ABSTRACT

GHANEM, SALLY SAMIR. Information Fusion: Scaling Subspace-Driven Approaches. (Under the direction of Dr. Hamid Krim).

Union of Subspaces (UoS) is a new paradigm for signal modeling and processing, which can identify more complex trends in data sets than simple linear models. Relying on a bi-sparsity pursuit framework and advanced non-smooth optimization techniques, the Robust Subspace Recovery (RoSuRe) algorithm was introduced in the recent literature as a reliable and numerically efficient strategy to unfold unions of subspaces.

The objective of this work is to exploit this inherent multiple subspace that can be imputed to data, to develop a multi-modal approach to object classification and identification using an ensemble of sensors. We consider a more realistic unsupervised learning scenario, where no training dataset is provided and adopt a data driven approach to determine key features extracted from each sensor modality. We subsequently combine the features from each sensor modality to generate a desirable universal feature thereby enabling an improved classification rate.

The UoS structure is unveiled by pursuing sparse self-representation of the given data, to yield the recovery of the underlying subspace structure in each sensor modality and obtain a finer level of classification by combining them. We subsequently use the resulting UoS structure to classify new observed data points and demonstrates the generalization power of our technique. In addition, for scaling capability purposes, we seek to exploit the deep structure of multi-modal data to robustly exploit the group subspace distribution of the information using the Convolutional Neural Network (CNN) formalism. Referring to it as to as deep Multimodal Robust Group Subspace Clustering (DRoGSuRe), this approach is compared against the independently developed state-of-the-art

approach named Deep Multimodal Subspace Clustering (DMSC). Experiments on different multimodal datasets show that our approach is competitive and more robust in the presence of noise. To further improve the complexity of this approach, we propose a deep structure encoder using the recently introduced Volterra Neural Networks (VNNs) to seek an efficient latent representation of multi-modal data whose features are jointly captured by a union of subspaces. The so-called self-representation embedding of the latent codes leads to a simplified fusion which is driven by a similarly constructed decoding. The Volterra Filter architecture achieved reduction in parameter complexity is primarily due to controlled non-linearities being introduced by the higher order convolutions in contrast to generalized activation functions.

Experimental results on two different datasets have shown a significant improvement in the clustering performance for VNN auto-encoder over conventional Convolutional Neural Network (CNN) auto-encoder. In addition, we also show that the proposed approach demonstrates a much-improved sample complexity over CNN-based auto-encoder with a superb robust classification performance.

Information Fusion: Scaling Subspace-Driven Approaches

by
Sally Samir Ghanem

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina
2021

APPROVED BY:

_____	_____
Hamid Krim Committee Chair	Brian Hughes
_____	_____
Huaiyu Dai	Kazufumi Ito

DEDICATION

To my everything, my son, Sajed, and my parents, Lila and Samir.

BIOGRAPHY

Sally Ghanem received her Bachelor of Science degree in Electrical Engineering from Alexandria University, Alexandria, Egypt, in 2013. She received her Master of Science degree in Electrical and Computer engineering from North Carolina State University, Raleigh, NC, USA in 2016, where she is currently pursuing the Ph.D. degree. In 2015, she joined the Vision, Information and Statistical Signal Theories and Applications (VISSTA) group. Her research interests include the areas of Computer Vision, Signal Processing, Image processing, Data Fusion, and Machine learning.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Hamid Krim for the constant support in my PhD study and research. Without his support this would have never been possible. I express my gratitude to the rest of my committee: Dr. Brian Hughes, Dr. Huaiyu Dai, and Dr. Kazufumi Ito for all your suggestions, comments, and questions.

My sincere thanks go to Dr. Hamilton Scott Clouse for his guidance, motivation, enthusiasm, and immense knowledge during my internship at the Air Force Research Laboratory (AFRL). I also acknowledge Dr. Ryan A. Kerekes for his assistance, expertise, and guidance during my internship at Oak Ridge National Laboratory (ORNL). My sincere thanks also go to Dr. Ashkan Panahi, and Dr. Siddharth Roheda, the post-doctoral researchers in the Vision, Information and Statistical Signal Theories and Applications (VISSTA) group for their guidance during my PhD. I would like to thank my colleagues: Erik Skau, Jennifer Gamble, Saba Emrani, Yuming Huang, Shahin Mahdizadehaghdam, Kenneth Tran, Wen Tang, Bo Jiang, and Tanmay Asthana for the valuable discussions and assistance during my time at NC State.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Robust Subspace Recovery (RoSuRe) algorithm was recently introduced as a principled and numerically efficient algorithm that unfolds underlying Unions of Subspaces (UoS) structure, present in the data. In contrast to simple linear models, the union of Subspaces (UoS) is capable of identifying more complex trends in data. We build on and extend RoSuRe to prospect the structure of different data modalities. We propose a novel multi-modal data fusion approach which learns a new joint representation of data modalities, in congruence with the underlying UoS model. We subsequently integrate the obtained structures to form a unified subspace structure. The proposed approach exploits the structural dependencies between the different modalities data to cluster the associated target objects. The resulting fusion of the unlabeled sensors' data has shown that our method is competitive with other state of the art subspace clustering methods. The resulting UoS structure is employed to classify newly observed data points, highlighting the abstraction capacity of the proposed method.

1.1 Review of State of Art

Unsupervised learning is a very challenging topic in machine learning which involves discovering hidden patterns in data with no given labels. Reliable clustering techniques will save time and effort typically required for labeling large datasets that might have thousands of observations. Subspace clustering was introduced as an efficient way of unfolding a union of low-dimensional subspaces underlying high dimensional data. Subspace clustering has been extensively studied in computer vision on account of the importance of interference applications as well as the

availability of vast visual data [1], [2], [3], and [4]. Union of Subspaces (UoS) is a relatively recent algorithmic approach[1] for identifying complex trends in datasets relative to simple linear models, like Robust Principal Component Analysis [5]. High dimensional data is rich with common and related features lying in corresponding subspaces, and also other non-conforming structures which may be errors or outlier sparse structures. Sparse modeling has been broadly employed in the machine learning literature to model noisy data [6]. Subspace clustering can be applied to group data points picked from a union of low-dimensional subspaces in an unsupervised fashion. Subspace clustering has become a highly researched topic in computer vision on account of the diversity of problems as well as the ready availability of large amounts of visual data, which have in turn, further motivated the development of such representations [7], [8]. Other applications of subspace clustering include image segmentation [9], image compression [10] and motion segmentation [11]. Union of Subspaces has been also incorporated into analysis dictionary learning to boost the classification performance in supervised learning [12]. In this work, we build on the work in [1], in which a bi-sparse model, known as Robust Subspace Recovery (RoSuRe) via Bi-sparsity Pursuit, is employed as a framework to recover the union of subspaces in the presence of sparse perturbations. The UoS structure is unveiled by pursuing sparse self-representation of the given data. We explore the flexibility of the UoS methodology for a new application: Identification and tracking of vehicles by way of multimodal data fusion. This application is similar to the emerging challenge of sensor networks and their role in advanced surveillance technology, as sensor networks can also use multimodal data fusion to obtain more information and insight from diverse data sources [13]. Multimodal data have increasingly become more accessible with the

[1]The idea of a Union of Subspaces was, to the best of our knowledge, first proposed in Berger [82].

proposed intelligent transportation systems allowing to gather more comprehensive and complementary information, and as a result, require optimized exploitation strategies to take advantage of all potential complementarity. While data fusion is not new, it has recently garnered a significant increase in research interest thereby unveiling the need of further foundational principles, in order to meet demands of the varied applications including image fusion [14], [15], target recognition [16], speaker recognition [17] and handwriting analysis [18].

The objective of this work is to develop a multi-modal sensing-based classification approach which we validate on a number of object classification and identification problems. The sensor network can either include similar or dissimilar sensors. Similar sensor fusion, which is the case when multiple sensors explore the same features, has been studied in [19] and [20]. On the other hand, dissimilar sensor fusion [21] employ different sensors to prospect different but potentially complementary features along various dimensions of the target.

Our framework primarily illustrates feature fusion from heterogeneous data modalities, while the approach is equally applicable to homogenous modalities, as it is principled and sufficiently general. Sensor fusion of heterogeneous data has long been of interest since it can explore different characteristics of a target, and provides information reinforcement for an increased resolvability of objects. This additional information from multiple modalities enriches target characterization and supports decision making. A comprehensive survey of data fusion is provided in [22] and [23]. In this work, we consider an operationally relevant unsupervised learning scenario, where no training data is provided, and we adopt a data driven approach to investigate vehicle signatures utilizing key features extracted from each sensor modality. We subsequently combine the features

from each sensor modality to generate a desirable universal feature, and increase the classification rate of specific vehicle classes. Convolutional Neural Networks (CNNs) have been widely used in deep learning for analyzing visual images in many applications. These applications include but not limited to image processing, segmentation, and segmentation. However, the complexity and price of implementing CNNs can be limiting in many applications. Inspired by Volterra non-linear system theory [23], an efficient Volterra Neural Network (VNN) has recently been proposed [24] to address the CNN over parametrization problem. The goal of this work was to control the non-linearities introduced in the network controlling the degree of the interactions between the delayed input samples of the data. The cascaded implementation proposed in [24] has shown to significantly reduce the number of parameters needed for training the network as compared to conventional neural networks. In addition to reducing the network complexity, Volterra Neural Networks (VNNs) have more tractable and comprehensible structure. The idea of using Volterra filter in neural networks had also been discussed in [25] and [26]. These other proposed approaches have a severely constrained order of non-linearity on account of an explosively increasing number of parameters.

1.2 Contributions

In Chapter 2, we build on the procedure studied in [1], in which a bi-sparse model, known as Robust Subspace Recovery (RoSuRe) via Bi-sparsity Pursuit, is employed as a framework to recover the union of subspaces in the presence of sparse perturbations. The UoS structure is unveiled by pursuing sparse self-representation of the given data. We explore the flexibility of the UoS methodology for a new application: Identification and tracking of vehicles by way of multimodal data fusion. Conventional vehicle identification methods such as license plate recognition and Radio Frequency Identification Tags (RFID) have long been widely used for that

purpose [27]. Unfortunately, such image-based methods are not always appropriate, for example, in applications that require low-power, low-cost, and robustness to environmental changes. Additionally, attention to privacy issues has raised more concerns over image acquisition. In contrast, further alternatives such as magnetic sensors and microphones are inexpensive and obviate privacy concerns [28]. The metallic structure of vehicles is primarily used as it perturbs the earth's magnetic field, and hence produces unique magnetic signatures that have served to discriminate between vehicles [29], [30]. Audio sensors have also been extensively employed in the area of vehicle classification for different applications and have proven their effectiveness and robustness [31], [32]. We employ two approaches for data fusion. The first approach relies on the bi-sparsity framework to recover the underlying subspace structure in each sensor modality separately, and to obtain a finer level of classification by combining them through consolidation [33]. The second approach relies on robust group sparsity, where the data from different modalities are jointly exploited to obtain a unified common subspace structure in one step.

In Chapter 3, we seek to exploit the deep structure of multi-modal data to robustly exploit the group subspace distribution of the information using the Convolutional Neural Network (CNN) formalism. We extract key features from each data modality using a CNN encoder, and subsequently combine those features to generate a common discriminative feature. Upon unfolding the set of subspaces invoking each data modality, and learning their corresponding encoders, an optimized integration of the generated inherent information is carried out to yield a characterization of various classes. Referred to as deep Multimodal Robust Group Subspace Clustering (DRoGSuRe), this approach is compared against the independently developed state-of-the-art approach named Deep Multimodal Subspace Clustering (DMSC) [34]. We test our

approach on two popular datasets, which we divide into learning and validation sets. The learned UoS structure is then utilized to classify new observed data points, which illustrates the generalization power of the proposed approach. By considering different scenarios by way of additive noise to either the training set, or the testing set, or both, we thoroughly investigate the robustness and resilience of the clustering approach performance.

Inspired by the success of Volterra Neural Networks (VNNs) in deep learning [24], we introduce an efficient implementation of the Deep Multi-modal Subspace clustering auto-encoder [34] in Chapter 4. We employ VNNs to seek a latent representation of multi-modal data whose features are jointly captured by a union of subspaces. The so-called self-representation embedding of the latent codes simplified the fusion which was driven by a similarly constructed decoding. More specifically, the CNNs are replaced with VNNs to control the introduced non-linearities via high order convolutions instead of using highly non-linear activation functions. Moreover, we propose additional ways to reduce the number of parameters needed to train the VNNs auto-encoder to a fraction of the number of parameters used by CNNs. Experimental results on two different datasets have shown a significant improvement in the clustering performance for VNNs auto-encoder over conventional Convolutional Neural Network (CNNs) auto-encoder. In addition, we also show that the proposed approach demonstrates a much-improved sample complexity over CNN-based auto-encoder with a superb robust classification performance.

CHAPTER 2

Robust Group Subspace Recovery: New Approach for Multi-Modality Data Fusion

2.1 Introduction

Union of Subspaces (UoS) is a novel approach for identifying complex trends in datasets relative to simple linear models. High dimensional data is rich with common and related features lying in corresponding subspaces and other nonconforming structures which may be errors or outlier sparse structures. Sparse modeling has been extensively utilized in the computer vision and machine learning literature to obtain linear models under the influence of perturbation [35] [36] [37]. In this chapter, we build on the procedure studied in [1], in which a bi-sparse model, known as Robust Subspace Recovery (RoSuRe) via Bi-sparsity Pursuit, is employed as a framework to recover the union of subspaces in the presence of sparse corruptions. The UoS structure is unveiled by pursuing sparse self-representation of the given data.

We explore the flexibility of the UoS methodology for a new application: Identification and tracking of vehicles by way of multimodal data fusion. This application is similar to the emerging challenge of sensor networks and their role in advanced surveillance technology, as sensor networks can also use multimodal data fusion to obtain more information and insight from multiple data sources. Multimodal data have increasingly become more accessible with the proposed intelligent transportation systems to gather more comprehensive and complementary information, which in turn, require optimized exploitation strategies taking advantage of system redundancy.

Recent developments in sensor technology have provided many possibilities in developing real-time transportation systems technologies. Traffic flow optimization, dynamic traffic management solutions, vehicle counting, travel time estimation, and other traffic modeling studies frequently require classification and identification of streams of vehicles. Moreover, accurate estimation of traffic parameters needs to be performed in real time for decision makers [38] [39]. Conventional vehicle identification methods such as license plate recognition and Radio Frequency Identification Tags (RFID) have been widely used for that purpose for so long [40] [41]. Unfortunately, such image-based methods are not appropriate for studies that require low-power consumption and low cost. Additionally, privacy issues becoming front and center have raised the concern over with image acquisition. On the contrary, further alternatives such as magnetic sensors and microphones are inexpensive and do not raise privacy concerns [42] [28]. Vehicles are primarily of metallic structure that perturb the earth's magnetic field, and hence produce unique magnetic signatures that have been served to discriminate between vehicles [29] [30]. Audio sensors have also been extensively employed in the area of vehicle classification for different applications and have proven their effectiveness and robustness [31] [32].

In this Chapter, we consider an operationally relevant unsupervised learning scenario, where no training dataset is provided, and we adopt a data driven approach to determine vehicle signatures utilizing key features extracted from each sensor modality. We subsequently combine the features from each sensor modality to generate a desirable universal feature and increase the classification rate of specific vehicle classes. We employ two approaches for data fusion. The first approach relies on the bi-sparsity framework to recover the underlying subspace structure in each sensor modality separately and obtain a finer level of classification by combining them through addition.

The second approach relies on robust group sparsity, where the data from different modalities are jointly exploited to obtain a unified common subspace structure in one step.

The idea of group sparsity was previously utilized in [43]. The Multi-task Low-rank Affinity Pursuit (MLAP) was proposed to boost region-based image segmentation by fusing different image features which have to have the same dimensionality and cannot be easily modified to address the present issue. On the other hand, our approach can address data modalities that have different dimensionalities and potentially incompatible in nature. MLAP was also based on LRR [44] while our approach is based on RoSuRe [1]. For fairness, we will compare our approach to other popular multimodal subspace clustering baselines such as LTMSC [45], kMLRR [46], kMSSC [46], MSSC [46] and MLRR [46]. LTMSC [46] exploits the complementary information of different views by jointly employing tensors to explore the high order correlations. The other methods were introduced in [46] as multimodal extension for Sparse Subspace Clustering (SSC) [2] and Low-rank Representation-based (LRR) [44] by enforcing a common representation across data modalities.

2.2 Information Subspace-Based Fusion

2.2.1 Problem Formulation

Consider a set of data realizations indexed by $k = 1, 2, \dots, n$. Furthermore, assume T data modalities, indexed by $t = 1, 2, 3, \dots, T$. Each data realization can be represented as a $m-$dimensional vector $\boldsymbol{x_k}(t) \in \mathbb{R}^m$, where $\mathbf{X}$(t) = $[\boldsymbol{x_1}(t)\ \boldsymbol{x_2}(t) \dots \boldsymbol{x_n}(t)]$. The goal is to partition a set of realizations into clusters whose respective measurements for each modality is well-represented by a low-dimensional subspace. Mathematically, this is tantamount to seeking a partitioning $\{\boldsymbol{X}^1, \boldsymbol{X}^2, \dots, \boldsymbol{X}^P\}$ of $[n]$ observations/realizations. Also, P is the number of clusters

indexed by i, such that there exist linear subspaces $S^I(t) \in \mathbb{R}^m$ with dim $(S^I(t)) \ll m$, where $\boldsymbol{x_k(t)} \in S^I(t)$ for all $t \in [T]$.

2.2.2 Robust Subspace Recovery via Bi-Sparsity Pursuit: Fusion of Subspace Information

The Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe) introduced in [1] was originally proposed for unimodal data. We present an overview for the RoSuRe structure, which our modified fusion approach exploits. We subsequently elaborate on the formulation of our problem and describe how we address the multi-modal data fusion. We assume that all data samples may be corrupted by additive sparse errors. Therefore, the UoS structure is often corrupted and each data sample deviates from its original subspace. Specifically, we consider a set of data samples $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}_{m,n}$ where n corresponds to the number of realizations and m specifies the number of variables or features in each realization. The columns of the matrix $\boldsymbol{X}$ may be partitioned such that each part $\boldsymbol{X^I}$ is decomposed into a low dimensional subspace and a sparse outlier (e.g., non-conforming data).

$$\boldsymbol{X^I} = \boldsymbol{L^I} + \boldsymbol{E^I}, \qquad I = 1, \ldots, P \tag{2.1}$$

where each $\boldsymbol{L^I}$ serves as a single low dimensional subspace of the original data, and $\boldsymbol{L} = [\boldsymbol{L^1}|\boldsymbol{L}^2| \ldots |\boldsymbol{L^P}]$ is the desired union of subspaces. Furthermore, the partition recovers the clustering of the original data samples corrupted by the error $\boldsymbol{E} = [\boldsymbol{E^1}|\boldsymbol{E^2}| \ldots |\boldsymbol{E^P}]$. The objective of this approach is to simultaneously retrieve the subspaces and the noiseless samples from the observed noisy data. The RoSuRe via Bi-Sparsity pursuit is based on the idea of self-representation. In other words, $\boldsymbol{l_i}$ can be represented by the other samples from the same subspace $S(\boldsymbol{l_i})$.

$$\boldsymbol{l_i} = \sum_{i \neq j,\, l_j \in S(l_i)} w_{ij} \boldsymbol{l_j} \tag{2.2}$$

The above relation can be represented in a matrix form as follows,

$$\boldsymbol{L} = \boldsymbol{LW} \tag{2.3}$$

Under a suitable arrangement/permutation of the data realizations, the sparse coefficient matrix $\boldsymbol{W}$ is an $n \times n$ block-diagonal matrix with zero diagonals provided that each sample is represented by other samples only from the same subspace. More precisely, $W_{ij} = 0$ whenever the indices i, j correspond to samples from different subspaces. As a result, the majority of the elements in $\boldsymbol{W}$ is equal to zero. After further approximations and relaxations, the problem is formulated as follows,

$$min_{W,E,L} \ \|\boldsymbol{W}\|_1 + \lambda \|\boldsymbol{E}\|_1,$$
$$\text{s.t. } \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{E},\ \boldsymbol{L} = \boldsymbol{LW},\ W_{ii} = 0. \tag{2.4}$$

where $\| \|_1$ denotes the l_1 norm, i.e. the sum of absolute values of the argument. The minimum of Eqn. (2.4) is approximated through linearized Alternating Direction Method of Multipliers ADMM [47] and the sparsity of both $\boldsymbol{E}$ and $\boldsymbol{W}$ is traced until convergence.

2.2.2.1 Finding Clusters in Data Using $\boldsymbol{W}$

The resulting $\boldsymbol{W}$ can be exploited to evaluate an affinity matrix. The affinity matrix is computed by,

$$\boldsymbol{A} = \boldsymbol{W} + \boldsymbol{W}^T \tag{2.5}$$

Subsequently, the spectral clustering method in [48] is utilized for data clustering. The method can be summarized as follows, a matrix $\boldsymbol{D}$ is defined to be a diagonal matrix whose i^{th} diagonal element is the degree of the i^{th} node, i.e., the sum of i^{th} row in $\boldsymbol{A}$. The standard graph Laplacian matrix is next constructed as follows,

$$\boldsymbol{G} = \boldsymbol{D}^{\frac{-1}{2}} \boldsymbol{A}\, \boldsymbol{D}^{\frac{-1}{2}} \tag{2.6}$$

Next, the eigenvectors $\boldsymbol{s_1, s_2, \ldots, s_k}$, of $\boldsymbol{G}$ corresponding to the largest k eigenvalues are computed, where k is the desired number of clusters. The matrix $\boldsymbol{S} = [\boldsymbol{s_1\ s_2 \ldots s_k}]$ is then formed by stacking the eigenvectors in columns. Each row of $\boldsymbol{S}$ is a point in $\mathbb{R}^k$, k-means is then used to cluster the rows of $\boldsymbol{S}$. Finally, the original point $\boldsymbol{x_i}$ is assigned to cluster j iff row i of the matrix $\boldsymbol{S}$ was assigned to cluster j.

2.2.2.2 Multi-Modal Subspace Recovery via RoSuRe

As previously stated, RoSuRe does not support multimodal data since the algorithm needs to be applied on each data modality individually. To overcome this problem, we apply RoSuRe on each data modality and then we integrate the resulting sparse coefficient matrices $\boldsymbol{W}(t)$'s, for $t = 1, 2, \ldots, T$ modalities, through adding them as follows,

$$\boldsymbol{W_{Total}} = \sum_{t=1}^{T} \boldsymbol{W(t)} \tag{2.7}$$

By doing so, we are reinforcing the relation between data points that exist in all data modalities as reflected by the elements of $\boldsymbol{W}(t)$. While this may be justified as ensuring a cross-sensor consistency, we are also reducing the noise variance introduced by the outliers. A similar strategy was explored in community detection in Social Networks [49], where an aggregation of multi-layer adjacency matrices was found to yield a better Signal to Noise ratio, and ultimately improved performance. In multi-layer networks, edges that exist in multiple layers, encode different but related relations among data points. The subspaces/clusters in our case, share model commonalities for given targets for which the relations among data observations are reflected by the sparse (non-zero) elements of $\boldsymbol{W}(t)$. We next introduce an alternative solution by a joint optimization framework over multiple data modalities at the same time, producing one common subspace structure instead of separately doing so with RoSuRe.

