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

Sally Ghanem, Student Member, IEEE Ashkan Panahi, Member, IEEE, Hamid Krim, Fellow, IEEE, and Ryan A. Kerekes, Member, IEEE

Abstract—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. The union of Subspaces (UoS) is capable of identifying more complex trends in data sets than simple linear models. We build on and extend RoSuRe to prospect the structure of different data modalities individually. We propose a novel multi-modal data fusion approach based on group sparsity which we refer to as Robust Group Subspace Recovery (RoGSSuRe). Relying on a bi-sparsity pursuit paradigm and non-smooth optimization techniques, the introduced framework learns a new joint representation of the time series from different data modalities, respecting an 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 from experiments on audio and magnetic cluster the associated target objects. The resulting fusion of the structural dependencies between the different modalities data to 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 from experiments on audio and magnetic 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.

Index Terms—Sparse learning, Unsupervised classification, Data fusion, Multimodal data.

I. INTRODUCTION

UNION of Subspaces (UoS) is a novel approach for identifying complex trends in datasets relative to simple linear models, like Robust Principal Component Analysis [1]. High dimensional data is rich with common and related features lying in corresponding subspaces and also other nonconforming structures which may be errors or outlier sparse structures. Sparse modeling has been broadly employed in the machine learning literature to model noisy data [2]. Subspace clustering can be utilized 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 widely available and large amounts of visual data, which has in turn, further motivated the development of such representations [3][4]. Other applications of subspace clustering include image segmentation [5], image compression [6] and motion segmentation [7]. Union of Subspaces has been also incorporated into analysis dictionary learning to boost the classification performance in supervised learning [8]. In this work, we build on the procedure studied in [9], 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 broad data sources [10]. 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. While data fusion is not new, it has recently garnered a significant increase in research interest for the need of further foundational principles, and for the varied applications including image fusion [11][12], target recognition [13], speaker recognition [14] and handwriting analysis [15]. Recent developments in sensor technology have provided more potential 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 [16]. Conventional vehicle identification methods such as license plate recognition and Radio Frequency Identification Tags (RFID) have long been widely used for that purpose [17]. Unfortunately, such image-based methods are not always appropriate, for example, in

An earlier version of this paper was presented at the 26th European Signal Processing Conference (EUSIPCO) in 2018 and was published in its Proceedings (https://ieeexplore.ieee.org/abstract/document/8553445).

This paper has been submitted and accepted in IEEE Sensors Journal. ©2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Sally Ghanem, Ashkan Panahi and Hamid Krim are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, 27606, USA
Ryan A. Kerekes is with Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA
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 [18]. Vehicles are primarily of metallic structure that perturb the earth’s magnetic field, and hence produce unique magnetic signatures that have served to discriminate between vehicles [19] [20]. Audio sensors have also been extensively employed in the area of vehicle classification for different applications, and have proven their effectiveness and robustness [21] [22].

The objective of this work is to develop a multi-modal classification approach which we validate on vehicle classification and identification. 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 [23] and [24]. On the other hand, dissimilar sensor fusion [25] employ different sensors to prospect different features along various dimensions of the target. Our framework assumes feature fusion from heterogeneous data modalities. However, the proposed approach is principled and sufficiently general, as no specific assumptions were made, thus making it equally applicable for homogeneous or heterogeneous data modalities. Sensor fusion of heterogeneous data has long been of interest since it can explore different characteristics of a target and provide information reinforcement for an increased resolvability of object identity. This additional information from multiple modalities enrich target characterization and support decision making. A comprehensive survey for data fusion is provided in [26] and [27].

In this work, 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, which we presented before in [28]. 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 [29]. 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 can not be easily modified to address the present issue. On the other hand, our approach can address data modalities that have different dimensionality and potentially incompatible in nature. MLAP was also based on LRR [30] while our approach is based on RoSuRe [9]. For fairness, we will compare our approach to other popular multimodal subspace clustering baselines such as LTMSC [31], kMLRR [32], kMSSC [32], MSSC [32] and MLRR [32]. LTMSC [31] exploits the complementary information of different views by jointly employing tensors to explore the high order correlations. The other methods were introduced in [32] as multimodal extension for Sparse Subspace Clustering (SSC) [33] and Low-rank Representation-based (LRR) [30] by enforcing a common representation across data modalities.

