622, January 1999. URL `https://www.microsoft.com/en-us/research/publication/probabilistic-principal-component-analysis/`.

M. Trabelsi and H. Frigui. Fuzzy and possibilistic clustering for multiple instance linear regression. pages 1–7, 07 2018. doi: 10.1109/FUZZ-IEEE.2018.8491540.

I. W. Tsang and J. T. Kwok. Large-scale sparsified manifold regularization. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1401–1408. MIT Press, 2007. URL `http://papers.nips.cc/paper/3005-large-scale-sparsified-manifold-regularization.pdf`.

D. Tuia and G. Camps-Valls. Kernel manifold alignment. 04 2015.

S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann. *Review of Classifier Combination Methods*, pages 361–386. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-76280-5. doi: 10.1007/978-3-540-76280-5_14. URL `https://doi.org/10.1007/978-3-540-76280-5_14`.

L. van der Maaten, E. Postma, and H. Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research - JMLR*, 10, 01 2007.

Praneeth Vepakomma, C. Tonde, and A. M. Elgammal. Supervised dimensionality reduction via distance correlation maximization. *CoRR*, abs/1601.00236, 2016. URL `http://arxiv.org/abs/1601.00236`.

G. Vivar, H. Burwinkel, A. Kazi, A Zwergal, N. Navab, and S. Ahmadi. Multi-modal graph fusion for inductive disease classification in incomplete datasets. *CoRR*, abs/1905.03053, 2019. URL `http://arxiv.org/abs/1905.03053`.

E. Vural and C. Guillemot. Out-of-sample generalizations for supervised manifold learning for classification. *IEEE Transactions on Image Processing*, 25(3):1410–1424, March 2016. ISSN 1057-7149. doi: 10.1109/TIP.2016.2520368.

E. Vural and C. Guillemot. A study of the classification of low-dimensional data with supervised manifold learning. *CoRR*, abs/1507.05880, 2018. URL `http://arxiv.org/abs/1507.05880`.

C. Wang and S. Mahadevan. Multiscale manifold alignment. 09 2010.

C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1541–1546. AAAI Press, 2011. ISBN 978-1-57735-514-4. doi: 10.5591/978-1-57735-516-8/IJCAI11-259. URL `http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-259`.

C. Wang, P. Krafft, and S. Mahadevan. Manifold alignment. 12 2011. doi: 10.1201/b11431-6.

R. Wang, S. Shan, X. Chen, Q. Dai, and W. Gao. Manifoldmanifold distance and its application to face recognition with image sets. *IEEE Transactions on Image Processing*, 21(10):4466–4479, Oct 2012. ISSN 1057-7149. doi: 10.1109/TIP.2012.2206039.

R. Wang, X. Wang, S. Kwong, and C. Xu. Incorporating diversity and informativeness in multiple-instance active learning. *IEEE Transactions on Fuzzy Systems*, 25(6):1460–1475, Dec 2017. ISSN 1063-6706. doi: 10.1109/TFUZZ.2017.2717803.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596, 2019. URL `http://arxiv.org/abs/1901.00596`.

X. Geng, D. Zhan, and Z. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(6):1098–1107, Dec 2005. ISSN 1083-4419.

Y. Xiao, B. Liu, and Z. Hao. Multiple-instance ordinal regression. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–16, 11 2017. doi: 10.1109/TNNLS.2017.2766164.

Y. Xiao, B. Liu, and Z. Hao. A sphere-description-based approach for multiple-instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):242–257, Feb 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2539952.

H. Xu, L. Yu, M. A. Davenport, and H. Zha. Active manifold learning via a unified framework for manifold landmarking. *CoRR*, abs/1710.09334, 2017. URL `http://arxiv.org/abs/1710.09334`.

H. L. Yang and M. M. Crawford. Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):51–64, Jan 2016. ISSN 0196-2892. doi: 10.1109/TGRS.2015.2449736.

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, Aug 2012. ISSN 2162-237X. doi: 10.1109/TNNLS.2012.2200299.

Z. Zhang, H. Zha, and M. Zhang. Spectral methods for semi-supervised manifold learning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2008. doi: 10.1109/CVPR.2008.4587381.

A. Zare and C. Jiao. Extended functions of multiple instances for target characterization. In *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4, June 2014. doi: 10.1109/WHISPERS.2014.8077525.

A. Zare, M. Cook, B. Alvey, and D. K. Ho. Multiple instance dictionary learning for subsurface object detection using handheld emi. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XX, 94540G*, Proc. SPIE, May 2015. doi: 10.1117/12.2179177. URL `https://doi.org/10.1117/12.2179177`.

A. Zare, C. Jiao, and T. C. Glenn. Multiple instance hyperspectral target characterization. *CoRR*, abs/1606.06354, 2016. URL `http://arxiv.org/abs/1606.06354`.

G. Zhang, P. Ghamisi, and X. Zhu. Fusion of heterogeneous earth observation data for the classification of local climate zones. *ArXiv*, abs/1905.12305, 2019.

J. Zhang. Multi-source remote sensing data fusion: Status and trends. *International Journal of Image and Data Fusion*, 1:5–24, 03 2010. doi: 10.1080/19479830903561035.

Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1073–1080. MIT Press, 2002.

Y. Zhang, X. Zheng, G. Liu, X. Sun, H. Wang, and K. Fu. Semi-supervised manifold learning based multigraph fusion for high-resolution remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters*, 11(2):464–468, Feb 2014. ISSN 1545-598X. doi: 10.1109/LGRS.2013.2267091.

J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun. Graph neural networks: A review of methods and applications. *CoRR*, abs/1812.08434, 2018. URL `http://arxiv.org/abs/1812.08434`.

B. Zitov and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977 – 1000, 2003. ISSN 0262-8856. doi: https://doi.org/10.1016/S0262-8856(03)00137-9.