2.2.3 Robust Group Subspace Recovery-Driven Fusion

In this subsection, we introduce a novel approach based on RoSuRe which we naturally adapt to multimodal data. We define Ω = $\{\boldsymbol{W}(t)\}_{t=1}^{T}$, where **W**(t) =$[w_{kj}(t)]_{k,j}$. We define the group norm $\|\Omega\|_{1,2}$:

$$\|\Omega\|_{1,2} = \sum\nolimits_{k,j} \sqrt{\sum_{t=1}^{T} w_{k,j}^{2}(t)} \tag{2.8}$$

We introduce the following optimization as the joint Sparse Subspace Clustering (JSSC) framework with group sparsity,

$$min_{\Omega/\, w_{kk}(t)=0} = \|\Omega\|_{1,2} + \rho \sum_{t=1}^{T} \|\boldsymbol{W}(t)\|_{1} \;\; s.t. \;\; \boldsymbol{X}(t) = \boldsymbol{X}(t)\boldsymbol{W}(t) \tag{2.9}$$

This procedure is justified by the observation that when each entity is represented by the other ones in the same cluster, the inter-class terms $\boldsymbol{w_{kj}(t)}$ are zero for every t, which implies that $\|\boldsymbol{\Omega}\|$ is group sparse along the dimension t. Moreover, minimizing group l-norm promotes a group-sparse solution. Equation (2.9) is separable in j, where j is the column index of $\boldsymbol{W}$ and can therefore be minimized for every j and thus re-written as follows,

$$min_{\Omega_j} = \sum_{k=1}^{n} \sqrt{\sum_{t=1}^{T} w_{k,j}^{2}} + \rho \sum_{k,t} \left| w_{k,j}(t) \right| \text{ s.t. } \boldsymbol{x}_j(t) = \boldsymbol{X}(t)\boldsymbol{w}_j(t) \;\forall\; w_{jj}(t) = 0 \tag{2.10}$$

Where $\Omega_j = \{\boldsymbol{w}_j(t)\}_{j=1}^{n}$ and Ω =$\{\Omega_j\}_{j=1}^{n}$. In order to validate our approach, our goal would be to prove that $w_{k,j}$ = 0 $\forall\; k \notin S_\alpha$ and $j \in S_\alpha$, where S_α is the index of the subspace containing $\boldsymbol{x}_j(t)$.

2.2.3.1 Robust Group Subspace Recovery

Similar to RoSuRe, we propose a robust and non-convex version of the above formulation in Eqn. (2.11). We assume that the data matrices include non-conforming elements to assume the structure $\boldsymbol{X}(t) = \boldsymbol{L}(t) + \boldsymbol{E}(t)$, where the columns of $\boldsymbol{L}(t)$ reside on their corresponding subspaces and $\boldsymbol{E}(t)$ is a sparse error matrix. The optimization problem is then rewritten as,

$$\begin{aligned} min_{\Omega/\, w_{kk}(t)=0} &= \|\Omega\|_{1,2} + \rho \sum_{t=1}^{T} \|\boldsymbol{W}(t)\|_1 + \lambda \sum_{t=1}^{T} \|\mathbf{E}(t)\|_1 \;\; s.t. \;\; \boldsymbol{X}(t) \\ &= \boldsymbol{X}(t)\boldsymbol{W}(t) \;\; and \; \boldsymbol{L}(t) = \boldsymbol{L}(t)\boldsymbol{W}(t) \end{aligned} \tag{2.11}$$

which can be approximately solved by a primal-dual method, with an appropriate introduction of an augmented Lagrange form. To proceed, first note that Eqn. (2.11) can be reduced to a two-variable problem by substituting $\boldsymbol{L}(t)$ with $\boldsymbol{X}(t) - \boldsymbol{E}(t)$ and using $\boldsymbol{L}(t) = \boldsymbol{L}(t)\boldsymbol{W}(t)$. Assuming T modalities, the Lagrangian objective functional now becomes:

$$\begin{aligned} & L(\boldsymbol{\Omega}, \boldsymbol{W}(t), \boldsymbol{E}(t), \boldsymbol{Y}(t), \mu) \\ &= \|\boldsymbol{\Omega}\|_{1,2} + \rho \sum_{t=1}^{T} \|\boldsymbol{W}(t)\|_1 + \lambda \sum_{t=1}^{T} \|\mathbf{E}(t)\|_1 \\ &+ \sum_{t=1}^{T} < \boldsymbol{L}(t)\boldsymbol{W}(t) - \boldsymbol{L}(t), \boldsymbol{Y}(t) > + \sum_{t=1}^{T} \frac{\mu}{2} \|\boldsymbol{L}(t)\boldsymbol{W}(t) - \boldsymbol{L}(t)\|_F^2 \end{aligned} \tag{2.12}$$

where $\boldsymbol{Y}(t)$ is a matrix of Lagrange multipliers and μ is a constant. Let $\widehat{\boldsymbol{W}}(t) = \boldsymbol{I} - \boldsymbol{W}(t)$, following the Chambolle-Pock algorithm for non-smooth primal dual algorithms, we update the following update rules for $\boldsymbol{W}(t)$ and $\boldsymbol{E}(t)$,

$$\boldsymbol{W_{k+1}}(t) = \arg min_{W(t)} \|\boldsymbol{\Omega}\|_{1,2} + \rho\|\boldsymbol{W}(t)\|_1 + < \boldsymbol{L_{k+1}}(t)\boldsymbol{W}(t) - \boldsymbol{L_{k+1}}(t), \boldsymbol{Y_k}(t) > + \frac{\mu_k}{2}\|\boldsymbol{L_{k+1}}(t)\boldsymbol{W}(t) - \boldsymbol{L_{k+1}}(t)\|_F^2 \quad (2.13)$$

$$\boldsymbol{E_{k+1}}(t) = \arg min_{E(t)}\ \lambda\|\mathbf{E}(t)\|_1 + < \boldsymbol{L_{k+1}}(t)\widehat{\boldsymbol{W}}_{k+1}(t) - \boldsymbol{L_{k+1}}(t), \boldsymbol{Y_k}(t) > + \frac{\mu_k}{2}\left\|\boldsymbol{L_{k+1}}(t)\widehat{\boldsymbol{W}}_{k+1}(t)\right\|_F^2 \quad (2.14)$$

The linearized ADMM in [47] is used to approximate Eqn. (2.13) and (2.14) as follows,

$$\boldsymbol{W_k^+}(t) = prox_{\frac{\rho}{\mu\eta_1}}\left(\boldsymbol{W_k}(t) + \frac{\boldsymbol{L_{k+1}^T}\left(\boldsymbol{L_{k+1}}\widehat{\boldsymbol{W}}_{\boldsymbol{k}}(t) - \frac{\boldsymbol{Y_k}(t)}{\mu_k}\right)}{\eta_1}\right) \quad (2.15)$$

$$\boldsymbol{W_{k+1}}(t) = \gamma_{\frac{\lambda}{\mu\eta_2}}(\boldsymbol{W_k^+}(t)) \quad (2.16)$$

$$\boldsymbol{E_{k+1}}(t) = \gamma_{\frac{\lambda}{\mu\eta_2}}\left(\boldsymbol{E_k}(t) + \frac{\left(\boldsymbol{L_{k+1}}\widehat{\boldsymbol{W}}_{\boldsymbol{k}}(t) - \frac{\boldsymbol{Y_k}(t)}{\mu_k}\right)\widehat{\boldsymbol{W}}_{\boldsymbol{k+1}}^{\boldsymbol{T}}}{\eta_2}\right) \quad (2.17)$$

Where $prox_\beta\left(A_{i,j}(t)\right) = A_{i,j}(t) * \max\left\{\left(\sqrt{\sum_{t=1}^{T} A_{i,j}(t)^2} - \beta\right), 0\right\} / \sqrt{\sum_{t=1}^{T} A_{i,j}(t)^2}$ and $\gamma_\tau\left(B_{i,j}\right) = sign\left(B_{i,j}\right) * \max\left\{\left(\left|B_{i,j}\right| - \tau\right), 0\right\}$. The Lagrange multipliers are updated as follows,

$$\boldsymbol{Y_{k+1}}(t) = \boldsymbol{Y_k}(t) + \mu_k(\boldsymbol{L_{k+1}}(t)\boldsymbol{W_{k+1}}(t) - \boldsymbol{L_{k+1}}(t)) \quad (2.18)$$

$$\mu_{k+1} = \epsilon\mu_k \quad (2.19)$$

2.2.3.2 Theoretical Discussion

Let $\boldsymbol{X}(t)$ represent the dataset with unit length data from the t^{th} modality for every $t \in [T]$. Moreover, *S(t)* = S^1(t) ∪ S^2 (t) ∪ . . . S^P (t) is the union of subspaces of the underlying structure,

where S^k(t) denotes the k^{th} subspace of the t^{th} modality. We seek the partitioning $\boldsymbol{X^1}$, $\boldsymbol{X}^2$, . . ., $\boldsymbol{X}^P$ of $[n]$ observations as previously elaborated. In the following, we state the theorem that will support our approach. We study the case for ρ= 0 for the sake of clarity and compare to the work in [50] for the deterministic model; in which the orientation of the subspaces as well as the distribution of the points in each subspace is deterministic. In particular, this theorem provides the required condition for the angle between subspaces to guarantee exact recovery. We show that our multi-modal approach provides looser and less restrictive bounds on the angle between the subspaces as compared to single-modal approach in [50]. More precisely, we prove that the RoGSuRe algorithm requires a smaller angle between different subspaces across all the modalities to guarantee their exact recovery, which explains the gain we achieve by leveraging multi-modal data fusion. Before proceeding, we will state some important definitions.

Definition 1 (Group Subspace Detection Property). The subspaces $\{S^I\}_{I=1}^P$ and points $\boldsymbol{X}(\boldsymbol{t})$ obey the group subspace detection if and only if it holds that for all i, the optimal solution to Eqn. (2.10) has nonzero entries only when the corresponding columns of $\boldsymbol{X}(t)$ are in the same subspace as $\boldsymbol{x}_i(t)$.

Definition 2. The inradius of a convex subset P of a finite dimension Euclidean space, denoted by $r(P)$, is defined as the radius of the largest Euclidean ball inscribed in P.

Definition 3. We take P_{-j}=$\left\{\left\{\xi(t)\boldsymbol{x}_q(t)\right\}_t \middle| \sum_{t=1}^T \xi^2(t) \leq 1, q \neq j\right\}$, where q belongs to the same subspace as j.

Theorem 1. Let $\boldsymbol{X}(\boldsymbol{t})$ represent the dataset with unit length data for every $t \in T$. Suppose that Eqn. (2.10) has a feasible solution Ω, where $w_{i,j}(t)$= 0 for all i, j not belonging to the same subspace. Let $\theta(t)$ be the smallest angle between vectors from distinct subspaces in the t^{th}

modality. For $\rho = 0$, if $max_t \cos^2 \theta(t) < min_j r^2(P_{-j})$, then the subspace detection property holds.

The theorem basically guarantees that the subspace detection property holds as long as for any two subspaces across all data modalities, the minimum angle is less than the minimum inradius of P_{-j} for all data modalities. It is easy to see that if the angle between a point on one subspace and an arbitrary direction on another (a dual direction) is small, these two subspaces will be close, hence, clustering becomes hard. Moreover, if the minimum inradius is small, which implies that the points are skewed towards specific direction and are not well spread throughout the subspace, therefore, the clustering will also be difficult. In short, the theorem affirms that as long as different subspaces for all data modalities are not likewise oriented and the points on each subspace are sufficiently spread and diverse, RoGSuRe will successfully cluster the data. By comparing our results to Theorem 2.5 in [50], it is easy to see that P_{-j} for $t > 1$ is much larger than the single modal approach in their paper and therefore the minimum inradius over t is larger. As a result, the angle between two subspaces has a smaller upper bound for multimodal data as compared to the single modal case in [50]. The proof of Theorem 1 is presented in Appendix A.

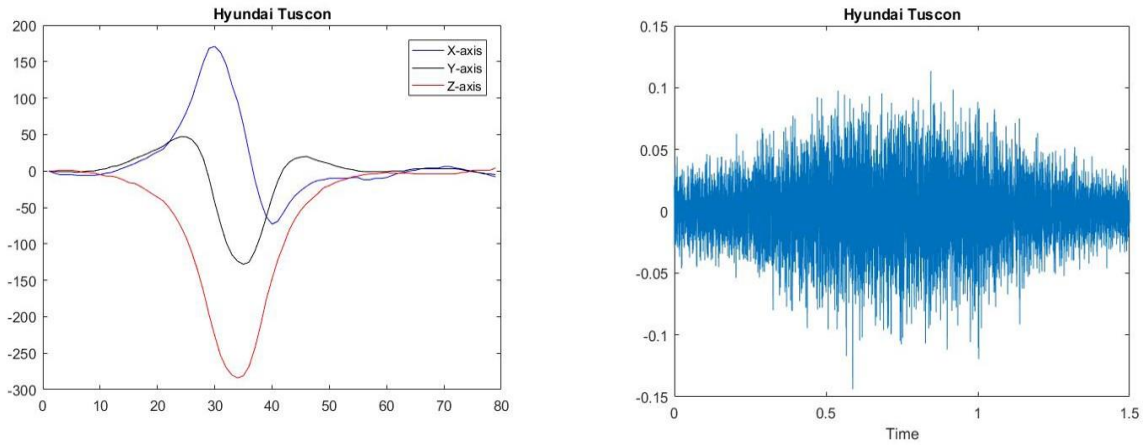
2.3 Multi-Modal Sensing in Vehicle Classification

As previously stated, the goal is to integrate the union of subspaces structure underlying the data measurements from each sensor modality to support decision making. A roadside sensor system was configured to collect data from passing vehicles using various sensors, including a camera, microphone, laser rangefinder, magnetometer, and low-frequency RF antenna. In this study, we are using the signatures captured using passive magnetic and acoustic sensors. The magnetic signatures are recorded using a single three-axis magnetic sensor, while the acoustic data is

collected by a single microphone. The sensors were mounted on a rigid rack for ease of deployment and management. The data collection was conducted in a park environment, with limited public interference. The data is collected for seven different vehicles: two SUVs, one sedan and four trucks. The two SUVs are GMC Yukon and Hyundai Tucson, the sedan is Honda Accord. The four trucks are Chevrolet pickup truck, 14 ft rental moving truck and two Ford F-150s, one has a mounted top on the bed and the other one does not. The seven different vehicles were driven by the system yielding a total of 546 observations per sensor. Our goal is to analyze this dataset and distinguish seven classes where each class corresponds to one car. Furthermore, our goal is to be able to classify a newly observed dataset, using the structure learned through the current unlabeled data. For this purpose, the observations were divided into training and testing as discussed in Table 2.1. As shown in the table, we used 50 observations for each car in the learning phase, and the rest of the observations for validation. Sample outputs of the magnetometer and microphone are shown in Fig. 2.1(a) and 2.1(b) respectively.

Table 2.1 Dataset Description

Vehicle	Training Points	Testing Points
Chevrolet Truck	50	29
Ford F-150 (Topper)	50	19
Ford F-150	50	31
GMC Yukon	50	24
Honda Accord	50	41
Hyundai Tuscon	50	20
Uhaul Truck	50	32



(a) (b)

Figure 2.1 Sample input data from both sensors. (a) Magnetometer data. (b) Audio data.

Acoustic sensors have been analyzed in various applications related to automatic transportation systems [51], [52]. Mel Frequency Cepstral Coefficients [53] are widely used in automatic speech recognition literature. They were introduced by Davis and Mermelstein in the 1980's and have been used extensively to date. The Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. We extract and process MFCCs from our audio data. In our experiments, a low pass filter was applied to the audio signals to remove noise. Signals were down sampled from 192 kHz to 64 kHz. The audio signals were divided into windows of size 0.025 seconds with a step size of 0.01 seconds to allow some overlap between the frames, and thereby get a reliable spectral estimate. Finally, the MFFCs were extracted for each window and the highest 25 coefficients are selected to result in 2900 log filter bank energies for each observation. This setting achieved the highest classification performance in our experiments. Magnetic sensors operate by detecting the variation in the magnetic inductance. Magnetic signatures can be characteristic of the vehicle of interest. Earth's magnetic field distortion can be used not only for

the detection, but also for the classification and recognition of transport vehicles [29], [30], [54]. The exploited three-axis system is capable of producing up to 154 Hz and outputs 16-bit values with 67 Gauss resolution. In our experiment, a sample rate of 40 Hz has been used. For calibration, the magnetic signatures were extracted from the magnetic signals by subtracting the value of the local magnetic field, which is measured when no car passed by the sensor. The beginning and the ending of the signal are subsequently determined. Each observation is then normalized and re-sampled to get a normalized length of 100 samples per axis for a total of 300 samples per observation. The X, Y and Z signal amplitudes are re-scaled/normalized to fall in the [−1,1] interval.

2.4 Experimental Results

2.4.1 RoSuRe-Based Fusion

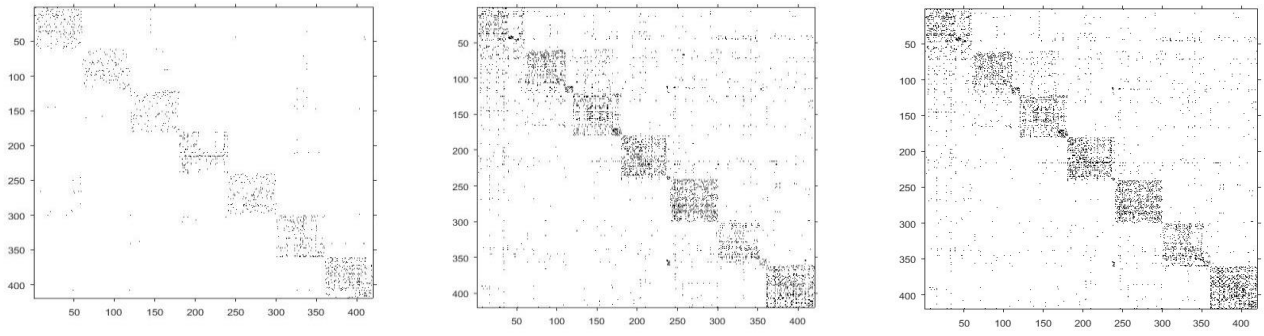
2.4.1.1 Applying Uni-Modal RoSuRe

In the following, we use the RoSuRe technique to recover the subspace structure embedded in the data associated with each of the magnetic and audio sensors. The sparse solution of the problem in Eqn. (2.4), $\boldsymbol{W}(\boldsymbol{t})$, provides important information about the relations among data points, which may be used to split data into individual clusters residing in a common subspace. Observations from each car can be seen as data points spanning one subspace. We first proceed to extract the principal components of each of the sensor data [55]. The largest 100 principal values for magnetometer data and the largest 11 principal components for audio data are selected to serve as representatives of the data in the principal component space. Using the enhanced lower dimensional representation of the data, the sparse UoS coefficient matrix is obtained using RoSuRe by way of Eqn. (2.4). The sparse coefficient matrix $\boldsymbol{W}(\boldsymbol{t})$ $\{t = 1$ for magnetometer, $t = 2$ for audio$\}$ is thus computed using PCA-based representation in lieu of $\boldsymbol{X}(t)$. The affinity matrix is

then calculated, and the spectral clustering classification technique explained in Section 2.2.2.1 is utilized to cluster the subspaces. The sparse coefficient matrices for magnetic and acoustic sensors are respectively illustrated in Fig. 2.2(a) and 2.2(b). The block-diagonal structure can be clearly seen from either of the matrices.

2.4.1.2 Bimodality RoSuRe-Driven Fusion (Multi-Modal RoSuRe)

Interpreting the subspace-based affinities based on $\boldsymbol{W(t)}$ as a layered set of networks, we proceed to carry out what amounts to modality fusion. The two sparse matrices $\boldsymbol{W_{audio}}$ and $\boldsymbol{W_{magnetic}}$ are added to produce one sparse matrix for both modalities, $\boldsymbol{W_{Total}}$, thereby improving performance. By doing so, we reinforce the contribution of similar representations that exist in both modalities as justified in Section 2.2.2.2. The overall sparse matrix, $\boldsymbol{W_{Total}}$ is displayed in Fig. 2.2(c). Observations belonging to one car are clustered as one subspace where the contribution of each sensor is embedded in the entries of the $\boldsymbol{W_{Total}}$. For clustering by $\boldsymbol{W_{Total}}$, we applied the same spectral clustering approach that we previously demonstrated in 2.2.2.1. As a result, the classification accuracy improved to 98.29% as highlighted in Table 2.2.



(a) (b) (c)

Figure 2.2 The sparse coefficient matrices for RoSuRe. (a) Coefficient matrix for magnetometer data. (b) Coefficient matrix for audio data. (c) Overall sparse coefficient matrix.

Table 2.2 Performance for Different Clustering Methods

	RoSuRe	k-means	GMM	HCA
Magnetometer data	86.71%	82.29%	77.14%	64.57%
Audio data	88.86%	52.1%	62.57%	40%
Fused	98.29%	82.29%	77.14%	64.57%

The performance of RoSuRe was compared against three widely used unsupervised clustering algorithms, namely, k-means, the Gaussian mixture model and hierarchical cluster analysis (HCA). k-means clustering, also referred to as the Lloyd-Forgy algorithm, is a computationally efficient method for cluster analysis in data mining [56]. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a representative of the cluster. A Gaussian mixture model is a probabilistic model which assumes all the data points generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Mixture models can be considered as a generalization for k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. Mixture models are in general less sensitive to the initialization of centroids. They have been used for feature extraction from speech data and object tracking [57], [58]. Hierarchical clustering is a technique which aims to build a hierarchy of clusters [59]. In our experiment, we used a bottom-up approach where all observations start in their own cluster, pairs of clusters are subsequently merged together according to their closeness. The Euclidean distance, $d(x_i, x_j) = \|x_i - x_j\|_2$, was used as a proximity measure between each pair of data points. We used complete-linkage criterion to measure the distance between clusters where the distance

$D(X,Y)$ between clusters X and Y is described as follows: $D(X,Y) = max_{x \in X, y \in Y} d(x,y)$. The results are displayed in Table 2.2. As shown in the table, RoSuRe has the highest classification accuracy for both audio and magnetometer data. Moreover, after fusing the two data modalities, RoSuRe shows a significant enhancement in the classification performance. Additionally, we compared the RoSuRe fusion performance with the other unsupervised clustering methods through linking the two modalities features. More specifically, we concatenated both magnetometer and audio observations in one vector and we then clustered the new representation of the data. The results in Table 2.2 show that, by concatenating the data, we are not gaining extra information. Moreover, the classification accuracy after concatenation is the same as that of the magnetometer data because of the dominant higher dimensionality of magnetometer observations as compared to audio observations. The results were therefore biased towards the former modality. Whereas, by integrating the sparse coefficient matrix corresponding to each modality, we have obviously boosted the performance of RoSuRe from approximately 86% to 98.29%.

2.4.2 Using Robust Group Subspace Recovery (RoGSuRe)

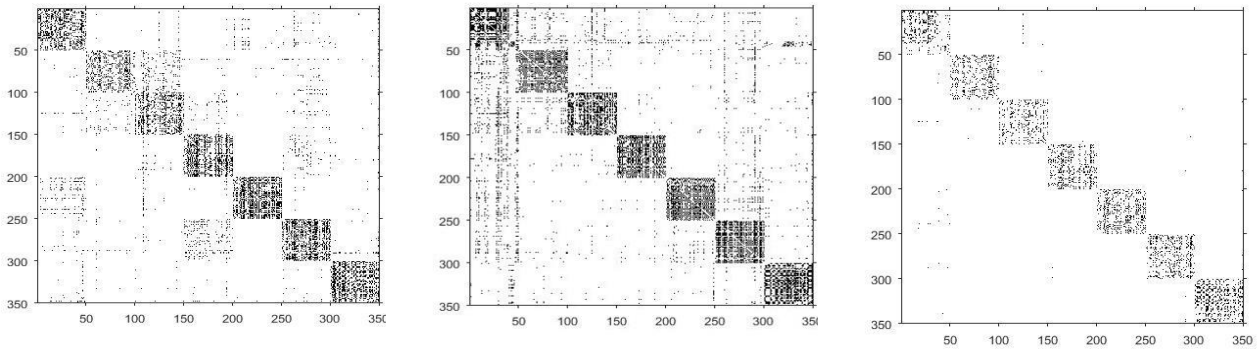
2.4.2.1 Fusing the Data Modalities with RoGSuRe:

In this subsection, we use the Robust Group Subspace Recovery technique to recover the subspace structure embedded in the data associated with each magnetic or audio observation. We follow the same data analysis explained in Section 2.3. We start by extracting the principal components of the data corresponding to each sensor [55] to serve as representatives of the data in the principal component space (in some sense denoised). The sparse coefficient matrix $\boldsymbol{W}(t)$ is computed by solving Eqn. (2.10). Next, we threshold $\boldsymbol{W}(t)$ by its median value. The sparse coefficient matrices for magnetic and acoustic sensors are respectively illustrated in Fig. 2.3(a) and 2.3(b). The block-

diagonal structure can be clearly seen from either of the matrices. Given that our UoS structure for the modalities are jointly obtained in this case, the cross-sensor consistency is achieved by their intersection and hence by a product of the two sparse binary matrices $\boldsymbol{W}(1)$ and $\boldsymbol{W}(2)$ to produce one sparse matrix for both modalities, $\boldsymbol{W_{Total}}$ resulting in an improved performance. Similar to ANDing process, the coefficient matrices from RoGSuRe are expected to share the same support and therefore their multiplication should yield a more reliable result. This is to be contrasted to the RoSuRe-based fusion when a weighted average for the sparse coefficient matrices worked much better than in that context since the data modalities might have different support so multiplication might lead to losing the unshared information.

$$\boldsymbol{W_{Total}} = \boldsymbol{W}(1).* \boldsymbol{W}(2) \tag{2.20}$$

The overall sparse matrix, $\boldsymbol{W_{Total}}$ is displayed in Fig. 2.3(c). Observations belonging to one car are clustered as one subspace in which the contribution of each sensor is embedded in the entries of the $\boldsymbol{W_{Total}}$. Using $\boldsymbol{W_{Total}}$, in clustering proceeded in the same way as previously pointed out in 2.2.2.1. As a result, the classification accuracy improved to 98.9% as highlighted in Table 2.3.



a) (b) (c)

Figure 2.3 The sparse coefficient matrices for RoGSuRe. (a) Sparse coefficient matrix for magnetometer data $\boldsymbol{W}(\mathbf{1})$. (b) Sparse coefficient matrix for audio data $\boldsymbol{W}(\mathbf{2})$**.** (c) The overall sparse coefficient matrix $\boldsymbol{W_{Total}}$.

Table 2.3 The Validation Performance for RoGSuRe and RoSuRe.

	Learning	Validation
RoGSuRe	99.14%	96.94%
RoSuRe	98.29%	94.82%

2.4.2.2 Experimental Validation with RoSuRe and RoGSuRe

After learning the structure of the data clusters, we validate our results on the test data. We first extract the principal components (eigen vectors of the covariance matrix) of each cluster in the original (training) dataset. We subsequently project each new test point onto the subspace corresponding to each cluster, spanned by its principal components. The l_2 norm of the projection is then computed, and the class with the largest norm is selected to correspond to the class of this test point. We use the coefficient matrix $\boldsymbol{W_{Total}}$ to cluster the test data points for both magnetometer and audio data. Classification on the new test data is jointly performed for both data modalities. The simulation results are listed in Table 2.3. From the results, it is clear that the Robust Group Subspace Recovery technique for the fused data remarkably outperforms RoSuRe.

In the following, we will compare the performance of RoGSuRe to some existing and known multimodal subspace clustering technique such as LTMSC [45], kMLRR [46], kMSSC [46], MSSC [46] and MLRR [46]. The results are depicted in Table 2.4. The reason our method outperforms the other methods is due to the group sparse term which does not enforce the similar structure across different modalities. Basically, the group-sparse term encourages different data modalities to communicate and share common information while at the same time each data modality maintains the relations between data points. The corresponding confusion matrices are

displayed in Fig. 2.4. Multimodal RoSuRe algorithm separately considers each view and ignores the correlation that might exist among different views. If we consider the case of low-quality data modalities, that might not share much commonality among their subspace structures, this will corrupt the support of the overall representation matrix, reduce the overall signal to noise ratio and dramatically degrade the performance. Similarly, MLRR, MSSC, KMSSC and KMLRR, will be negatively impacted in case of low-quality data modalities since they enforce the same structure among different data views. On the other hand, RoGSuRe will be minimally affected in this case, as it provides a T-factor (assuming T modalities) improvement by allowing modalities to strengthen repeated relations. In particular, when addressing a large number of modalities, the clustering improvement will be significant, and the gap between RoGSuRe and other approaches performance will be substantial. Other techniques, such as LTMSC, minimize the convex combination of the nuclear norms of all subspace representation matrices by seeking the lowest rank of the self-representation via a joint collaboration of multiple views. It, however, does not seem to provide richer information than unimodal data for our dataset.

Table 2.4 Performance for Different Multi-Modal Subspace Clustering Methods.

Method	Accuracy
LTMSC	75.43%
KMLRR	77.71%
KMSSC	79.14%
MLRR	89.14%
MSSC	98.29%
RoSuRe	98.29%
RoGSuRe	99.14%

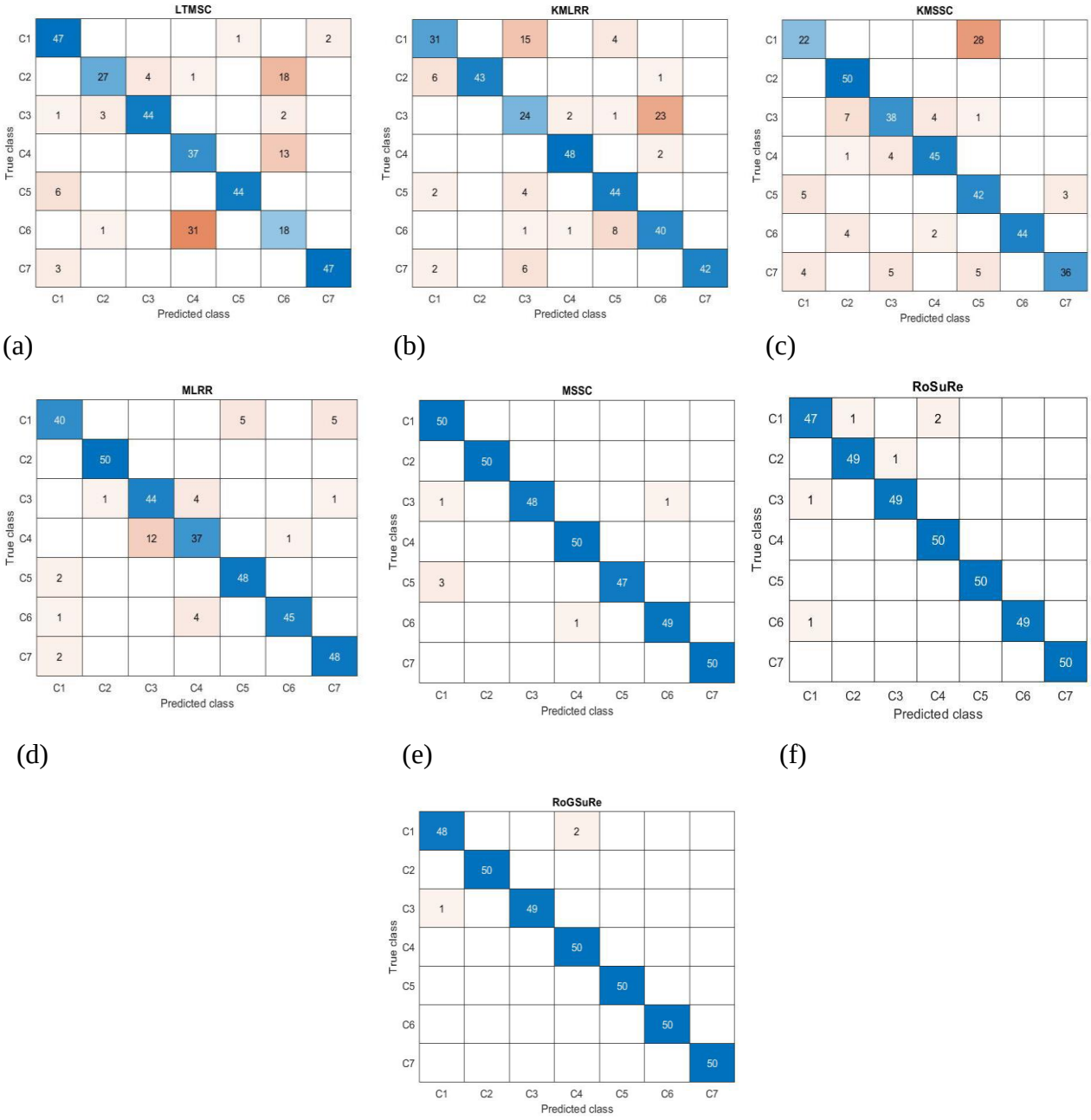


Figure 2.4 The confusion matrices for different multi-modal subspace clustering methods. (a) Confusion matrix for LTMSC. (b) Confusion matrix for KMLRR. (c) Confusion matrix for KMSSC. (d) Confusion matrix for MLRR. (e) Confusion matrix for MSSC. (f) Confusion matrix for RoSuRe. (g) Confusion matrix for RoGSuRe.

2.5 Conclusion

In this Chapter, we proposed two different approaches to fuse passive signal measured by low power instruments. The proposed approach recovers the underlying subspaces of data samples from measured data possibly corrupted by sparse errors. The multi-modal RoSuRe method is used to reliably and separately recover the subspace for each modality while the RoGSuRe manages to jointly optimize the subspace cluster structure. Both approaches provide a natural way to fuse multi-modal data by employing the self-representation matrix as an embedding in a shared domain. Experiments on real data are presented to demonstrate the effectiveness of this newly proposed method in solving the problem of subspace fusion with sparsely corrupted unlabeled data. Experimental results show a significant improvement for RoGSuRe over other state of the art subspace clustering techniques.

CHAPTER 3

Deep Robust Group Subspace Recovery

3.1 Introduction

Unsupervised learning is a very challenging topic in machine learning and involves discovering hidden patterns in data with no given labels. Reliable clustering techniques will save time and effort required for labeling large datasets that might have thousands of observations. Subspace clustering have been introduced as an efficient way for unfolding union of low-dimensional subspaces underlying high dimensional data. Subspace clustering has been extensively studied in computer vision due to the vast availability of visual data as in [1], [2], [3], and [4]. It has been deployed in many applications such as image segmentation [9], image compression [10], and object clustering [61].

Multimodal data have become more accessible with recent advances in sensor technology and on account of the pervasive use of sensors in every application. Different sensing often provides complementary information and offer richer information than unimodal data. A principled combination of the information contained in the different sensors and at different scales is hence likely to enhance understanding of the structure of the data. Uncovering the principles and laying out the fundamentals for multimodal data has become an important topic in research in light of many applications in diverse fields including image fusion [15], target recognition [16], speaker recognition [17] and handwriting analysis [18]. Convolutional neural networks have been widely used on multimodal data as in [62], [63], and [64]. The objective of this work is to devise a multi-modal framework for object clustering in an unsupervised learning scenario. We extract key

features from each data modality using a CNN encoder, and subsequently combine those features to generate a common discriminative feature. Our framework is based on the assumption that each data modality lies in a union of low dimensional subspaces that captures underlying hidden features. The UoS structure is unveiled by pursuing sparse self-representation of the given data modality. We subsequently aggregate the subspace structures corresponding to each data modality to a jointly unified characteristic subspace. We test our approach on two popular datasets, which we divide into learning and validation sets. The learned UoS structure is then utilized to classify new observed data points, which illustrates the generalization power of the proposed approach. By considering different scenarios by way of additive noise to either the training set, or the testing set, or both, we thoroughly investigate the robustness and resilience of the clustering approach performance.

3.2 Deep Robust Group Subspace Clustering

3.2.1 Problem Formulation

Consider a set of data realizations indexed by $k = 1, 2, \ldots, n$. Furthermore, assume T data modalities, indexed by $t = 1, 2, 3, \ldots, T$. Each data realization can be represented as a m-dimensional vector $\boldsymbol{x_k}(t) \in \mathbb{R}^m$, where $\mathbf{X}(t) = [\boldsymbol{x_1}(t)\ \boldsymbol{x_2}(t) \ldots \boldsymbol{x_n}(t)]$. The goal is to partition a set of realizations into clusters whose respective measurements for each modality is well-represented by a low-dimensional subspace. Mathematically, this is tantamount to seeking a partitioning $\{\boldsymbol{X^1}(t), \boldsymbol{X^2}(t), \ldots, \boldsymbol{X}^P(t)\}$ of $[n]$ observations, where P is the number of clusters indexed by p, such that there exist linear subspaces $S^p(t) \subset \mathbb{R}^m$ with dim $(S^p(t)) \ll$ m. Let $\boldsymbol{x}_k(t) \in S^p(t)\ \forall\ t$ and $k \in X^p$.

We will exploit the self-expressive property presented in [1] and [2], which entails that the representation of each sample as a linear combination of all samples from the same subspace. To proceed, we address the multi-modal data fusion by building on Deep Subspace Clustering (DSC) [65]. We next formulate a robust fusion of multi-modal sensor data, which we subsequently assess for potential scaling, and as a result propose an iterative solution achieved by a multiple/deep scale search using Convolutional Neural Networks. A diagram showing our algorithm is depicted in Figure 3.1. Our algorithm consists of three main stages; the first stage is the encoder which encodes the input modalities into a latent space. The encoder consists of T parallel CNN networks, where T is the number of data modalities. Each modality data is fed into one network, and the output of each network represents the modality data projection into its corresponding hidden/latent space. The second component of the auto encoder is T self-expressive layers, the goal of which is to enforce the self-expressive property among the data observations of each data modality. Each self-expressive layer is a fully connected layer which independently operates on the output of each encoder. The last stage is the decoder which reconstructs input data from the self-expressive layers' output. The objective function sought through this approximation network is reflected in Eqn. (3.3). The group sparsity introduced in [66] requires the minimization of the group norm of matrices $\boldsymbol{W}(\boldsymbol{t})$, which in turn, entails a smaller angle between the different spaces across all modalities, thus promoting the goal of obtaining a common latent space. Note that minimizing group norm provides as well a group sparse solution along data modalities. If we in addition, constrain the coefficient matrices corresponding to each data modality to commute, therefore, we ensure their sharing the same eigen vectors. The idea of commutation has been used in [67], [68], and [69]. We define $\boldsymbol{\Omega} = \{\boldsymbol{W}(t)\}_{t=1}^{T}$, where $\boldsymbol{W}(t) = [\boldsymbol{w_{kj}}(t)]_{k,j}$ and the group l-norm $\|\boldsymbol{\Omega}\|_{1,2}$ is defined as:

$$\|\boldsymbol{\Omega}\|_{1,2} = \sum_{k,j} \sqrt{\sum_{t=1}^{T} w_{k,j}^2(t)}. \tag{3.1}$$

We also define $[\boldsymbol{W}(t_1), \boldsymbol{W}(t_2)]$ as,

$$[\boldsymbol{W}(t_1), \boldsymbol{W}(t_2)] = \boldsymbol{W}(t_1)\,\boldsymbol{W}(t_2) - \boldsymbol{W}(t_2), \boldsymbol{W}(t_1) = 0 \tag{3.2}$$

The loss function is then rewritten as,

$$min_{\mathbf{W}(t)/\, w_{kk}(t)=0} \sum_{t_1,t_2}^{T} \|[\boldsymbol{W}(t_1), \boldsymbol{W}(t_2)]\|^2 + \|\boldsymbol{\Omega}\|_{1,2}$$

$$+ \frac{\gamma}{2}\sum_{t=1}^{T} \|\boldsymbol{X}(t) - \boldsymbol{X_r}(t)\|_F^2 + \rho \sum_{t=1}^{T} \|\boldsymbol{W}(t)\|_1 + \frac{\mu}{2}\sum_{t=1}^{T} \|\boldsymbol{L}(t) - \boldsymbol{L}(t)\boldsymbol{W}(t)\|_F^2 \tag{3.3}$$

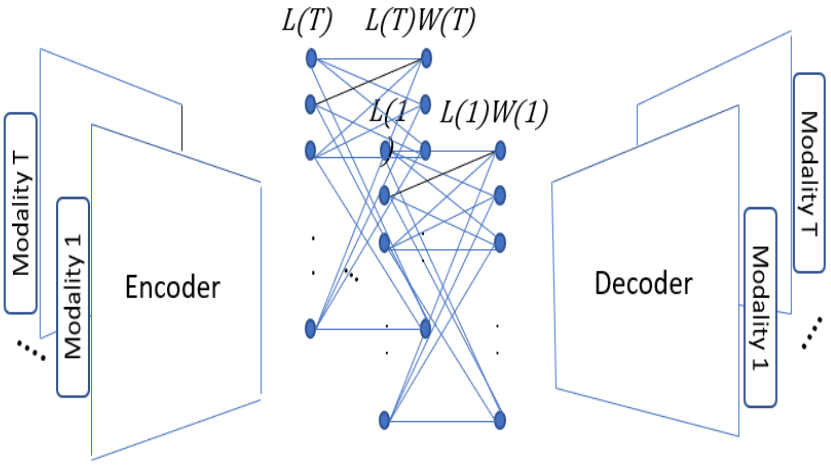


Figure 3.1 Deep Robust Group Subspace Clustering Diagram.

Where $\boldsymbol{X_r}(t)$ represent the reconstructed data corresponding to modality t, and $\boldsymbol{L}(t)$ is the output of the t^{th} encoder with input $\boldsymbol{X}(t)$. $\boldsymbol{W}(t)$ is the sparse weight function that ties the data observation for modality t. Solving DRoGSuRe in Tensorflow and using the adaptive momentum based gradient descent method (ADAM) [70] results in minimizing the loss function. For each data

modality, the weights of the encoder, the self-expressive layer and the decoder are individually calculated, however, fine-tuning the weights is based on the loss function, which is a function of the group norm and the pairwise product difference between sparse coefficient matrices. Under a suitable arrangement/permutation of the data realizations, the sparse coefficient matrix $\boldsymbol{W}(t)$ is an $n \times n$ block-diagonal matrix with zero diagonals provided that each sample is represented by other samples only from the same subspace. More precisely, $\boldsymbol{W_{ij}}(t) = 0$ whenever the indices i, j correspond to samples from different subspaces. As a result, most of the elements in $\boldsymbol{W}$ are equal to zero. $\| \|_1$ denotes the l_1 norm, i.e., the sum of absolute values of the argument. The Lagrangian objective functional may be rewritten as,

$$
\begin{aligned}
L\big(\boldsymbol{W}(t)\big) &= \sum_{t_1,t_2}^{T} \|[\boldsymbol{W}(t_1), \boldsymbol{W}(t_2)]\|^2 + \|\boldsymbol{\Omega}\|_{1,2} \\
&+ \rho \sum_{t=1}^{T} \|\boldsymbol{W}(t)\|_1 + \frac{\gamma}{2} \sum_{t=1}^{T} \|\boldsymbol{X}(t) - \boldsymbol{X_r}(t)\|_F^2 + \sum_{t=1}^{T} \frac{\mu}{2} \|\boldsymbol{L}(t)\boldsymbol{W}(t) - \boldsymbol{L}(t)\|_F^2 \\
&+ \sum_{t=1}^{T} < \boldsymbol{L}(t)\boldsymbol{W}(t) - \boldsymbol{L}(t), \boldsymbol{Y}(t) >
\end{aligned}
\tag{3.4}
$$

Assume $\widehat{\boldsymbol{W}}(t) = \boldsymbol{I} - \boldsymbol{W}(t)$, we update $\boldsymbol{W}(\boldsymbol{t})$ as follows,

$$
\begin{aligned}
\boldsymbol{W_{k+1}}(t) = \arg min_{\boldsymbol{W}(t)} &\sum_{t_1,t_2}^{T} \|[\boldsymbol{W}(t_1), \boldsymbol{W}(t_2)]\|^2 + \|\boldsymbol{\Omega}\|_{1,2} + \rho\|\boldsymbol{W}(t)\|_1 + \\
&< \boldsymbol{L_{k+1}}(t)\boldsymbol{W}(t) - \boldsymbol{L_{k+1}}(t), \boldsymbol{Y_k}(t) > + \frac{\mu_k}{2} \|\boldsymbol{L_{k+1}}(t)\boldsymbol{W}(t) - \boldsymbol{L_{k+1}}(t)\|_F^2
\end{aligned}
\tag{3.5}
$$

Similar to [1], we utilize linearized ADMM [47] to approximate the minimum of Eqn. (3.5) since the algorithmic solution is complicated and yields a non-convex optimization functional. It has been shown that linearized ADMM is very effective for l_1 minimization problems and the

augmented Lagrange multiplier (ALM) method can take care of the non-convexity of the problem [71] [72]. Therefore, utilizing an appropriate augmented Lagrange multiplier μ_k, we can compute the global optimizer by solving the dual problem. The solution to Eqn. (3.5) can be approximated, using linearized soft thresholding, as follows,

$$\begin{aligned} \boldsymbol{W_k^+}(t) = prox_{\frac{\rho}{\mu\eta_1}}(\boldsymbol{W_k}(t) + \frac{\boldsymbol{L_{k+1}^T}(\boldsymbol{L_{k+1}}\widehat{\boldsymbol{W}}_{\boldsymbol{k}}(t) - \frac{\boldsymbol{Y_k}(t)}{\mu_k})}{\eta_1} \\ + \sum_{t_1,t_2=1,t_1\neq t_2}^{T} \{\big(\boldsymbol{W_k}(t_1)\boldsymbol{W_k}(t_2) - \boldsymbol{W_k}(t_2)\boldsymbol{W_k}(t_1)\big)W_k^T(t_2) \\ + W_k(m)(\boldsymbol{W_k}(t_1)\boldsymbol{W_k}(t_2) - \boldsymbol{W_k}(t_2)\boldsymbol{W_k}(t_1))\} \end{aligned} \tag{3.6}$$

$$\boldsymbol{W_{k+1}}(t) = \gamma_{\frac{\rho}{\mu\eta_2}}(\boldsymbol{W_k^+}(t)) \tag{3.7}$$

Where $\eta_1 \geq \|L\|_2^2$. We alternatively update $\boldsymbol{L}(\boldsymbol{t})$ as,

$$\boldsymbol{L_{k+1}}(t) = \boldsymbol{L_k}(t) + \mu_k\left(\boldsymbol{L_k}(t)\widehat{\boldsymbol{W}}_{k+1}(t) - \frac{Y_k(t)}{\mu_k}\right)\widehat{W}_{k+1}^T(t). \tag{3.8}$$

Where $prox_\beta\left(A_{i,j}(t)\right) = A_{i,j}(t) * \max\{\left(\sqrt{\sum_{t=1}^T A_{i,j}(t)^2} - \beta\right), 0\}/\sqrt{\sum_{t=1}^T A_{i,j}(t)^2}$ and $\gamma_\tau(B_{i,j}) = sign(B_{i,j}) * \max\{(|B_{i,j}| - \tau), 0\}$. The Lagrange multipliers are updated as follows,

$$\boldsymbol{Y_{k+1}}(t) = \boldsymbol{Y_k}(t) + \mu_k(\boldsymbol{L_{k+1}}(t)\boldsymbol{W_{k+1}}(t) - \boldsymbol{L_{k+1}}(t)) \tag{3.9}$$

$$\mu_{k+1} = \epsilon\mu_k \tag{3.10}$$

After computing the gradient of the loss function, the weights of each multi-layer network, that corresponds to one modality, are updated while other modalities' networks are fixed. In other words, after constructing the data during the forward pass, the loss function determines the updates that back-propagates through each layer. The encoder of the first modality is updated, afterwards,

the self-expressive layer of that modality gets updated and finally the decoder. Since the weights corresponding to each modality are dependent on other modalities, we update each part of the network corresponding to each modality with the assumption that all other networks' components corresponding to other modalities are fixed.

3.2.2 Class Partitioning

To proceed with distinguishing the various classes in an unsupervised manner, we first evaluate the affinity matrix as detailed in [48]. The affinity matrix is computed as,

$$\mathbf{A} = \boldsymbol{W} + \boldsymbol{W}^T \tag{3.11}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Briefly, a matrix $\boldsymbol{D}$ is defined to be a diagonal matrix whose i^{th} diagonal element is the degree of the i^{th} node, i.e., the sum of i^{th} row in $\boldsymbol{A}$. The standard graph Laplacian matrix is next constructed as follows,

$$\mathbf{G} = \boldsymbol{D}^{-1/2}\mathbf{A}\boldsymbol{D}^{-1/2} \tag{3.12}$$

where $\mathbf{G} \in \mathbb{R}^{n \times n}$. Next, the eigenvectors $\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_r}$ of $\boldsymbol{G}$ corresponding to the largest r eigenvalues are computed, where r is the desired number of clusters. The matrix $\boldsymbol{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_r}]$ is then formed by stacking the eigenvectors in columns. Each row of $\boldsymbol{E}$ is a point in $\mathbb{R}^r$, k-means clustering is then used to cluster the rows of $\boldsymbol{E}$. Finally, the original point $\boldsymbol{x}_i$ is assigned to cluster j iff row i of the matrix $\boldsymbol{E}$ was assigned to cluster j. The resulting sparse coefficient matrices $\boldsymbol{W}(t)$'s, for $t = 1, 2, \dots, T$ are then integrated as follows,

$$\boldsymbol{W_{Total}} = \sum_{t=1}^{T} \boldsymbol{W}(t) \tag{3.13}$$

By doing so, we are reinforcing the relation between data points that exist in all data modalities as reflected by the elements of $\boldsymbol{W}(t)$. While this may be justified as ensuring a cross-sensor

consistency, we are also reducing the noise variance introduced by the outliers. A similar strategy was explored [49] in community detection in Social Networks, where an aggregation of multi-layer adjacency matrices was found to yield a better Signal to Noise ratio, and ultimately improved performance. In multilayer networks, edges that exist in multiple layers, encode different but related relations among data points. The subspaces/clusters in our case, share model commonalities for given targets for which the relations among data observations are reflected by the sparse (nonzero) elements of $\boldsymbol{W}(t)$.

3.2.3 Theoretical Discussion

In order to justify the multiple banks of self-expressive layers, we assume that each modality $\boldsymbol{X}(t)$ may be expressed as a private information contribution $\boldsymbol{X}_p(t)$ and a shared information $\boldsymbol{X}_s(t)$ such that,

$$\boldsymbol{X}(t) = \boldsymbol{X_s}(t) + \boldsymbol{X_p}(t) \tag{3.14}$$

The shared information can be represented as follows,

$$\boldsymbol{X_s}(t) = \sum_{t=1}^{T} F(\boldsymbol{W}(t)(\Pi_s \boldsymbol{X}(t))) \tag{3.15}$$

Where $\Pi_s = \cap_{t=1,\ldots,T} \Pi_s^t$. $\boldsymbol{X_s}(t)$ and $\boldsymbol{X_p}(t)$ are distinct and will hence lie in different subspaces, which will hence be mapped to different components in $\boldsymbol{W}(t)$. Similarly for the subspaces spanned by $\boldsymbol{X_p}(t_i)$ and $\boldsymbol{X_p}(t_j)$, $i \neq j$, the corresponding components of $\boldsymbol{W}(t_i)$ and $\boldsymbol{W}(t_j)$ will almost surely not coincide. On the other hand, the components of $\boldsymbol{W}(t_i)$ and $\boldsymbol{W}(t_j)$ corresponding to $\boldsymbol{X_s}(t_i)$ and $\boldsymbol{X_s}(t_j)$ will almost surely coincide, thus justifying the construction of a layered $\boldsymbol{W}_{Total}$, and thereby improving the SNR. In addition, the decoder will help protect and maintain the private information corresponding to each modality $\boldsymbol{X}_p(t)$ by ensuring that data can be reconstructed

again from the latent space with minimal loss. In the following, we will elaborate more on how aggregating affinity matrices should impact the overall clustering performance. The idea of aggregating affinity matrices is not new, in fact, it has been used extensively in clustering and community detection field. For example, in [73], the authors proposed a method that combines the self-similarity matrices of the eigenvectors after applying a Singular Value Decomposition on clusters. In [74], they proposed merging the information provided by the multiple modalities by combining the characteristics of individual graph layers using tools from subspace analysis on a Grassmann manifold. In [75], they propose a multilayer spectral graph clustering (SGC) framework that performs convex layer aggregation. In Appendix B, we show that by perturbing one or more data modalities, our proposed approach introduces less error to the overall affinity matrix as compared to DMSC. Hence, preserving the performance and yielding a graceful degradation of the clustering accuracy as an increasing number of modalities get corrupted by noise.

3.3 Affinity Fusion Deep Multimodal Subspace Clustering

For completeness, we provide a brief overview of the Deep Multimodal Subspace Clustering algorithm which was proposed in [34]. As noted earlier for DRoGSuRe and similarly for Affinity Fusion Deep Multimodal Subspace clustering (AFDMSC), the network is composed of three main parts: a multimodal encoder, a self-expressive layer, and a multimodal decoder. The output of the encoder contributes to a common latent space for all modalities. The self-expressiveness property applied through a fully connected layer between the encoder and the decoder results in one common set of weights for all the data sensing modalities. This marks a divergence in defining the latent space with DRoGSuRe. Our proposed approach, as a result, safeguards the private information $\boldsymbol{X_p}(t); t = 1, \dots, T$ individually for each of the sensors, i.e., dedicating more degrees

of freedom for each of the sensors. This is in contrast to AFDMSC. The reconstruction of the input data by the decoder, can yield the following loss function to secure the proper training of the self-expressive network,

$$min_{W|w_{kk}=0}\|\boldsymbol{W}\|_2 + \frac{\gamma}{2}\sum_{t=1}^{T}\|\boldsymbol{X}(t) - \boldsymbol{X_r}(t)\|_F^2 + \frac{\mu}{2}\sum_{t=1}^{T}\|\boldsymbol{L}(t) - \boldsymbol{L}(t)\boldsymbol{W}\|_F^2, \tag{3.16}$$

where $\boldsymbol{W}$ represents the parameters of the self-expressive layer, $\boldsymbol{X}(t)$ is the input to the encoder, $\boldsymbol{X_r}(t)$ denote the output of the decoder and $\boldsymbol{L}(t)$ denotes the output of the encoder. μ and γ are regularization parameters. An overview for the DMSC approach is illustrated in Figure 3.2.

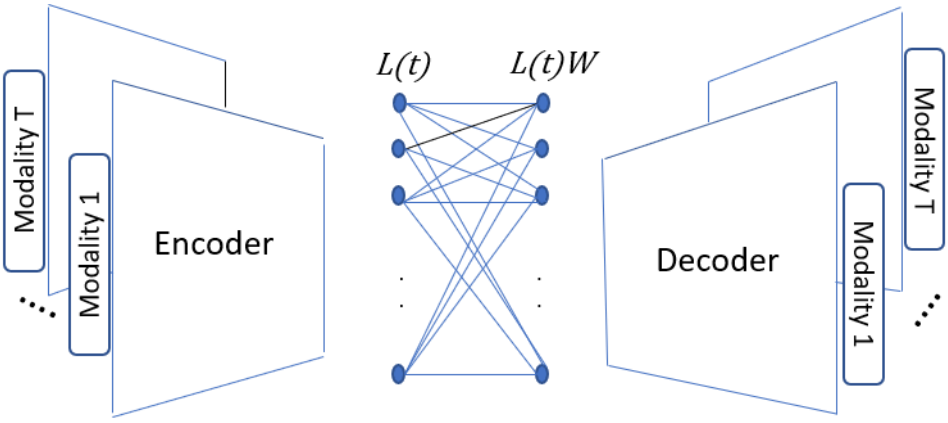


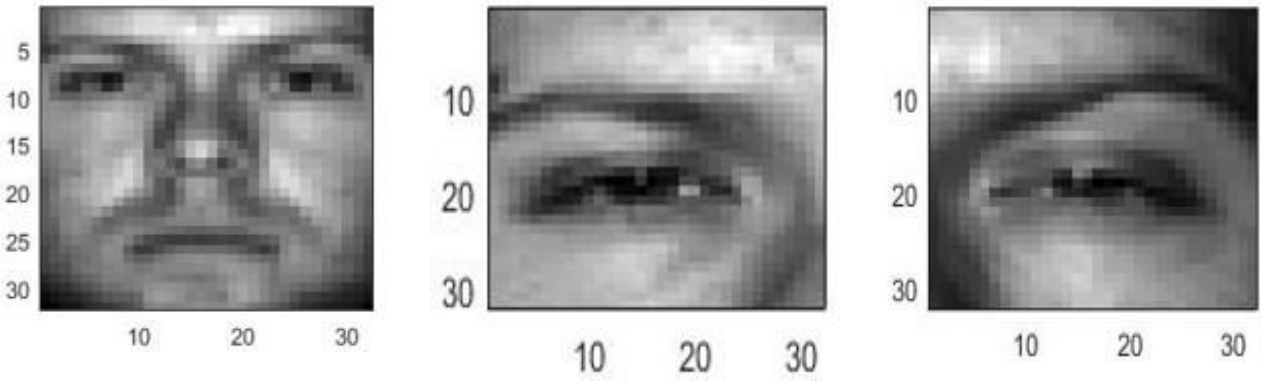
Figure 3.2 Deep Multimodal Subspace Clustering Diagram.

3.4 Experimental Results

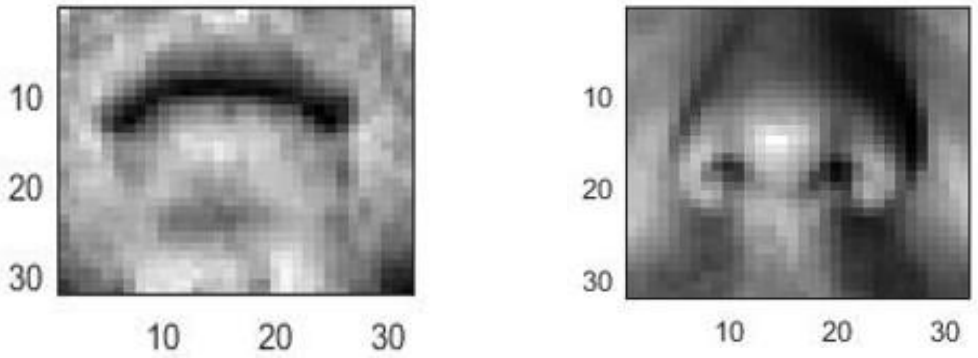
3.4.1 Dataset Description

We will evaluate our approach on two different datasets. The first dataset we will use is the Extended Yale-B dataset [76]. The same dataset has been used extensively in subspace clustering as in [2], [77]. The dataset is composed of 64 frontal images of 38 individuals under different

illumination conditions. In this work, we will use the augmented data used in [34], where facial components such as left eye, right eye, nose and mouth have been cropped to represent four additional modalities which are spatially unrelated. Images corresponding to each modality have been cropped to a size of 32×32. A sample image for each modality is shown in Figure (3.3).



(a) (b) (c)

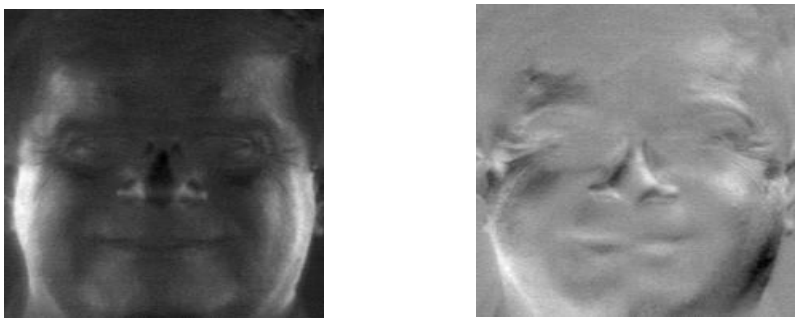


(d) (e)

Figure 3.3 Sample Images from the Augmented Extended Yale-B Dataset. (a) Face. (b) Left eye. (c) Right eye. (d) Mouth. (e) Nose.

The second validation dataset we use is the ARL polarimetric face dataset [78]. This consists of facial images for 60 individuals in the visible domain and in four different polarimetric states. All the images are spatially aligned for each subject. We have also resized the images to 32×32 pixels. Sample images from this dataset are shown in Figure (3.4).

(a) (b) (c)



(d) (e)

Figure 3.4 Sample Images from the ARL Polarimetric Dataset. (a) Visible. (b) DoLP. (c) S0. (d) S1. (e) S2.

3.4.2 Network Structure

In the following, we will elaborate on how we construct the neural network for each dataset. Similarly to [34], we implemented DRoGSuRe with Tensorflow and used the adaptive momentum based gradient descent method (ADAM) [70] to minimize the loss function in Eqn. (3.3) with a learning rate of 10^{-3}.

3.4.2.1 Army Research Laboratory (ARL) Dataset

In case of ARL dataset, we have five data modalities and will therefore have 5 different encoders, self-expressive layers and decoders. Each encoder is composed of three neural layers. The first layer consists of 5 convolutional filters of kernel size 3. The second layer has 7 filters of kernel size 1. The last layer has 15 filters with kernel size equals 1.

3.4.2.2 Extended Yale-B (EYB) Dataset

For EYB dataset, we also have five data modalities, therefore, we have 5 different encoders, self-expressive layers and decoders. Each encoder is composed of three neural layers. The first layer consists of 10 convolutional filters of kernel size 5. The second layer has 20 filters of kernel size 3. The last layer has 30 filters of kernel size 3.

3.4.3 Noiseless Results

In the following, we compare the performance of our approach versus the DMSC approach when learning the union of subspaces structure of noise-free data. First, we divide each dataset into training and validation sets to be able to classify a newly observed dataset, using the structure learned through the current unlabeled data. The ARL expression dataset used for training consists of 2160 images per modality. The validation baseline images include 720 images total per modality. For the EYB, we randomly selected 1520 images per modality for training and 904 images for validation. The sparse solution $\boldsymbol{W}(\boldsymbol{t})$ corresponding to each data modality, provides important information about the relations among data points, which may be used to split data into individual clusters residing in a common subspace. Observations from each object can be seen as data points spanning one subspace. Interpreting the subspace-based affinities based on $\boldsymbol{W}(\boldsymbol{t})$ as a layered set of networks, we proceed to carry out what amounts to modality fusion. The T sparse matrices are added to produce one sparse matrix for both modalities, $\boldsymbol{W_{Total}}$, thereby improving

performance. Observations associated with one object are clustered as one subspace where the contribution of each sensor is embedded in the entries of $\boldsymbol{W_{Total}}$. For clustering by $\boldsymbol{W_{Total}}$, we applied the same spectral clustering approach that we previously demonstrated in Section 3.2.2. After learning the structure of the data clusters, we validate our results on the validation set. We extract the principal components (eigen vectors of the covariance matrix) of each cluster in the original (training) dataset, to act as a representative subspace of its corresponding class.

We subsequently project each new test point onto the subspace corresponding to each cluster, spanned by its principal components. The l_2 norm of the projection is then computed, and the class with the largest norm is selected to be the class of this test point. For DRoGSuRe, we use the coefficient matrix $\boldsymbol{W_{Total}}$ in Equation (3.5) to cluster the test data points coming from all data modalities. We compare the clustering output labels with the ground truth for each dataset. The results for ARL and EYB datasets are depicted in Tables (3.1) and (3.2) respectively. From the results, it is clear that DRoGSuRE technique for the fused data remarkably outperforms DMSC in case of ARL dataset. However, in case of EYB dataset and in the noiseless case, DMSC performed better than DRoGSuRe.

Table 3.1 Performance Comparison for ARL Dataset.

	Learning	Validation
DMSC	97.59%	98.33%
DRoGSuRe	100%	100%

Table 3.2 Performance Comparison for EYB Dataset.

	Learning	Validation
DMSC	98.82%	98.89%
DRoGSuRe	98.42%	98.76%

3.4.4 Noisy training with single and multiple modalities

In the following, we test the robustness of our approach in the case of noisy learning. We distort one modality at a time by shuffling the pixels of all images in that particular modality during the training phase. By doing so, we are perturbing the structure of the sparse coefficient matrix associated with that modality, thus impacting the overall $\boldsymbol{W}$ matrix for both DRoGSuRe and DMSC. Testing with clean data, i.e., no distortion, demonstrates the impact of perturbing the training and hence performing an inadequate training, e.g., insufficient data or non-convergence. This can also be considered as augmenting the training data with new information or a new view for one modality which might not necessarily contained in the testing or the validation data. Moreover, we repeat the same experiment with the distortion of two modalities before learning the sparse coefficient matrices for both DMSC and DRoGSuRe. The results for the ARL dataset are depicted in Table (3.3) and (3.4), while results for the EYB dataset are shown in Table (3.5) and (3.6). For ARL dataset, we refer to Visible, S0, S1, S2 and DoLP as Mod 0, 1, 2, 3 and 4 respectively. For EYB Dataset, we refer to Face, left eye, nose, mouth and right eye as mod 0, 1, 2, 3, and 4. We refer to each modality as Mod, where L denotes learning and V denotes validation results. From the results, it is clear that DRoGSuRe is showing a significant improvement in the clustering accuracy as compared to DMSC for both learning and validation set. The reason for

that, is again, since perturbing one or two modalities would have less impact on the overall performance for DRoGSuRe in comparison to DMSC.

Table 3.3 ARL Dataset: Distorting One Modality.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0	87.17%	86.67%	95.37%	95%
Mod 1	91.67%	90%	98.29%	98.33%
Mod 2	92.77%	92.78%	99.17%	99.44%
Mod 3	90.55%	90.57%	99.31%	99.44%
Mod 4	92.78%	91.11%	96.44%	96.67%

Table 3.4 ARL Dataset: Distorting Two Modalities.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0 & 1	82.22%	82.78%	92.27%	94.58%
Mod 1 & 2	91.11%	91.11%	97.22%	97.36%
Mod 0 & 3	85.51%	82.56%	93.01%	95.42%
Mod 1 & 4	91.67%	89.44%	97.22%	97.36%
Mod 2 & 3	90%	89.72%	97.69%	97.78%

Table 3.5 EYB Dataset: Distorting One Modality.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0	87.96%	88.5%	93.29%	94.69%
Mod 1	91.84%	91.15%	95.79%	97.46%
Mod 2	89.01%	88.72%	98.03%	97.57%
Mod 3	92.69%	91.81%	95.59%	96.68%
Mod 4	91.45%	91.59%	97.17%	97.35%

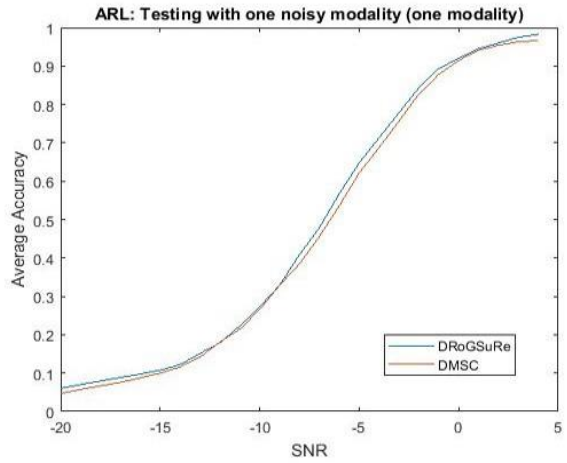
Table 3.6 EYB Dataset: Distorting Two Modalities.

	DMSC L	DMSC V	DRoGSuRe L	DRoGSuRE V
Mod 0 & 2	86.64%	85.18%	96.84%	96.13%
Mod 0 & 4	87.83%	89.16%	94.54%	95.8%
Mod 1 & 4	86.38%	86.06%	94.21%	95.8%
Mod 2 & 3	88.22%	84.96%	91.58%	93.92%
Mod 3 & 4	88.03%	86.28%	94.08%	95.35%

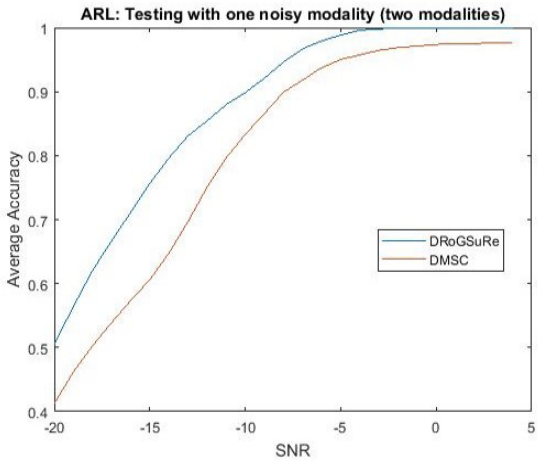
3.4.5 Testing with limited noisy data

In the following, we study the effect of using noiseless data for training while validating with noisy and missing data. We add Gaussian noise to one data modality in the validation set and vary the SNR by varying the noise variance. We subsequently assume that we only have one modality available at testing. Then, we keep increasing the number of available noiseless data modalities beside the noisy modality. We average the results considering all different combinations of data modalities for ARL and EYB datasets. The results are depicted in Figure (3.5) and (3.6) respectively. For the ARL dataset, we note the increasing gap between DMSC and DRoGSuRe as

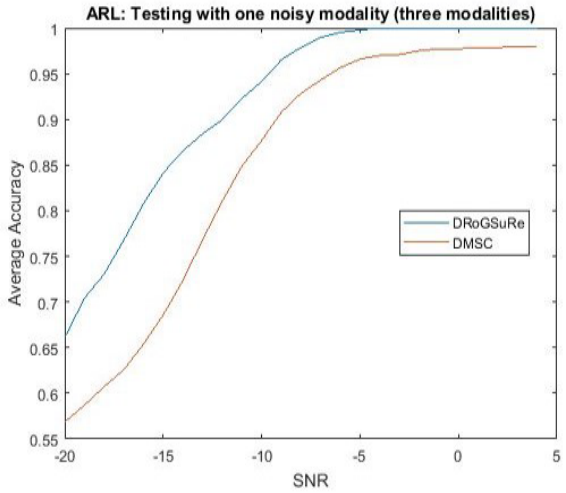
we augment the sensing capacity with noise free modalities. On the other hand, for the EYB dataset and at lower SNR, the performance of DRoGSuRe is slightly worse than DMSC which might be explained by the results in Table (3.2); as the training accuracy for DMSC is slightly better than DRoGSuRe in the case of clean training. However, at higher SNR, the performance of the two approaches is very close.



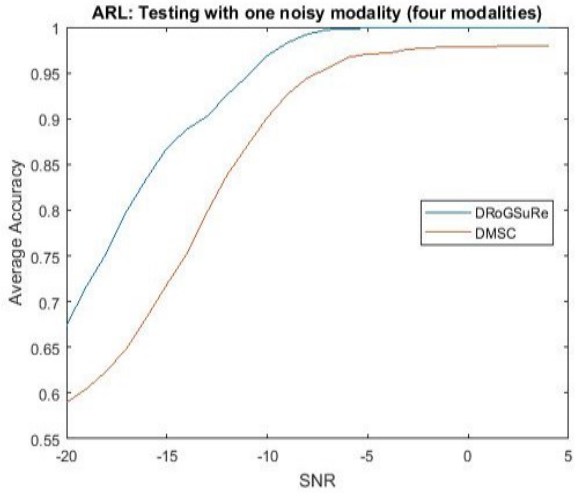
(a)



(b)



(c)



(d)

Figure 3.5 ARL Noiseless Training and validating on Limited Noisy Data.

3.4.6 Missing modalities during testing

In the following, we evaluate the performance of DRoGSuRe and DMSC in case of missing data modalities during testing. It is not uncommon to have one or more sensors that might be silent during testing, thus justifying this experiment for further assessment. We try different combinations of available modalities during testing, and we average the clustering accuracy for each trial. Results are depicted in Figures (3.7) and (3.8) for ARL and EYB data respectively. Again, we notice a significant improvement for DRoGSuRe over DMSC for ARL Dataset. For

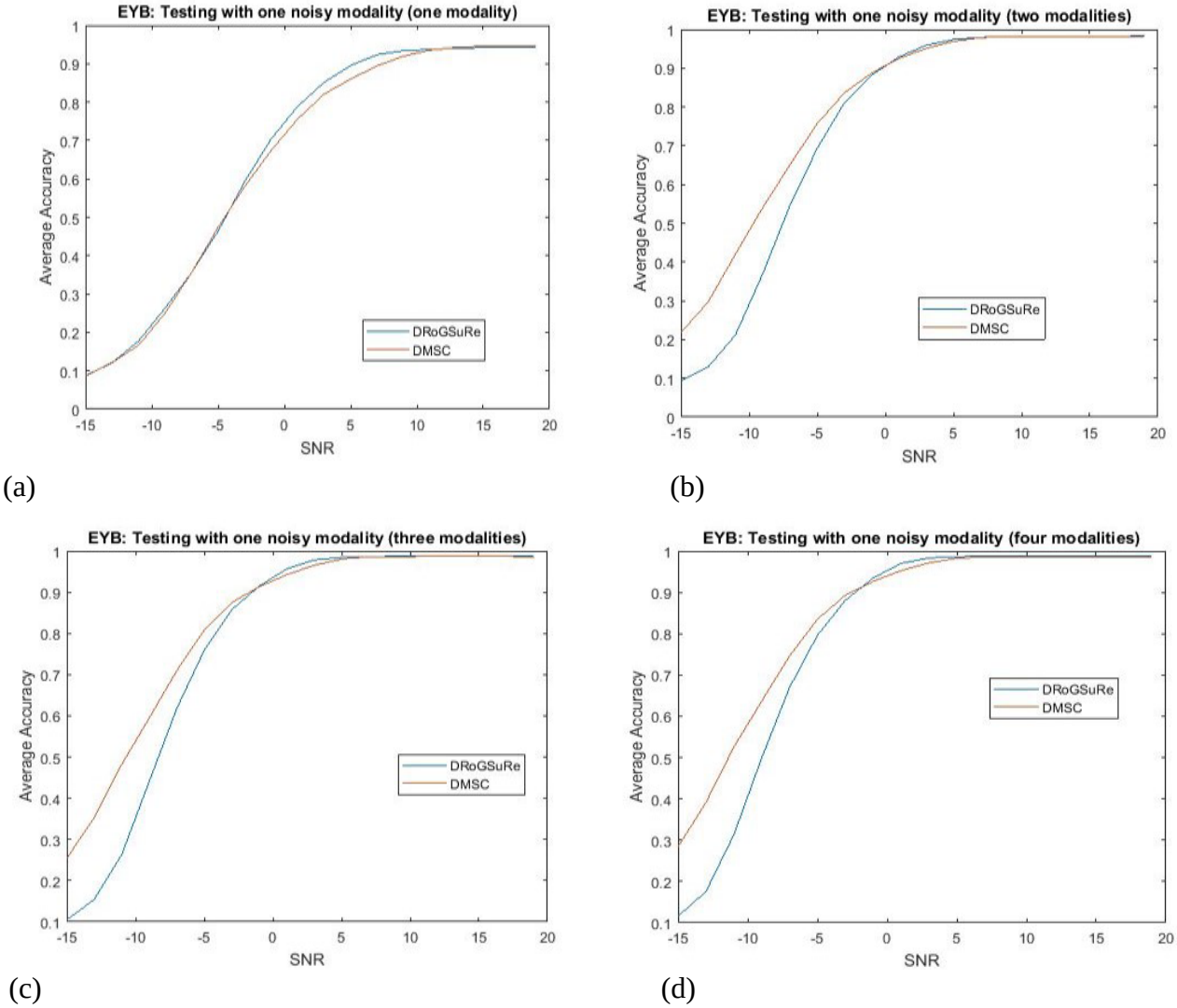


Figure 3.6 EYB Noiseless Training and Validating on Limited Noisy Data.

EYB dataset, there is a slight improvement for DRoGSuRe over DMSC, however, the performance of DRoGSuRe is gracefully degrading while more modalities become unavailable during testing, which emphasizes the reliability and robustness of our proposed approach.

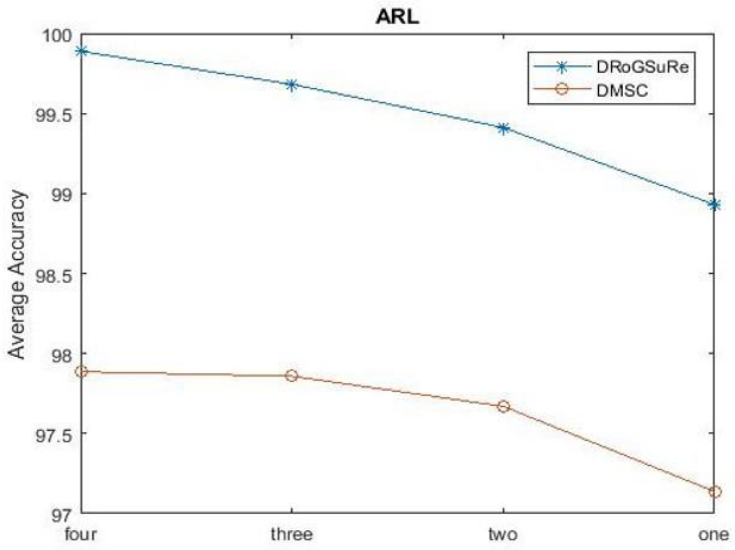


Figure 3.7 Missing Modalities during Testing for ARL Dataset.

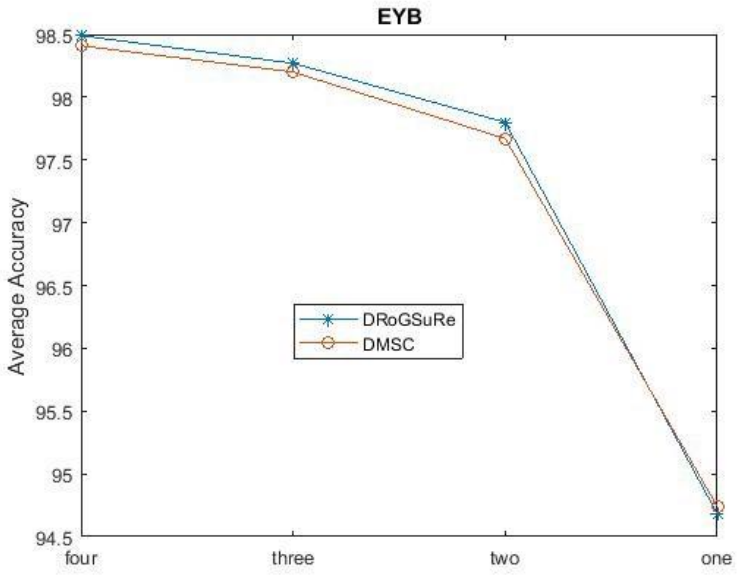


Figure 3.8 Missing Modalities during Testing for EYB Dataset.

3.5 Feature Concatenation

Here we propose a rationale along with an alternative solution for enhancing the performance for EYB multi-modal data. Due to the specific structure of the EYB multi-modal data, the concatenation of the features corresponding to each modality is a reasonable alternative. By doing so, we are adjoining together the features representing each part of the face. Since the four modalities correspond to non-overlapping partitions of the face, the feature set corresponding to each partition will solely provide complementing information. A similar idea is proposed in [34] and is referred to as Late concatenation, where the multi-modal data is integrated in the last stage of the encoder. Their resulting decoder structure remains the same for either affinity fusion or late concatenation. This entails deconcatenating the multi-modal data prior to decoding it. Our proposed approach on the other hand, results in a self-expressive layer being driven by the concatenated features from the M encoder branches. Afterwards, we feed the self-expressive layer

output to each branch of the decoder. The concatenated information results in a more efficient code for the data, thereby resulting in an overall parsimonious with a sparse structure of the decoder, results in a decoder composed of three neural layers. The first layer consists of 150 filters of kernel size 3. The second layer consists of 20 layers of kernel size 3. The third layer consists of 10 layers of kernel size 5. Our approach is illustrated in Figure (3.9). We optimize the weights of the auto-encoder as follows,

$$min_{W|w_{kk}=0}\rho\|\boldsymbol{W}\|_1+\frac{\gamma}{2}\sum_{t=1}^{T}\|\boldsymbol{X}(t)-\boldsymbol{X_r}(t)\|_F^2+\frac{\mu}{2}\sum_{t=1}^{T}\|\boldsymbol{N}-\boldsymbol{NW}\|_F^2, \tag{3.17}$$

Where $N=[L(1)\|L(2)\|L(3)\|L(4)\|L(5)]$. We compared the performance of our proposed approach against the late concatenation approach in [34] and the results are depicted in Table (3.7).

Table 3.7 Concatenation Performance for EYB Dataset.

	Learning	Validation
DMSC Late Concatenation	95.66%	94.7%
CNN Concatenation Network	99.28%	99.3%

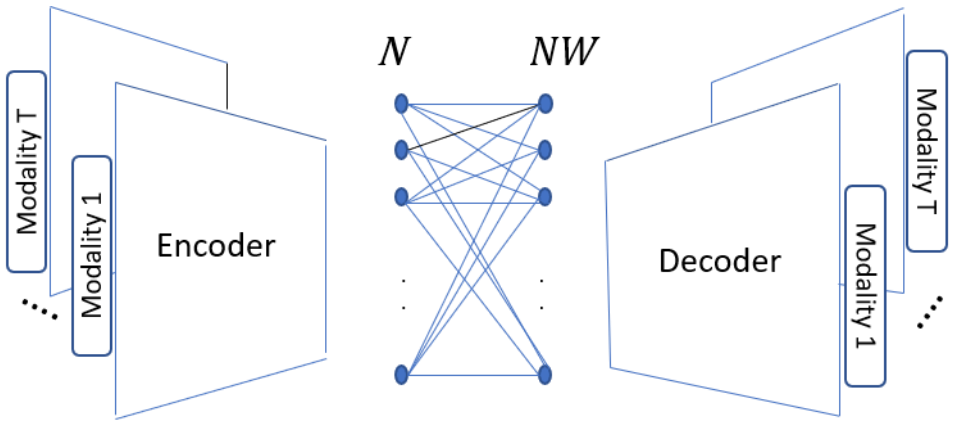


Figure 3.9 CNN Concatenation Network.

From the previous table, we can conclude that concatenating the features from the encoder and feeding the concatenated information to each decoder branch achieves a better performance for this type of multi-modal data structure. The reason behind this enhancement is the combination of efficient extraction of the basic features from the whole face and finer features from each part of the face. Promoting more efficiency as noted, this concatenation may also be intuitively viewed as adequate mosaicking, in which different patterns complement each other. In the following, we will show how our proposed approach performs in two cases: missing and noisy test data. The results of the new proposed approach, which we refer to as CNN concatenation network, is compared to the state-of-the-art DMSC network [34]. We start by training the auto-encoder network using 75% of the data and then we test on the rest of the data. In Figure (3.10), we show how the performance degrades by decreasing the number of available modalities at testing from five to one. From the results, it is clear how the CNN concatenation network outperforms the DMSC network. Additionally, we repeated the same experiment we performed in subsection 3.3.5. We train the network with noiseless data and then add Gaussian noise to one data modality at the testing. Additionally, we vary the number of available modalities at testing from one to four. The results are depicted in Figure (3.11). From the results, it is clear how the concatenated CNN network is more robust to noise than DMSC. In addition, we have utilized the Concatenation network to perform object clustering on the ARL data. We compare the clustering performance of the concatenation network with both DMSC and DRoGSuRe. The results are depicted in Table (3.8).

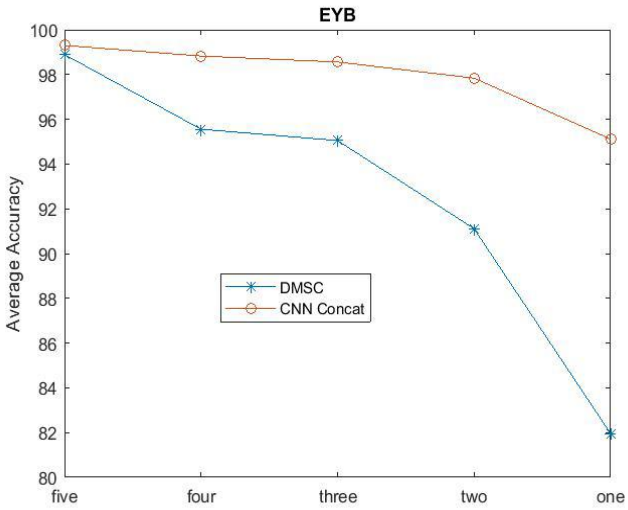
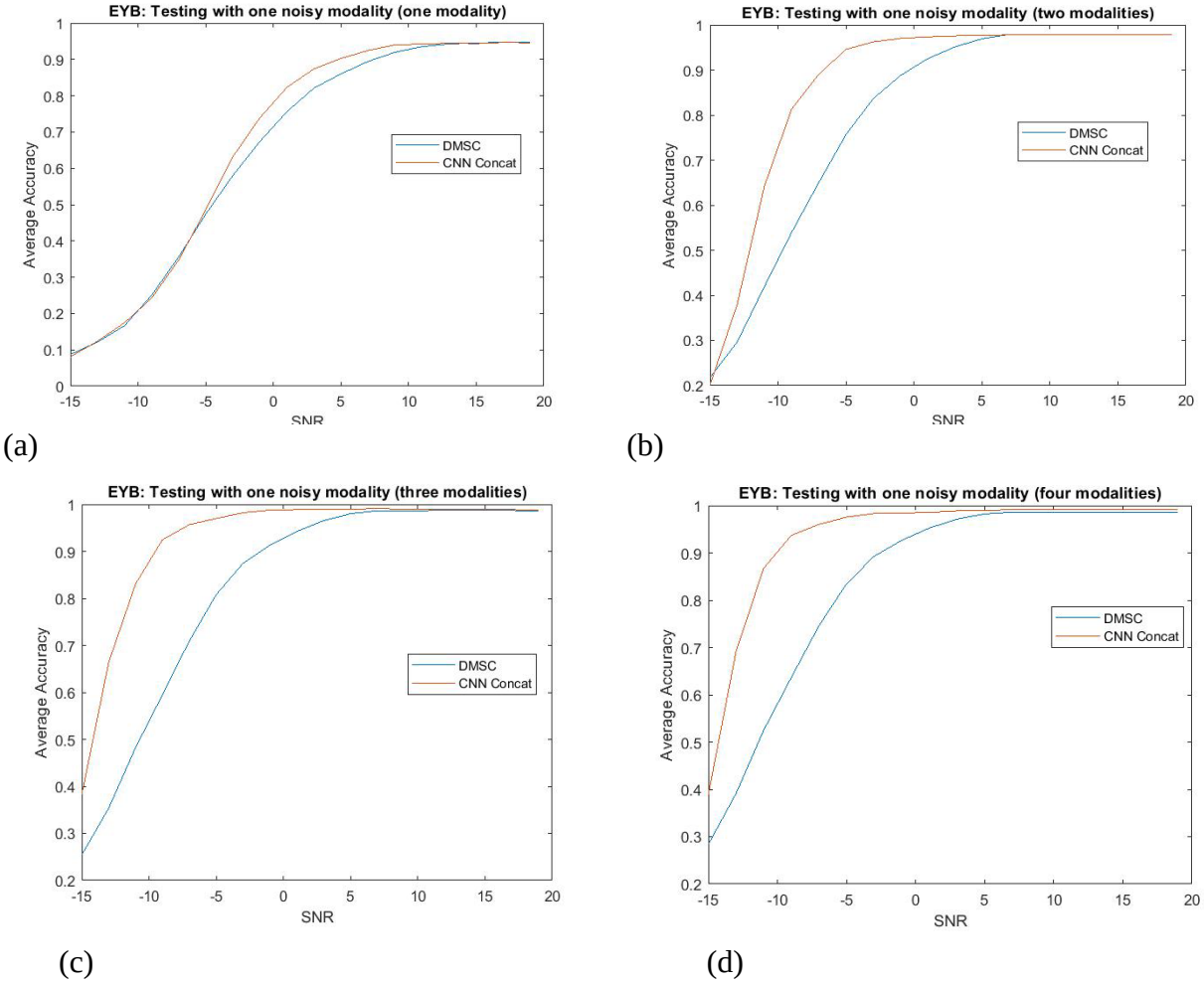


Figure 3.10 Missing Modalities during Testing for EYB Dataset.



(a) (b)

(c) (d)

Figure 3.11 EYB Noiseless Training and Validating on limited Noisy Data.

Table 3.8 Concatenation Performance for ARL Dataset.

	Learning	Validation
DMSC	97.59%	98.33%
DRoGSuRe	100%	100%
CNN Concat	99.44%	99.17

From the results, we conclude that DRoGSuRe still outperforms the other approaches for the ARL dataset. Although the number of parameters involved in training the DRoGSuRe network is higher than other approaches, since there are multiple self-expressive layers, however, DRoGSuRe is more robust to noise and limited data availability during testing.

3.6 Conclusion

In this chapter, we proposed a deep multimodal approach to fuse data through recovering the underlying subspaces of data observations from data corrupted by noise to scale to complex data scenarios. DRoGSuRe provides a natural way to fuse multimodal data by employing the self-representation matrix as an embedding for each data modality. Experimental results show a significant improvement for DRoGSuRe over DMSC under different types of potential limitations and provides robustness with limited sensing modalities. We also proposed the concatenated CNN network model, which can work better for different multi-modal data structures.

CHAPTER 4

Latent Code-Based Fusion: A Volterra Neural Network Approach

4.1 Introduction

Convolutional Neural Networks (CNNs) have widely been used in deep learning for analyzing visual images in many applications. However, the complexity and cost of implementing CNNs can be limiting in some applications. The efficient and recently proposed Volterra Neural Networks (VNNs) [24] was aimed to primarily address these limitations and also overcome the CNN over parametrization problem. To control the non-linearities intentionally induced in the network, a judicious setting of the degree of interactions between the delayed input samples of the data is induced. The cascaded implementation proposed in [24] has shown to significantly reduce the number of parameters needed for training the network in comparison to conventional neural networks. In addition to reducing the network complexity, Volterra Neural Networks (VNNs) have a more tractable and comprehensible structure. This represents a significant departure from the work in [25] and [26], as the strategy of convolution together with an understanding of the inherent complexity of a naive approach resulted in the success of this new outlook. These prior approaches had limited the degree of the non-linearities up to certain value to avoid the explosive complexity, but thereby defeating the initial goal of overcoming the limitations of existing CNNs. The cascaded implementation in [24] has been shown to alleviate this limitation through applying the second order filter repeatedly applied until the desired order is attained.

On account of recent advances in sensor technology, multimodal data have become widely available and useful. Additional modalities can grant additional integral information as compared

to unimodal data. A principled integration of multi-modal sensor data may indeed often boost the data reach of the structure and improve the quality of extracted features. Multi-modal fusion has been extensively used in different applications including but not limited to image fusion [15], target recognition [16], speaker recognition [17] and handwriting analysis [18]. In addition, CNNs have been extensively utilized for multimodal data analysis as in [62], [80] and [81]. However, the complexity of implementing multi-modal fusion network still persists. Inspired by the success of VNNs in deep networks [24], we propose an efficient implementation of the Deep Multi-modal Subspace clustering auto-encoder [34], using Volterra filters. More specifically, CNNs are replaced with VNNs which to more carefully control the introduced non-linearities via high order convolutions instead of using "blind" non-linear activation functions as carried out in architectures found in [34]. Moreover, we propose additional ways to significantly reduce the number of parameters needed to train the VNN auto-encoder, up to a fraction of the number of parameters used by CNNs while retaining a comparable clustering performance.

4.2 Volterra Multi-Modal Subspace Clustering (VMSC)

4.2.1 Problem Formulation

Similarly to the problem formulation presented in Section 3.2.1, we start by considering a set of data realizations indexed by $k = 1, 2, \ldots, n$. Furthermore, assume T data modalities, indexed by $t = 1, 2, 3, \ldots, T$. Each data realization can be represented as a m-dimensional vector $\boldsymbol{x_k}(t) \in \mathbb{R}^m$, where $\mathbf{X}(t) = [\boldsymbol{x_1}(t)\ \boldsymbol{x_2}(t) \ldots \boldsymbol{x_n}(t)]$. The goal is to partition a set of realizations into clusters whose respective measurements for each modality is well-represented by a low-dimensional subspace. Mathematically, this is tantamount to seeking a partitioning $\{\boldsymbol{X^1}(t), \boldsymbol{X^2}(t), \ldots, \boldsymbol{X}^P(t)\}$

of [n] observations, where P is the number of clusters indexed by p, such that there exist linear subspaces $S^p(t) \subset \mathbb{R}^m$ with dim ($S^p(t)$) $\ll$ m. Let $\boldsymbol{x}_k(t) \in S^p(t) \; \forall \; t \; and \; k \in X^p$.

4.2.2 Volterra Multi-Modal Subspace Clustering Auto-encoder (VMSC-AE)

In this section, we provide the fundamental concepts of our proposed approach which we refer to as Volterra Multimodal Subspace clustering auto-encoder (or VMSC). Our framework learns an efficient latent representation of multi-modal data whose features are jointly captured by a union of subspaces. Our approach was inspired by the Volterra series, which is a model for non-linear behavior similar to the Taylor series. Volterra series represents a functional expansion of nonlinear and time-invariant systems. It differs from the Taylor series in its ability to capture "memory" effects. In other words, The Taylor series can be used for approximating the response of a nonlinear system if the output of this system depends strictly on the input at that particular time. In contrast to Volterra series, whose output depends on the input to the system at all other times. Volterra filters (VF) are based on the Volterra series, VF describe a non-linear system via higher order convolutions. The relation between the output and input of the Volterra filter can be expressed as follows,

$$y_t = \sum_{\tau_1=0}^{L-1} \boldsymbol{H}^1_{\tau_1} x_{t-\tau_1} + \sum_{\tau_1,\, \tau_2=0}^{L-1} \boldsymbol{H}^2_{\tau_1,\tau_2} x_{t-\tau_1}\, x_{t-\tau_2} + \cdots + \sum_{\tau_1,\tau_2,\ldots,\tau_K=0}^{L-1} \boldsymbol{H}^K_{\tau_1,\tau_2,\ldots,\tau_K} x_{t-\tau_1} x_{t-\tau_2} \ldots x_{t-\tau_K}, \tag{4.1}$$

where L is the number of terms in the filter memory (also referred to as the filter length), H^K are the weights for the K^{th} order term. The linear term in the previous equation is actually similar to a convolutional layer in CNNs. Generally speaking, Nonlinearities in CNNs are often introduced by activation functions, and not in the convolutional layer, while VNNs introduce non-linearities

in the convolutional layers. Volterra Neural Network (VNNs) has been recently proposed to control the non-linearities introduced in the network and to overcome the CNN over parametrization problem. The VF architecture introduces controlled non-linearities in the Neural Networks in contrast to non-linear activation functions, which provide infinite non-linearity into the neural network which cannot be easily truncated.

Non-linear activation functions, such as Relu and Sigmoid, are often utilized in CNNs to act as a "gate" in between the input feeding the current neuron and its output going to the next layer. However, in some cases, the unnecessary nonlinearities introduced by those activation functions in CNN networks induce useless or irrelevant information to the network, which might confuse the classifier. In this Chapter, we propose a multi-modal autoencoder using the recently introduced Volterra Neural Networks (VNNs) [24] to seek a latent representation of multi-modal data whose features are jointly captured by a union of subspaces. More specifically, we replace the CNNs in our network with VNNs to control the introduced non-linearities and to non-linearly map the data points to a latent space that is well-adapted to subspace clustering in an unsupervised manner.

The Volterra Multi-Modal Subspace Clustering Auto-Encoder (VMSC-AE) exploits the self-expressive property presented in [1] and [2] to acquire the latent space structure that reveals the relationships between data points in each cluster. The self-representation property entails the representation of each sample as a linear combination of all other samples from the same subspace/ cluster. In the following, we elaborate on the structure of the Volterra structure-based multi-modal auto-encoder. As noted earlier for AFDMSC in Section 3.3 and similarly for VMSC-AE, the network has 3 main components, namely a multi-modal encoder, a self-expressive layer, and a

multimodal decoder. The encoder in this work replaces the standard CNNs with the Volterra Neural Network (VNN) developed in [24].

The multi-modal encoder consists of T parallel Volterra NNs. Each branch of the encoder processes one of the modalities and extracts relevant features. The T feature maps are subsequently concatenated promoting the goal of obtaining a common latent space.

The second component of the auto-encoder is the self-expressive layer, the goal of which is to enforce the self-expressive property [2] among the concatenated features. This is enforced by a fully connected layer which operates on the concatenated output of the encoder. The last stage is the decoder which reconstructs the input data from the self-expressive layers' output. The objective functional sought through this approximation network is reflected in Eqn. (4.2),

$$min_{W|w_{kk}=0}\|\boldsymbol{W}\|_1 + \frac{\gamma}{2}\sum_{t=1}^{T}\|\boldsymbol{X}(t) - \boldsymbol{X_r}(t)\|_F^2 + \frac{\mu}{2}\|\boldsymbol{L_{concat}} - \boldsymbol{L_{concat}}\boldsymbol{W}\|_F^2, \tag{4.2}$$

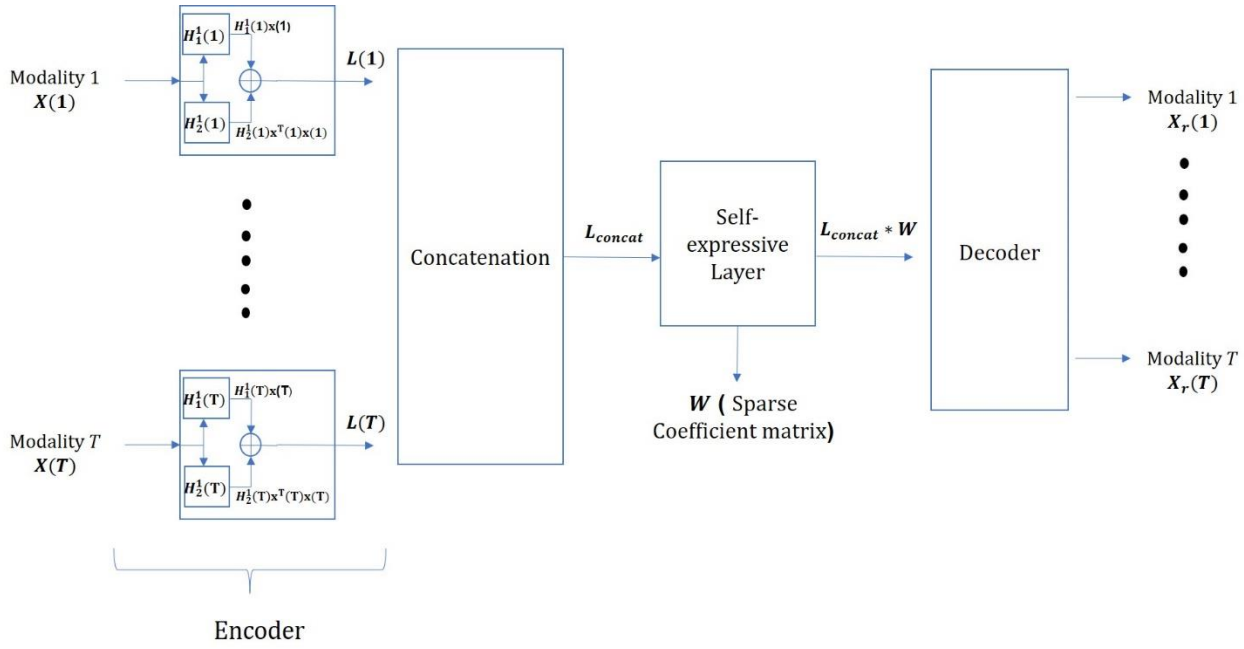


Figure 4.1 Volterra Multimodal Subspace Clustering Auto-Encoder.

where $\boldsymbol{X_r}(t)$ represents the reconstructed data corresponding to modality t, and $\boldsymbol{L_{concat}}$ is the concatenation of $\boldsymbol{L}(1), \boldsymbol{L}(2)\ \ldots, \boldsymbol{L}(T)$, where $\boldsymbol{L}(t)$ is the output of the encoder corresponding to modality t. $\boldsymbol{W}$ is the sparse weight function that ties the concatenated features. The above cost functional is optimized in Tensorflow using the adaptive momentum based gradient descent method (ADAM) [70]. Under a suitable arrangement/permutation of the data realizations, the sparse coefficient matrix $\boldsymbol{W}$ is an $n \times n$ block-diagonal matrix with zero diagonals provided that each sample is represented by other samples only from the same subspace. More precisely, W_{ij} = 0 whenever the indices i, j correspond to samples from different subspaces. As a result, the majority of the elements in $\boldsymbol{W}$ are equal to zero. $\| \|_1$ denotes the l_1 norm, i.e., the sum of absolute values of the argument. Upon computing the gradient of the loss function, the weights of each multi-layer network, that corresponds to one modality, are updated while other modalities' networks are fixed. In other words, after constructing the data during the forward pass, the loss

function determines the updates that back-propagate through each layer. The encoder of the first modality is updated, following which, the self-expressive layer of that modality gets updated and finally the decoder. A diagram showing our algorithm is depicted in Figure 4.1.

4.2.3 Class Partitioning

Similarly to the methodology explained in Section 3.2.2, we proceed with distinguishing the various classes in an unsupervised manner. First, we evaluate the affinity matrix as detailed in [48]. The affinity matrix is computed as,

$$\boldsymbol{A} = \boldsymbol{W} + \boldsymbol{W}^T, \tag{4.3}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. Briefly, a matrix $\boldsymbol{D}$ is defined to be a diagonal matrix whose i^{th} diagonal element is the degree of the i^{th} node, i.e., the sum of i^{th} row in **A**. The standard graph Laplacian matrix is next constructed as follows,

$$\boldsymbol{G} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}, \tag{4.4}$$

where $\boldsymbol{G} \in \mathbb{R}^{n \times n}$. The eigenvectors $\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_r}$ of $\boldsymbol{G}$ corresponding to the largest r eigenvalues are next computed, where r is the desired number of clusters. The matrix $\boldsymbol{E} = [\boldsymbol{e_1}, \boldsymbol{e_2}, \dots, \boldsymbol{e_r}]$ is then formed by stacking the eigenvectors in columns. Each row of $\boldsymbol{E}$ is a point in $\mathbb{R}^r$, *k*-means clustering is then used to cluster the rows of $\boldsymbol{E}$. Finally, the original point $\boldsymbol{x_i}$ is assigned to cluster *j* iff row *i* of the matrix $\boldsymbol{E}$ was assigned to cluster *j*.

4.3 Experimental Results

4.3.1 Dataset Description

To substantiate the discussed approach along with the various steps, we select two different datasets. The same datasets were utilized in Chapter 3. The first dataset is the Extended Yale Dataset [76]. This dataset has been used extensively in subspace clustering as in [2], [44]. The dataset is composed of 64 frontal images of 38 individuals under different illumination conditions. In this work, we will use the augmented data used in [34], where facial components such as left eye, right eye, nose and mouth have been cropped to represent four additional modalities which are spatially unrelated. Images corresponding to each modality have been cropped to a size of 32×32.

The second validation dataset we use is the ARL polarimetric face dataset [78]. This consists of facial images for 60 individuals in the visible domain and in four different polarimetric states. The dataset was collected using a polarimetric long-wave infrared imager, to facilitate cross-spectrum face recognition research. Different polarization states of thermal emissions can provide additional geometric and textural facial details, which can be used to improve face identification. The Stokes parameters S0, S1, S2, and S3 are often used to represent polarization-state information. They are collected by measuring the radiant intensity transmitted through a polarizer that rotates at different angles. S0 represents the conventional total intensity thermal image, S1 captures the horizontal and vertical polarimetric information, and S2 captures the diagonal polarimetric information. S1 and S2 capture orthogonal, yet complementary, polarimetric information. The degree of linear polarization (DoLP) describes the portion of an electromagnetic wave that is linearly polarized.

All the images are spatially aligned for each subject. We have also resized the images to 32×32 pixels.

4.3.2 Network Architecture

In the following, we will elaborate on how we construct the VNN for each dataset. For both datasets, the data corresponding to each modality goes into the corresponding encoder. The encoder projects the input modality into the feature space. Features are extracted from each modality independently and are subsequently concatenated before going through the self-expressive layer. The input to the self-expressive represents the data modalities' projection into the latent space. The second component of the Volterra filter auto-encoder is the self-expressive layer. The goal of this layer is to enforce the self-representation property among the features extracted from each data modality, utilizing a fully connected layer which operates on the merged features. The last stage is the decoder which reconstructs input data from the self-expressive layers' output and has the same structure as the encoder. We implemented the Volterra Multi-Modal Subspace Clustering autoencoder with Tensorflow and used the adaptive momentum based gradient descent method (ADAM) [70] to minimize the loss function in Eqn. (4.1) with a learning rate of 10^{-3} for ARL dataset and 10^{-4} for EYB dataset. For DMSC, we used the same network structure they used in [34]. In the following, we will elaborate on the structure we constructed for the VMSC applied on ARL and EYB dataset.

4.3.2.1 ARL Dataset

The ARL dataset consists of five data modalities, therefore, the auto-encoder has five different encoders, one self-expressive layers, and five decoders. Each encoder is composed of a 2^{nd} order Volterra Filter. The Volterra filter consists of 3 filters of kernel size 1, and 2 filters of kernel size 3. The decoder has the same structure as the encoder.

4.3.2.2 EYB Dataset

For EYB dataset, we use five data modalities, therefore, we have an encoder for each modality, one self-expressive layers and five decoders. Each encoder is composed of a 2^{nd} order Volterra Filter. The Volterra filter consists of 3 filters of kernel size 1, 3 filters of kernel size 3, and 4 filters of kernel size 5.

4.3.2.3 Fusion Results

We evaluate the performance of our proposed VMSC against the convolutional auto-encoder DMSC network. We divide each dataset into learning and validation sets. For both datasets, we train each auto-encoder using only 75% of the data, following which, we test on 25% of the data. The Union of Subspace structure learned during training is then utilized to classify new observed data points in the test set. The sparse solution $\boldsymbol{W}$ provides important information about the relations among data points, which may be used to split data into individual clusters residing in a common subspace. Observations from each object can be seen as data points spanning one subspace. Interpreting the subspace-based affinities based on $\boldsymbol{W}$, we proceed to carry out what amounts to modality fusion. For clustering by $\boldsymbol{W}$, we applied the same spectral clustering approach that we previously demonstrated in Section 4.3.3.

We compare the performance of VMSC against DMSC using the clustering accuracy accuracy, normalized mutual information (NMI) [84], and Adjusted Rand Index (ARI) [85] metrics and the results are depicted in Tables (4.1) and (4.2) for EYB and ARL dataset respectively. From the results, it can be seen that the VMSC auto-encoder outperforms the DMSC network, all the while reducing the number of parameters needed to carry out the clustering task. The reason behind this improvement is the fact that the Volterra-based network maintains a tractable structure that controls the non-linearities introduced in the system in contrast to the CNN network that can introduce undesirably infinite non-linearities.

Table 4.1 Fusion Results for EYB Dataset.

	Accuracy	NMI	ARI	No. of parameters
DMSC	98.82%	98.81%	98.08%	2,367,400
VMSC-AE	99.34%	99.15%	98.63%	2,332,800

Table 4.2 Fusion Results for ARL Dataset.

	Accuracy	NMI	ARI	No. of parameters
DMSC	97.59%	99.42%	97.53%	4,667,720
VMSC-AE	99.95%	99.94%	99.90%	4,666,650

4.3.3 Training with Less Data

We proceed to evaluate the performance of the proposed Volterra filter auto-encoder with limited training data. A major challenge for any deep neural network may be the availability of inadequately sufficient data to train the network. In the following, we will assess our proposed data fusion network versus the convolutional deep neural network, DMSC, in case of limited data availability during training. We train the auto-encoder structure using portions of the available data, i.e., 25%, 40%, 50%, 60%, and 75%. The Results are depicted in Table (4.3) and (4.4) for ARL and EYB dataset, respectively. From the results, it is clear that fusing the data using Volterra Neural Networks autoencoder significantly boosts the clustering accuracy while using less parameters than DMSC. In addition, the Volterra filter autoencoder is more robust and less sensitive to limited data availability during training.

Table 4.3 ARL Dataset: Training with Less Data

Data Ratio	VMSC			No. Parameters	DMSC			No. Parameters
	ACC	NMI	ARI		ACC	NMI	ARI	
25%	99.32%	99.72%	98.19%	519,450	93.33%	97.86%	88.25%	520,520
40%	99.42%	99.78%	99.49%	1,328,154	94%	98.5%	91.98%	1,329,224
50%	99.56%	99.86%	99.63%	2,074,650	94.17%	98.6%	92.7%	2,075,720
60%	99.9%	99.94%	99.88%	2,987,034	95.69%	98.64%	93.87%	2,988,104
75%	99.95%	99.95%	99.9%	4,666,650	97.59%	99.42%	97.53%	4,667,720

Table 4.4 EYB Dataset: Training with Less Data

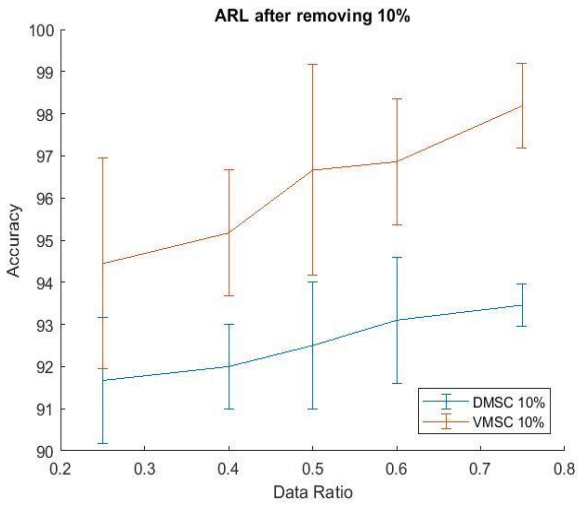
Data Ratio	VMSC			No. Parameters	DMSC			No. Parameters
	ACC	NMI	ARI		ACC	NMI	ARI	
25%	95.58%	96.8%	94.42%	333,784	93.03%	92.34%	86.83%	368,384
40%	97.25%	97.9%	96.81%	854,144	93.42%	96.34%	93.21%	888,744
50%	98.4%	98.26%	96.9%	1,322,000	97.34%	97.41%	95.05%	1,356,600
60%	98.96%	98.97%	97.99%	1,893,824	98.72%	98.47%	97.18%	1,928,424
75%	99.34%	99.15%	98.63%	2,332,800	98.82%	98.81%	98.08%	2,367,400

4.3.4 Network Pruning by Random Removal of Edges

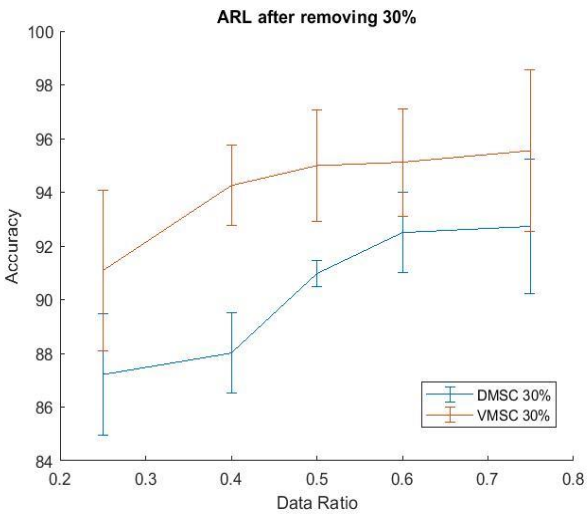
In this section, we introduce a solution to reducing the number of parameters required in training the auto-encoder network. It is clear from Table (4.1) and (4.2) that the total number of parameters needed is dominated by the self-expressive layer parameters, which is a function of O(N^2), where

N is the number of samples in the dataset. As a result, this may lead to a longer training time and require a lot of computational resources. In the following, we will attempt to reduce the number of parameters needed by randomly removing a ratio of the edges in the self-expressive layer and train the network with the remaining edges. By eliminating the appropriate number of edges and setting them equal to zero, the clustering performance should not be highly affected. The reason is the fact that the self-representation coefficient matrix $\boldsymbol{W}$ should be sparse with a block-diagonal structure. Therefore, most of the edges will eventually be equal to zero. In addition to training the network with less data, we have reduced the number of edges that needed to be trained by setting a fixed ratio of those edges to be equal zero and ignore them while training as if they do not exist. In Figures (4.2) and (4.3)**,** we illustrate the error bar after removing 10%, 30%, 50% and 70% of the edges from the self-expressive layer for the ARL and EYB datasets, respectively.

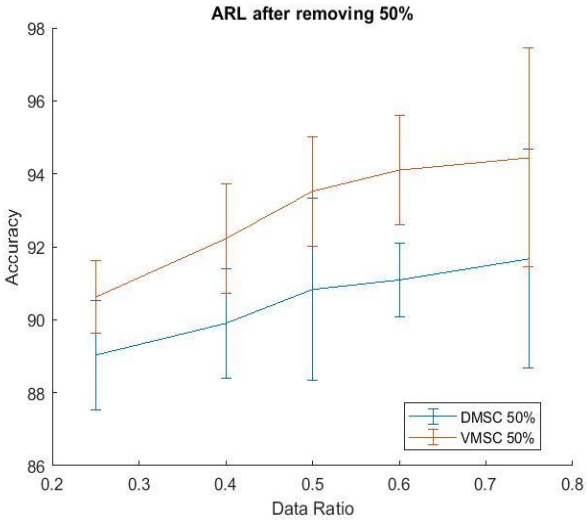
The results have been averaged over 10 trials. From the results, we can conclude that the Volterra-Based Network, VMSC, is more robust to the changes in the self-expressive layer connections as compared to DMSC. As we remove edges from the self-expressive layer, the performance degradation in case of VMSC is more graceful. In addition, the Volterra Filter auto-encoder is less sensitive to training with less data as compared to DMSC network, because of the lower number of parameters in the encoder and decoder which also prevents overfitting when lower number of samples are available.



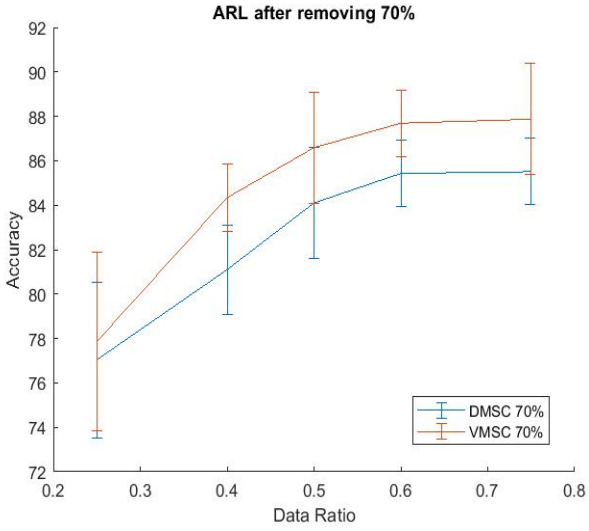
(a)



(b)

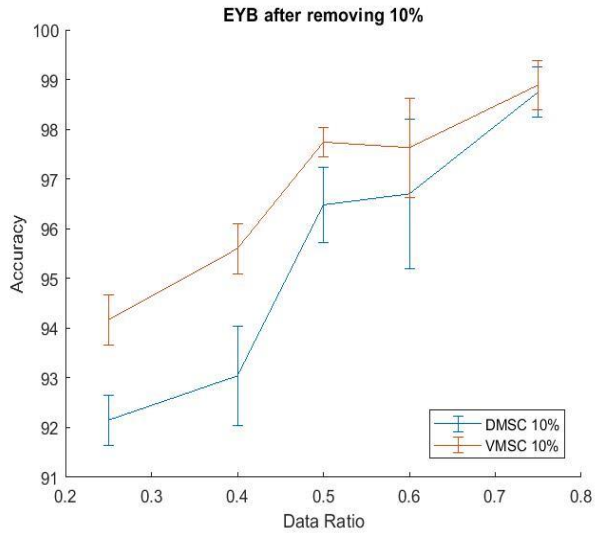


(c)

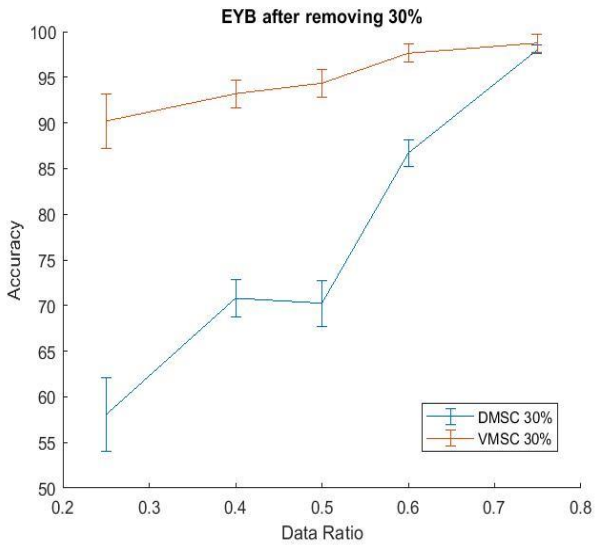


(d)

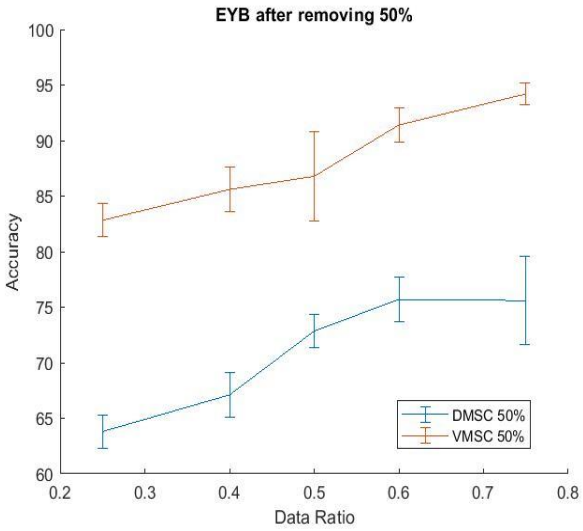
Figure 4.2 Pruning the Auto-encoder Network while Using Different Portions of the ARL Dataset.



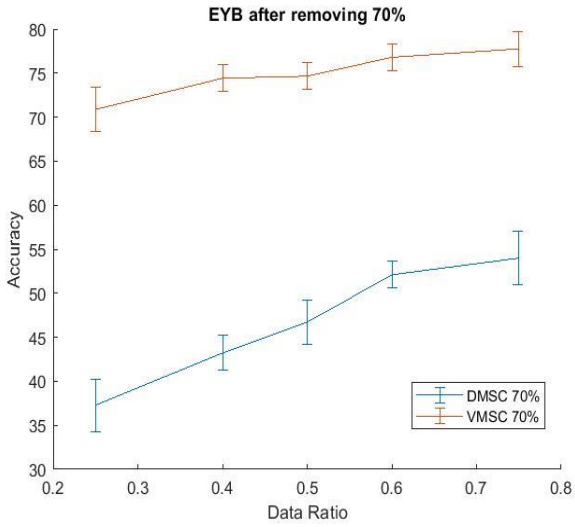
(a)



(b)



(c)



(d)

Figure 4.3 Pruning the Auto-encoder Network while Using Different Portions of the EYB Dataset.

4.3.5 Network Pruning Using Cyclic Sparse Connected Layers

The second approach we evaluate, is referred to as cyclic sparsely connected (CSC) layers [83]. This approach was employed as an overlay for fully connected (FC) layers whose number of parameters, $O(N^2)$, can dominate the parameters of the entire deep neural network model. The CSC layers are composed of a few sequential layers, referred to as support layers, which result in full connectivity between the Inputs and Outputs of each CSC layer. Models trained with CSC layers follow a bottom-up approach by incrementally increasing the parameters such as connectivity and number of synapses, to achieve a desired accuracy. The sparsely connected layers are composed of a sequence of L layers: an Input layer, $L-1$ layers in between referred to as support layers, and an output layer. All layers are connected via edges (synapses). Assume having N nodes and the fan-out of every node, as well as the fan-in of every layer, is equal to F. Therefore, the total number of edges, E, in the CSC layers' structure is as follows,

$$E = NFL\,, \tag{4.5}$$

Every layer in the CSC structure is defined by an adjacency matrix, $\boldsymbol{A_i}$, $s.t.\, i = 1, \dots, L$, whose length and width are equal to N, where $\boldsymbol{A}(i,j)$ indicates the number of edges that connect the input node i to the output node j. As a result, NF out of N^2 elements of each adjacency matrix corresponding to one support layer are equal to 1 and the rest of the elements are zeros. The connectivity, C, is defined as the number of paths between every pair of nodes chosen arbitrarily from the diagram Input and Output layers. Therefore,

$$F^L = NC\,, \tag{4.6}$$

From Eqns. (4.5) and (4.6), it can be deduced that the complexity decreased from $O(N^2)$ to $O(NlogN)$. The relationship between the number of edges and the number of nodes is as follows,

$$E = NFlog_F(NC)\,, \tag{4.7}$$

An adjacency matrix, $\boldsymbol{A_i}$, is defined for every support layer i in the CSC layers we use to replace the fully connected layer. Every support layer has a generator polynomial $p_i(x)$, which is composed of *F* terms to generate a cyclic adjacency matrix of block length *N* as explained in [83]. The generator polynomial corresponding to each support cyclic matrix constructs the first row. Every next row of the matrix is a cyclic right shift of its previous row. In [83], two different factorization approaches were proposed for constructing the support layers. For our problem, we assume that the connectivity, *C*, is equal to 1. We follow the first approach proposed in [83] to construct the generator polynomial $p_T(\mathrm{x}) = \sum_{i=0}^{N-1} x^i$ as follows,

$$\begin{cases} \sum_{i=0}^{N-1} x^i = \prod_{i=0}^{L-1} p_i(x) \\ p_i(x) = \sum_{j=0}^{F-1} x^{S_i \cdot j}\,, \qquad S_i = F^i \end{cases}, \tag{4.8}$$

The CSC layers are described by the polynomial function by assigning each $p_i(x)$ to each support layer i. S_i is the stride value which specifies the distance between elements of value 1s in the first row of the support matrix of layer i. In the following, we evaluate the CSC layers to further prune the auto-encoder network. In our experiment, we assume that $C = 1$ and given N, we compute the number of edges in each support layer accordingly using Eqn. (4.5). For the ARL and EYB dataset, we assume that C = 1 and *L* = 2. We found out that this structure achieves the best performance while retaining a high compression rate. The experimental results are depicted in Table (4.5) and (4.6) for ARL and EYB datasets respectively. We utilized different data ratios for training the autoencoder network and subsequently compare the results from the VMSC-AE network to the DMSC network. In addition, we list the number of parameters required for training the auto-

encoder with the fully connected self-expressive layer versus the number of parameters needed for training the auto-encoder utilizing the CSC layers. The results are also illustrated in Figures 4.4 and 4.5 for ARL and EYB datasets respectively.

Table 4.5 ARL Dataset: Impact of CSC Layers when Training with Less Data.

	Fully connected			No. of parameters	After Pruning			No. of parameters
	ACC	NMI	ARI		ACC	NMI	ARI	
DMSC 25%	93.33%	97.86%	88.25%	520,520	92.34%	97.54%	85.1%	56,840
VMSC 25%	99.32%	99.72%	98.19%	519,450	99.03%	99.16%	97.19%	55,770
DMSC 40%	94%	98.5%	91.98%	1,329,224	93.2%	98.13%	90.24%	117,320
VMSC 40%	99.42%	99.78%	99.49%	1,328,154	99.28%	99.27%	97.67%	116,250
DMSC 50%	94.17%	98.6%	92.7%	2,075,720	93.67%	98.22%	92.05%	174,920
VMSC 50%	99.56%	99.86%	99.63%	2,074,650	99.48%	99.55%	99.01%	153,690
DMSC 60%	95.69%	98.64%	93.87%	2,988,104	95.66%	98.4%	92.89%	203,960
VMSC 60%	99.9%	99.94%	99.88%	2,987,034	99.77%	99.71%	99.53%	202,890
DMSC 75%	97.59%	99.42%	97.53%	4,667,720	97.55%	99.47%	97.81%	282,920
VMSC 75%	99.95%	99.95%	99.9%	4,666,650	99.91%	99.93%	99.8%	281,850

Table 4.6 EYB Dataset: Impact of CSC Layers when Training with Less Data.

	Fully connected			No. of parameters	After Pruning			No. of parameters
	ACC	NMI	ARI		ACC	NMI	ARI	
DMSC 25%	93.03%	92.34%	86.83%	368,384	92.75%	91.74%	82.25%	84,900
VMSC 25%	95.58%	96.8%	94.42%	333,784	95.14%	95.92%	91.85%	50,300
DMSC 40%	93.42%	96.34%	93.21%	888,744	93.01%	93.41%	86.43%	98,040
VMSC 40%	97.25%	97.9%	96.81%	854,144	96.33%	97.43%	96.12%	63,440
DMSC 50%	97.34%	97.41%	95.05%	1,356,600	94.89%	94.78%	88.43%	114,000
VMSC 50%	98.4%	98.26%	96.9%	1,322,000	98.33%	98.22%	96.71%	79,400
DMSC 60%	98.72%	98.47%	97.18%	1,928,424	95.51%	95.72%	91.59%	199,272
VMSC 60%	98.96%	98.97%	97.99%	1,893,824	98.68%	98.5%	97.42%	131,840
DMSC 75%	98.82%	98.81%	98.08%	2,367,400	96.29%	98.32%	96.93%	224,200
VMSC 75%	99.34%	99.15%	98.63%	2,332,800	99.07%	98.81%	98.09%	174,400

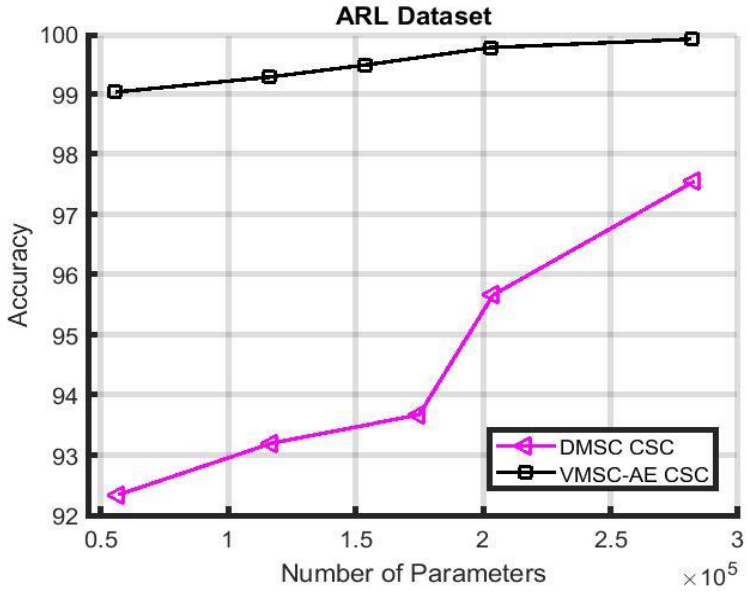


Figure 4.5 Performance vs Number of Parameters of VFSC CSC, and DMSC CSC for ARL dataset.

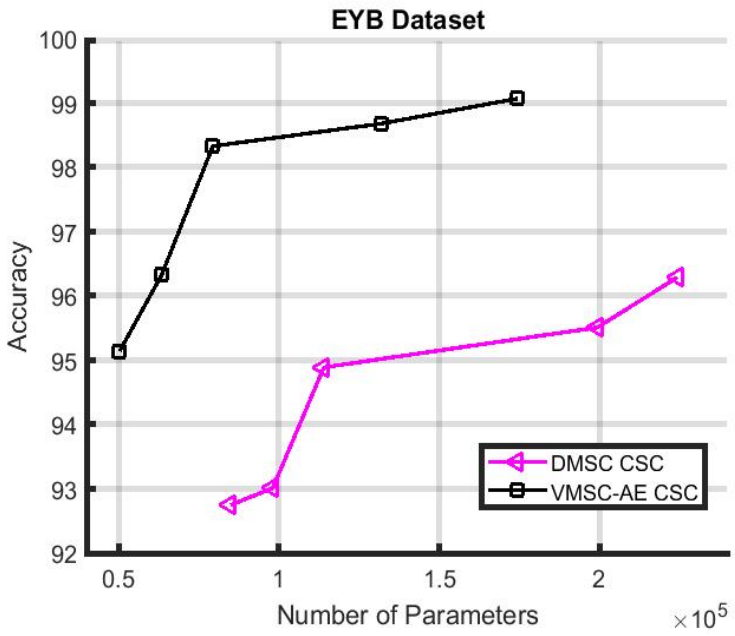


Figure 4.6 Performance vs Number of Parameters of VFSC CSC, and DMSC CSC for EYB dataset.

4.4 Conclusion

In this Chapter, we presented an efficient Volterra Neural Network auto-encoder for multi-modal data fusion. The introduced framework extracted the underlying embedding of each data modality under the assumption of data self-representation. Experimental results show a significant improvement for Volterra Neural Network over the Convolutional Neural Network auto-encoder. In addition, we evaluated multiple approaches to further prune the network structure and reduce the model complexity of the multi-modal subspace clustering auto encoder method. The experimental results showed that our proposed approach provides better sample complexity over CNN-based auto-encoder and demonstrates a robust classification performance.

CHAPTER 5

Conclusion

In this dissertation, we explored multiple approaches for multi-modal data fusion. In Chapter 2, a principled and robust framework has been introduced to unfold the underlying Unions of Subspaces (UoS) structure present each data modality. The recovered subspaces can identify more complex trends in data sets corrupted by noise. Referred to as Robust Group Subspace Recovery, the introduced framework learns a new joint representation of the data from different sources. The goal of this process is to exploit the structural dependencies between the different modalities data to cluster the associated target objects. Experimental results show a significant improvement for RoGSuRe over other state of the art subspace clustering techniques.

In Chapter 3, the deep structure of multi-modal data has been exploited to robustly utilize the group subspace distribution of the information using the Convolutional Neural Network (CNN) formalism. Referred to as deep Multimodal Robust Group Subspace Clustering (DRoGSuRe), this approach provides a natural way to fuse multimodal data by employing the self-representation matrix as an embedding for each data modality. The set of subspaces constituting each data modality is unfolded through learning their corresponding encoders. An optimized integration of the generated inherent information is subsequently carried out to yield a characterization of various classes. Experimental results show a significant improvement for DRoGSuRe over the independently developed state-of-the-art approach named Deep Multimodal Subspace Clustering (DMSC). DRoGSuRe provides robustness under different types of potential limitations and is

more competitive with limited sensing modalities as compared to the state-of-the art deep structure.

In Chapter 4, a deep multi-modal auto-encoder network has been proposed to fuse data, to recover the underlying subspaces of data observations, and help scaling the data fusion approach. The introduced framework fuses multi-modal data by employing the self-representation matrix as an embedding for each data modality. The Volterra Filter architecture leads to a reduction in the required number of parameters on account of controlled nonlinearities being introduced by the higher order convolutions as opposed to activation functions. Experimental results on two different datasets have shown a significant improvement in the clustering performance for Volterra Multimodal Subspace Clustering Autoencoder (VMSC-AE) over conventional convolutional neural network auto-encoder, DMSC. In addition, we have been able to show that the Volterra Neural Network auto-encoder is less sensitive to small data training and network pruning in comparison to Convolutional Neural Network auto-encoder.

REFERENCES

[1] X. Bian, A. Panahi, and H. Krim, “Bi-sparsity pursuit: A paradigm for robust subspace recovery,” Signal Processing, vol. 152, pp. 148–159, 2018.

[2] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 11, pp. 2765–2781, 2013.

[3] P. Favaro, R. Vidal, and A. Ravichandran, “A closed form solution to robust subspace estimation and clustering,” in CVPR 2011. IEEE, 2011, pp. 1801–1807.

[4] C.-G. Li and R. Vidal, “Structured sparse subspace clustering: A unified optimization framework,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 277–286.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” J. ACM, vol. 58, no. 3, p. 11, May 2011.

[6] M. Elad, M. A. T. Figueiredo, and Y. Ma, “On the role of sparse and redundant representations in image processing,” Proc. IEEE, vol. 98, no. 6, pp. 972–982, Jun. 2010.

[7] V. M. Patel and R. Vidal, “Kernel sparse subspace clustering,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2014, pp. 2849–2853.

[8] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi, “Deep subspace clustering with sparsity prior,” in Proc. IJCAI, 2016, pp. 1925–1931.

[9] A. Y. Yang, S. R. Rao, and Y. Ma, “Robust statistical estimation and segmentation of multiple subspaces,” in Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW), May 2006, p. 99.

[10] W. Hong, J. Wright, K. Huang, and Y. Ma, “Multiscale hybrid linear models for lossy image representation,” IEEE Trans. Image Process., vol. 15, no. 12, pp. 3655–3671, Dec. 2006.

[11] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, “An algebraic geometric approach to the identification of a class of linear hybrid systems,” in Proc. 42nd IEEE Int. Conf. Decision Control, vol. 1, May 2003, pp. 167–172.

[12] W. Tang, A. Panahi, H. Krim, and L. Dai, “Analysis dictionary learning based classification: Structure for robustness,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 6035–6046, Dec. 2019.

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

[14] R. K. Sharma, “Probabilistic model-based multisensor image fusion,” Ph.D. dissertation, Dept. Elect. Comput. Eng., Oregon Graduate Inst. Sci. Technol., Hillsboro, OR, USA, Oct. 1999.

[15] O. Hellwich and C. Wiedemann, “Object extraction from high-resolution multisensor image data,” in Proc. 3rd Int. Conf. Fusion Earth Data, vol. 115, 2000, pp. 1–10.

[16] Z. Korona and M. M. Kokar, “Model theory based fusion framework with application to multisensor target recognition,” in Proc. IEEE/SICE/RSJ Int. Conf. Multisensor Fusion Integr. for Intell. Syst., Oct. 1996, pp. 9–16.

[17] F. K. Soong and A. E. Rosenberg, “On the use of instantaneous and transitional spectral information in speaker recognition,” IEEE Trans. Acoust., Speech, Signal Process., vol. 36, no. 6, pp. 871–879, Jun. 1988.

[18] L. Xu, A. Krzyzak, and C. Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” IEEE Trans. Syst., Man, Cybern., vol. 22, no. 3, pp. 418–435, 1992.

[19] L. Li, Z.-Q. Luo, K. M. Wong, and E. Bosse, “Convex optimization approach to identify fusion for multisensor target tracking,” IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, vol. 31, no. 3, pp. 172–178, May 2001.

[20] M. C. Florea and E. Bosse, “Critiques on some combination rules for probability theory based on optimization techniques,” in Proc. 10th Int. Conf. Inf. Fusion, Jul. 2007, pp. 1–8.

[21] L. Li, “Data fusion and filtering for target tracking and identification,” Ph.D. dissertation, Dept. Elect. Comput. Eng., MCMASTER Univ., Hamilton, ON, Canada, 2003.

[22] D. L. Hall and J. Llinas, “An introduction to multisensor data fusion,” Proc. IEEE, vol. 85, no. 1, pp. 6–23, Jan. 1997.

[23] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, “Multisensor data fusion: A review of the state-of-the-art,” Inf. Fusion, vol. 14, no. 1, pp. 28–44, Jan. 2013.

[23] V. Volterra, “Theory of functionals and of integral and integrodifferential equations,” 1959.

[24] S. Roheda and H. Krim, “Conquering the cnn over-parameterization dilemma: A Volterra filtering approach for action recognition,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 11 948–11 956.

[25] G. Zoumpourlis, A. Doumanoglou, N. Vretos, and P. Daras, “Non-linear convolution filters for cnn-based learning,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4761–4769.

[26] R. Kumar, A. Banerjee, B. C. Vemuri, and H. Pfister, “Trainable convolution filters and their application to face recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 7, pp. 1423–1436, 2011.

[27] Y. A. Kathawala and B. Tueck, “The use of RFID for traffic management,” Int. J. Technol., policy Manage., vol. 8, no. 2, pp. 111–125, 2008.

[28] R. O. Sanchez, C. Flores, R. Horowitz, R. Rajagopal, and P. Varaiya, “Vehicle re identification using wireless magnetic sensors: Algorithm revision, modifications and performance analysis,” in Proc. IEEE Int. Conf. Veh. Electron. Saf., Jul. 2011, pp. 226–231.

[29] A. Haoui, R. Kavaler, and P. Varaiya, “Wireless magnetic sensors for traffic surveillance,” Transp. Res. C, Emerg. Technol., vol. 16, no. 3, pp. 294–306, Jun. 2008.

[30] C. T. Christou and G. M. Jacyna, “Vehicle detection and localization using unattended ground magnetometer sensors,” in Proc. 13th Int. Conf. Inf. Fusion, Jul. 2010, pp. 1–8.

[31] H. Krim and M. Viberg, “Two decades of array signal processing research: The parametric approach,” IEEE Signal Process. Mag., vol. 13, no. 4, pp. 67–94, Jul. 1996.

[32] Y. Ding, B. Banitalebi, T. Miyaki, and M. Beigl, “RFTraffic: A study of passive traffic awareness using emitted RF noise from the vehicles,” EURASIP J. Wireless Commun. Netw., vol. 2012, no. 1, Dec. 2012, Art. no. 8.

[33] S. Ghanem, A. Panahi, H. Krim, R. A. Kerekes, and J. Mattingly, “Information subspace-based fusion for vehicle classification,” in Proc. 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018, pp. 1612–1616.

[34] M. Abavisani and V. M. Patel, “Deep multimodal subspace clustering networks,” IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 6, pp. 1601–1614, 2018.

[35] R. Rubinstein, T. Faktor, and M. Elad. ” K-SVD dictionary-learning for the analysis sparse model.” In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5405-5408. IEEE, 2012.

[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong.” Locality constrained linear coding for image classification.” In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3360- 3367. IEEE, 2010.

[37] Z. Zhang, M. Zhao, and T. Chow.” Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification.” IEEE Transactions on Knowledge and Data Engineering 25, no. 10, pp. 2192-2205, 2013.

[38] V. Khorani, F. Razavi, and V. R. Disfani.” A mathematical model for urban traffic and traffic optimization using a developed ICA technique.” IEEE Transactions on intelligent transportation systems 12, no. 4, pp. 1024-1036, 2011.

[39] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C.Van Lint. ” Prediction intervals to account for uncertainties in travel time prediction.” IEEE Transactions on Intelligent Transportation Systems, no. 2, pp. 537-547, 2011.

[40] C. Anagnostopoulos, T. Alexandropoulos, V. Loumos, and E. Kayafas. ” Intelligent traffic management through MPEG-7 vehicle flow surveillance.” In IEEE International Symposium on Modern Computing, pp. 202-207. IEEE, 2006.

[41] Y. A. Kathawala, and B. Tueck.” The use of RFID for traffic management.” International journal of technology, policy and management 8, no. 2, pp. 111-125, 2008.

[42] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya.” Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors.” Transportation Research Part C: Emerging Technologies 17, no. 6, pp. 586-606, 2009.

[43] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, “Multi-task low-rank affinity pursuit for image segmentation,” in Proc. Int. Conf. Comput. Vis., May 2011, pp. 2439–2446.

[44] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 171–184, Jan. 2013.

[45] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao, “Low-rank tensor constrained multiview subspace clustering,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1582–1590.

[46] M. Abavisani and V. M. Patel, “Multimodal sparse and low-rank subspace clustering,” Inf. Fusion, vol. 39, pp. 168–177, Jan. 2018.

[47] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” in Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 612–620.

[48] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in Proc. Adv. Neural Inf. Process. Syst., 2002, pp. 849–856.

[49] D. Taylor, S. Shai, N. Stanley, and P. J. Mucha, “Enhanced detectability of community structure in multilayer networks through layer aggregation,” Phys. Rev. Lett., vol. 116, no. 22, Jun. 2016, Art. no. 228301.

[50] M. Soltanolkotabi and E. J. Candés, “A geometric analysis of subspace clustering with outliers,” Ann. Statist., vol. 40, no. 4, pp. 2195–2238, Aug. 2012.

[51] B. Malhotra, I. Nikolaidis, and J. Harms, “Distributed classification of acoustic targets in wireless audio-sensor networks,” Comput. Netw., vol. 52, no. 13, pp. 2582–2593, Sep. 2008.

[52] J. Kell, I. Fullerton, and M. Mills, “Traffic detector handbook,” Federal highway administration, Washington, DC, USA, Tech. Rep FHWA HRT- 06-108, Oct. 2006.

[53] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” in Readings Speech Recognition. Amsterdam, The Netherlands: Elsevier, 1990, pp. 65–74.

[54] S. Charbonnier, A.-C. Pitton, and A. Vassilev, “Vehicle re-identification with a single magnetic sensor,” in Proc. IEEE Int. Instrum. Meas. Technol. Conf. Process., May 2012, pp. 380–385.

[55] I. T. Jolliffe, “A note on the use of principal components in regression,” Appl. Statist., vol. 10, pp. 300–303, Oct. 1982.

[56] E. W. Forgy, “Cluster analysis of multivariate data: Efficiency versus interpretability of classifications,” Biometrics, vol. 21, no. 3, pp. 768–769, 1965.

[57] Z. Zivkovic, “Improved adaptive Gaussian mixture model for background subtraction,” in Proc. 17th Int. Conf. Pattern Recognit. (ICPR), 2004, pp. 28–31.

[58] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal processing, vol. 10, nos. 1–3, pp. 19–41, 2000.

[59] L. Rokach and O. Maimon, “Clustering methods,” in Data Mining Knowledge Discovery Handbook. Boston, MA, USA: Springer, 2005, pp. 321–352.

[60] R. Brandenberg, A. Dattasharma, P. Gritzmann, and D. Larman, “Isoradial bodies,” Discrete Comput. Geometry, vol. 32, no. 4, pp. 447–457, Nov. 2004.

[61] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, “Clustering appearances of objects under varying illumination conditions,” in 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., vol. 1. IEEE, 2003, pp. I–I.

[62] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, “Multimodal deep learning (pp. 689–696),” in International conference on machine learning (ICML), Bellevue, WA, 2011.

[63] D. Ramachandram and G. W. Taylor, “Deep multimodal learning: A survey on recent advances and trends,” IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 96–108, 2017.

[64] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, “Deep multispectral semantic scene understanding of forested environments using multimodal fusion,” in International Symposium on Experimental Robotics. Springer, 2016, pp. 465–477.

[65] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, “Deep subspace clustering networks,” in Advances in Neural Information Processing Systems, 2017, pp. 24–33.

[66] S. Ghanem, A. Panahi, H. Krim, and R. A. Kerekes, “Robust group subspace recovery: A new approach for multi-modality data fusion,” IEEE Sensors Journal, 2020.

[67] S. Roheda, H. Krim, and B. S. Riggan, “Robust multi-modal sensor fusion: An adversarial approach,” IEEE Sensors Journal, vol. 21, no. 2, pp. 1885–1896, 2020.

[68] ——, “Commuting conditional gans for multi-modal fusion,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3197–3201.

[69] S. Roheda, H. Krim, Z.-Q. Luo, and T. Wu, “Event driven fusion,” arXiv preprint arXiv:1904.11520, 2019.

[70] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[71] R. T. Rockafellar, “Augmented lagrange multiplier functions and duality in nonconvex programming,” SIAM Journal on Control, vol. 12, no. 2, pp. 268–285, 1974.

[72] D. G. Luenberger, Y. Ye et al., Linear and nonlinear programming. Springer, 1984, vol. 2.

[73] H.-H. Gabriel, M. Spiliopoulou, and A. Nanopoulos, “Eigenvector-based clustering using aggregated similarity matrices,” in Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 1083–1087.

[74] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, “Clustering on multi-layer graphs via subspace analysis on grassmann manifolds,” IEEE Transactions on signal processing, vol. 62, no. 4, pp. 905–918, 2013.

[75] P.-Y. Chen and A. O. Hero, “Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms,” IEEE Transactions on Signal and Information Processing over Networks, vol. 3, no. 3, pp. 553–567, 2017.

[76] K.-C. Lee, J. Ho, and D. J. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 5, pp. 684–698, 2005.

[77] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, “Robust recovery of subspace structures by low-rank representation,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 171–184, 2012.

[78] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan, “A polarimetric thermal database for face recognition research,” in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 187–194.

[79] B. Hunter and T. Strohmer, “Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements,” arXiv preprint arXiv:1011.0997, 2010.

[80] S. Roheda, B. S. Riggan, H. Krim, and L. Dai, “Cross-modality distillation: A case for conditional generative adversarial networks,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 2926–2930.

[81] S. Roheda, H. Krim, Z.-Q. Luo, and T. Wu, “Decision level fusion: An event driven approach,” in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2598–2602.

[82] Berger, Roger L., and Dennis F. Sinclair. “Testing Hypotheses Concerning Unions of Linear Subspaces.” Journal of the American Statistical Association, vol. 79, no. 385, 1984, pp. 158-163.

[83] M. Hosseini, M. Horton, H. Paneliya, U. Kallakuri, H. Homayoun, and T. Mohsenin, “On the complexity reduction of dense layers from $O(N^2)$ to $O(NlogN)$ with cyclic sparsely connected layers,” in 2019 56th ACM/IEEE Design Automation Conference (DAC). IEEE, 2019, pp. 1–6.

[84] Vinh, Nguyen Xuan, Julien Epps, and James Bailey. "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance." The Journal of Machine Learning Research 11 (2010): 2837-2854.

[85] Rand, William M. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association 66, no. 336 (1971): 846-850.

APPENDICES

APPENDIX A

Proof of Theorem 1

We recall that we only study for the case $\rho = 0$ in Eqn. (2.10). We denote by $\boldsymbol{z_j(t)}$ the dual vector associated with the constraints in Eqn. (2.10). In the following, we construct a solution w and a dual variable $\boldsymbol{z_j(t)}$ satisfying the optimality conditions of Eqn. (2.10) and the subspace detection property in Definition 1. For this purpose, we introduce the following optimization problem,

$$min_{\{\widetilde{w}_j(t)\}_{t=1}^T} \sum_l \sqrt{\sum_{t=1}^T \widetilde{w}_{l,j}^2(t)} \quad s.t. \;\; \boldsymbol{x_j(t) = X_{-j}^{\alpha}(t)\widetilde{w}_j(t)} \tag{A.1}$$

where α indicates the subspace of the j^{th} datapoint, $X_{-j}^{\alpha}(t)$ is every point in $\boldsymbol{X(t)}$ from the α subspace except $\boldsymbol{x_j(t)}$ and $\boldsymbol{\widetilde{w}_{l,j}(t)}$ denotes the l^{th} element in $\boldsymbol{\widetilde{w}_j(t)}$. With an abuse of notation, we also take $\boldsymbol{\widetilde{w}_j(t)}$ to be the optimal solution for Eqn. (A.1). Moreover, we denote its corresponding optimal dual vector with the smallest l_2 norm by $\boldsymbol{\tilde{z}_j(t)}$. Hence, $\boldsymbol{\tilde{z}_j(t)} \in col\{\boldsymbol{X_{-j}^{\alpha}(t)}\}$, where col represent the column space, since otherwise, the projection of $\boldsymbol{\tilde{z}_j(t)}$ onto $col\{\boldsymbol{X_{-j}^{\alpha}(t)}\}$ serves as another dual vector with strictly smaller l_2 norm. Then, the optimality condition of Eqn. (A.1) yields $(\boldsymbol{X_{-j}^{\alpha}(t)})^T\boldsymbol{\tilde{z}_j(t)} \in \partial \sum_l \sqrt{\sum_{t=1}^T \widetilde{w}_{l,j}^2(t)}$ where ∂ denote the sub-differential set such that,

$$(\boldsymbol{(X_{-j}^{\alpha}(t))^T\tilde{z}_j(t)})_l = \frac{\widetilde{w}_{l,j}(t)}{\sqrt{\sum_t \widetilde{w}_{l,j}(t)}} \text{ if } \sum_t \widetilde{w}_{l,j}(t) \neq 0,$$

$$\sqrt{\sum_{t=1}^T (\boldsymbol{(X_{-j}^{\alpha}(t))^T\tilde{z}_j(t)})_l^2} \leq 1 \; if \; \sum_t \widetilde{w}_{l,j}(t) = 0 \tag{A.2}$$

Now, we construct $\boldsymbol{w_j}(t)$ by appropriately appending zero entries to $\widetilde{\boldsymbol{w}}_{\boldsymbol{j}}(t)$ whenever the indices i, j of datapoints correspond to samples from different subspaces and when $i = j$, since each sample can be represented by other samples from the same subspace and can't represent itself. Furthermore, we take $\boldsymbol{z_j}(t) = \tilde{\boldsymbol{z}}_{\boldsymbol{j}}(t)$. In the following, we prove that $\boldsymbol{w_j}(t)$ and $\boldsymbol{z_j}(t)$ satisfy the optimality condition of Eqn. (2.10), hence being an optimal solution with subspace detection property. The optimality condition for Eqn. (2.10) can be written as,

$$(\boldsymbol{X}(t)\boldsymbol{z_j}(t))_k = \frac{\boldsymbol{w_{kj}(t)}}{\sqrt{\sum_t \boldsymbol{w_{kj}(t)}}} \text{ if } \boldsymbol{w_{kj}(t)} \neq 0 \text{ for some t,}$$

$$\sum_{t=1}^{T}((\mathbf{X}(t))^T \boldsymbol{z_j}(t))_k^2 \leq 1 \; if \; \sum_t w_{kj}^2(t) = 0 \tag{A.3}$$

It is simple to check that the conditions in Eqn. (A.3) for $k \in S_j$ are satisfied by the definition in Eqn. (A.2). Also, note that for $k \notin S_j$, we have $w_{k,j}(t) = 0$. Therefore, in order to prove that the subspace detection property holds, it remains to check that $\forall\, k \notin S_j$, we have,

$$\sum_{t=1}^{T} < \boldsymbol{x_k}(t), \boldsymbol{z_j}(t) >^2 < 1 \tag{A.4}$$

To proceed, we observe that,

$$\sum_{t=1}^{T}(\boldsymbol{x_k^T}(t)\boldsymbol{z_j}(t))^2 = \sum_{t=1}^{T}\left\|\boldsymbol{x_k^T}(t)\right\|^2 \left\|\boldsymbol{z_j}(t)\right\|^2 cos^2(\theta_{k,j}(t)) \tag{A.5}$$

Where $\boldsymbol{\theta_{kj}(t)}$ is the angle between $\boldsymbol{x_k}(t)$ and $\boldsymbol{z_j}(t)$. Since the data is normalized, $\left\|\boldsymbol{x_k^T}(t)\right\|^2 = 1$. Then, we have,

$$\sum_{t=1}^{T}\left\|\boldsymbol{z_j}(t)\right\|^2 cos^2(\theta(t)) \leq max_t cos^2(\theta(t)) \sum_{t=1}^{T}\left\|\boldsymbol{z_j}(t)\right\|^2 \tag{A.6}$$

The polar of a convex body C is given by,

$$C^o = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x^T} c \leq 1 \forall\, c \in C\} \tag{A.7}$$

Therefore, we observe that the polar P_{-j}^o of P_{-j} is given by:

$$P_{-j}^o = \{\{\boldsymbol{y}(t)\}_{t=1}^T : \sum_{t=1}^{T} (\boldsymbol{x_q^T}(t)\boldsymbol{y}(t))^2 \leq 1\} \tag{A.8}$$

As a result, from Eqn. (A.2), we conclude that $\{\boldsymbol{z_j}(t)\}_{t=1}^T \in P_{-j}^o$. We define the circumradius of a convex subset P of the finite dimension Euclidean space as the radius of the smallest Euclidean ball containing P and denote it by $R(P)$. Hence,

$$\sum_{t=1}^{T} \left\|\boldsymbol{z_j}(t)\right\|^2 \leq R^2(P_{-j}^o) \tag{A.9}$$

For a symmetric convex body P, the following relationship between the inradius of P and circumradius of its polar P^o holds [60]:

$$r(P)R(P^o) = 1 \tag{A.10}$$

Therefore,

$$\sum_{t=1}^{T} \left\|\boldsymbol{z_j}(t)\right\|^2 \leq \frac{1}{r^2(P_{-j})} \tag{A.11}$$

In summary, Eqns. (A.6) and (A.10) imply that it suffices to verify that $\forall k \notin S_j$, we have,

$$max_t\, cos^2\theta(t) < min_t\, r^2(P_{-j}) \tag{A.12}$$

Then, the condition in Eqn. (A.4) is satisfied and $w_j(t)$ is a solution for Eqn. (A.10) when $\rho = 0$ which implies that the subspace detection property holds.

APPENDIX B

Parameter Perturbation Analysis

To theoretically compare our proposed variational scaling fusion approach (DRoGSuRe) to DMSC, we proceed by way of a first order perturbation analysis on the parameter set $\boldsymbol{W}^i$ of respectively either technique $i = 1, 2$. This will, in turn impact the associated affinity matrix $\boldsymbol{A}^i$, which as we will later elaborate directly impacts the subspace clustering procedure which is central to the inference following the fusion procedure.

Adopting the original formulation for the first persistently differential scaling approach, namely that T modalities are jointly exploited, results in, $\boldsymbol{X^1}(t) = \left[\boldsymbol{x_1^1}(t)\ \boldsymbol{x_2^1}(t)\ \ldots\ \boldsymbol{x_n^1}(t)\right]$, where $\boldsymbol{x_k^1}(t) \in \mathbb{R}^m, t = 1, 2, \ldots, T$ represents the k^{th} observation. The second approach only effectively uses only one subspace structure of the fused modalities $\boldsymbol{X^2}(t) = \left[\boldsymbol{x_1^1}\ \boldsymbol{x_2^1}\ \ldots\ \boldsymbol{x_n^1}\right]$. A first order perturbation on the data may be due to noise or to a degradation of a given sensor, and results in a perturbation of the UoS parameters,

$$\widetilde{\boldsymbol{W}}_{\boldsymbol{1}}^{\boldsymbol{i}} = \boldsymbol{W}_{\boldsymbol{1}}^{\boldsymbol{i}} + \delta^i \tag{B.1}$$

For the first method, each modality will have an associated subspace cluster parameter set $\{\boldsymbol{W}_t^1\}_{t=1,\ldots,T}$, with $\boldsymbol{W_t^1} \in \mathbb{R}^{n\times n}$. The overall parameter set for DRoGSuRe can then be written as,

$$\widetilde{\boldsymbol{W}}^{\boldsymbol{1}} = \widetilde{\boldsymbol{W}}_1^1 + \boldsymbol{W}_2^1 + \cdots + \boldsymbol{W}_m^1 \tag{B.2}$$

Where the unperturbed overall sparse coefficient matrix is written as follows, $\boldsymbol{W_{tot}^1} = \boldsymbol{W_1^1} + \boldsymbol{W_2^1} + \cdots + \boldsymbol{W_m^1}$. A similar development follows for method 2, with the difference that the contributing modalities are fused a priori.

Proposition. The persistent differential scaling of m-modal Group Robust Subspace Clustering Fusion yields an order m-improvement resilience over the singly differential scaling fusion.

Proof. We first write the affinity matrix associated with DRoGSuRE as,

$$\widetilde{\boldsymbol{A}}^{1} = \widetilde{\boldsymbol{W}}^{1}_{tot} + (\widetilde{\boldsymbol{W}}^{1}_{tot})^{T} \tag{B.3}$$

$$\widetilde{\boldsymbol{A}}^{1} = \widetilde{\boldsymbol{W}}^{\boldsymbol{1}}_{\boldsymbol{1}} + \boldsymbol{W}^{\boldsymbol{1}}_{\boldsymbol{2}} + \cdots + \boldsymbol{W}^{\boldsymbol{1}}_{\boldsymbol{m}} + (\widetilde{\boldsymbol{W}}^{\boldsymbol{1}}_{\boldsymbol{1}} + \boldsymbol{W}^{\boldsymbol{1}}_{\boldsymbol{2}} + \cdots + \boldsymbol{W}^{\boldsymbol{1}}_{\boldsymbol{m}})^{T} \tag{B.4}$$

Where the superscript T denotes transpose. This is equivalent to,

$$\widetilde{\boldsymbol{A}}^{1} = \widetilde{\boldsymbol{A}}^{1}_{1} + \sum_{i=2}^{T} \boldsymbol{A}^{1}_{i} \tag{B.5}$$

Where $0 \leq \widetilde{\boldsymbol{A}}^{1}_{1}(i,j) \leq 1 + \delta^{1}$. The unperturbed collective affinity matrix $\boldsymbol{A}^{\boldsymbol{1}}$ can be similarly written $\boldsymbol{A}^{1} = \sum_{i=1}^{T} \boldsymbol{A}^{1}_{i}$ with the unity constraint on each entry of all matrices. We may also write the magnitude of the difference,

$$\left|\boldsymbol{A}^{\boldsymbol{1}} - \widetilde{\boldsymbol{A}}^{\boldsymbol{1}}\right| = \delta^{1} + (\delta^{1})^{T} \tag{B.6}$$

Letting $\Delta = \delta^{1} + (\delta^{1})^{T} \in \mathbb{R}^{n \times n}$, and assuming $\epsilon = max_{i,j}[\Delta]_{i,j}$, we can write,

$$\left\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\right\|_{F} \leq n\epsilon \tag{B.7}$$

Given the Δ matrix individual entry bounds , we conclude that,

$$0 \leq \epsilon \leq \frac{1}{t} \tag{B.8}$$

Since DMSC assumes having one sparse coefficient matrix W for all data modalities, which is equivalent to only one subspace structure of the fused modalities $\boldsymbol{X}^{\boldsymbol{2}}(t) = [\boldsymbol{x}^{\boldsymbol{2}}_{\boldsymbol{1}} \ldots \boldsymbol{x}^{\boldsymbol{2}}_{\boldsymbol{n}}]$. Therefore, the UoS parameters will be perturbed by δ^{2} as follows,

$$\widetilde{\boldsymbol{W}}^{2} = \boldsymbol{W}^{\boldsymbol{2}} + \delta^{2} \tag{B.9}$$

The affinity matrix associated with DMSC can be written as follows, $\widetilde{\boldsymbol{A}}^{\boldsymbol{2}} = \widetilde{\boldsymbol{W}}^{\boldsymbol{2}} + (\widetilde{\boldsymbol{W}}^{\boldsymbol{2}})^T$, which is equivalent to,

$$\widetilde{\boldsymbol{A}}^{\boldsymbol{2}} = \boldsymbol{W}^{\boldsymbol{2}} + \delta^2 + (\boldsymbol{W}^{\boldsymbol{2}})^T + (\delta^2)^T \tag{B.10}$$

Similarly, the unperturbed affinity matrix will be as follows,

$$\boldsymbol{A}^{\boldsymbol{2}} = \boldsymbol{W}^{\boldsymbol{2}} + (\boldsymbol{W}^2)^T \tag{B.11}$$

The magnitude of the difference can be written as follows,

$$|\boldsymbol{A}^{\boldsymbol{2}} - \widetilde{\boldsymbol{A}}^{\boldsymbol{2}}| = \delta^2 + (\delta^2)^T \tag{B.12}$$

Letting $\gamma = \delta^2 + (\delta^2)^T \in \mathbb{R}^{n\times n}$, i.e., $\left|\boldsymbol{A}^{\boldsymbol{2}} - \widetilde{\boldsymbol{A}}^2\right| = \gamma$, and assuming $\Psi = max_{i,j}[\,\gamma]_{i,j}$, we can write $\left\|\boldsymbol{A}^2 - \widetilde{\boldsymbol{A}}^2\right\|_F \leq n\Psi$. Given the γ matrix individual entry bounds, we conclude $0 \leq \Psi \leq 1$. If we only perturb one modality, knowing that $0 \leq A(i,j) \leq 1$, therefore the error could lie between $0 \leq \Psi \leq 1$, which entails either creating a fake relation between two data points or erasing an existing relation. ϵ and Ψ are random variables that do not have to follow a specific distribution, however, in any case $E(\epsilon^2) \ll E(\,\Psi^2)$ and therefore $SNR_{DRoGSuRe} \gg SNR_{DMSC}$.

In light of the above two bounds, and the results of [79], where it is shown that the spectral clustering dependent on the respective projection operators P_{W^1} and $\tilde{P}_{\widetilde{W}^1}$ onto the vector subspaces spanned by the principal eigenvectors of $\boldsymbol{W}^{\boldsymbol{1}}_{\boldsymbol{tot}}$ and $\widetilde{\boldsymbol{W}}^{\boldsymbol{1}}_{\boldsymbol{tot}}$ of may be written as,

$$\left\|P_{W^1} - \tilde{P}_{\widetilde{W}^1}\right\|_F \leq \frac{\sqrt{2}}{\alpha^1}\left\|\boldsymbol{A}^{\boldsymbol{1}} - \widetilde{\boldsymbol{A}}^1\right\|_F \tag{B.13}$$

where α^1 is the spectral gap between the k^{th} and $(k+1)^{st}$eigen value of $\boldsymbol{A}^{\boldsymbol{1}}$, $|\lambda^1_k - \lambda^1_{k+1}|$. Similarly, for DMSC, the bound on the projection operators is,

$$\left\|P_{W^2} - \tilde{P}_{\widetilde{W}^2}\right\|_F \leq \frac{\sqrt{2}}{\alpha^2}\left\|\boldsymbol{A}^{\boldsymbol{2}} - \widetilde{\boldsymbol{A}}^{\boldsymbol{2}}\right\|_F \tag{B.14}$$

Where $\alpha^2 = |\lambda_k^1 - \lambda_{k+1}^1|$. Since $\boldsymbol{W_1^1, W_2^1, \ldots, W_T^1}$ happen to commute and if they happen to be diagonalizable, therefore, they share the same eigen vectors. As a result, the eigenvectors of $\boldsymbol{W_1^1 + W_2^1 + \cdots + W_T^1}$ are also the same and the corresponding eigenvalue that is the sum of the corresponding eigenvalues of $\boldsymbol{W_1^1, W_2^1, \ldots and\ W_T^1}$. Therefore, $\lambda_k^1 \gg \lambda_k^2$. From all the above, we can conclude that smaller error yielding to better clustering, hence preserving the performance, yields the improvement by the T-factor noted in the proposition and shown in the two perturbation developments.