The paper is organized as follows. In Section 2, we provide the fundamental concepts of the two approaches RoSuRe and RoGSuRe. In Section 3, we introduce the different sensing modalities that will be utilized for experimentation along with the pre-processing, feature selection and extraction techniques. In Section 4, we present the experimental results of our approach, while Section 5 provides concluding remarks.

II. INFORMATION SUBSPACE-BASED FUSION

A. Problem Formulation

Consider a set of data realizations indexed by \( k = 1, 2, ..., n \). Furthermore, assume \( T \) data modalities, indexed by \( t = 1, 2, 3, ..., T \). Each data realization can be represented as a \( m \)-dimensional vector \( x_k(t) \in \mathbb{R}^m \), where \( X(t) = [x_1(t) \ x_2(t) \ ... \ x_n(t)] \). 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 \( \{X^1, X^2, ..., X^P\} \) of \( n \) observations/realizations. Also, \( P \) is the number of clusters indexed by \( I \), such that there exist linear subspaces \( S^I(t) \subset \mathbb{R}^m \) with \( \text{dim}(S^I(t)) < m \), where \( x_k(t) \in S^I(t) \) for all \( t \in [T] \).

B. Robust subspace recovery via Bi-sparsity Pursuit: Fusion of Subspace Information

The Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe) introduced in [9] 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 \( X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{m \times n} \), 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 \( X \) may be partitioned such that each part \( X^I \) is decomposed into a low dimensional subspace and a sparse outlier (e.g., non-conforming data).

\[
X^I = L^I + E^I, I = 1, ..., P,
\]

where each \( L^I \) serves as a single low dimensional subspace of the original data, and \( L = [L^1 | L^2 | ... | L^P] \) is the desired union of subspaces. Furthermore, the partition recovers the clustering of the original data samples corrupted by the error \( E = [E^1 | E^2 | ... | E^P] \). 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, \( I_s \) can be represented by the other samples from the same subspace \( S(I_s) \).

\[
I_s = \sum_{i \neq j, j \in S(I_s)} w_{ij} I_j.
\]

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

\[
L = LW.
\]

Under a suitable arrangement/permutation of the data realizations, the sparse coefficient matrix \( W \) 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 indexes \( i, j \) correspond to samples from different subspaces. As a result, the majority of the elements in \( W \) is equal to zero. After further approximations and relaxations, the problem is formulated as follows,

\[
\min_{W \in L} \|W\|_1 + \lambda \|E\|_1,
\]

s.t. \( X = L + E, \ L = LW, \ W_{ii} = 0 \),

where \( \|\cdot\|_1 \) denotes the \( l_1 \) norm, i.e. the sum of absolute values of the argument. The minimum of Eqn.(4) is approximated through linearized Alternating Direction Method of Multipliers (ADMM) and the sparsity of both \( E \) and \( W \) is traced until convergence. See [9] for more details.

1) Finding clusters in data using \( W \): The resulting \( W \) can be exploited to evaluate an affinity matrix. The affinity matrix is computed by,

\[
A = W + W^T.
\]

Subsequently, the spectral clustering method in [35] is utilized for data clustering. The method can be summarized as follows: a matrix \( D \) 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,

\[
G = D^{-1/2} AD^{-1/2}.
\]

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

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 \( W(t) \)’s, for \( t = 1, 2, \ldots, T \) modalities, through adding them as follows,

\[
W_{Total} = \sum_{t=1}^{T} W(t).
\]

By doing so, we are reinforcing the relation between data points that exist in all data modalities as reflected by the elements of \( W(t) \). 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 [36], 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 \( W(t) \). 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.

C. 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 \( \Omega = \{W(t)\}_{t=1}^{T} \), where \( W(t) = [w_{kj}(t)]_{k,j} \). We define the group \( l_1 \)-norm \( \|\Omega\|_{1,2} \) as:

\[
\|\Omega\|_{1,2} = \sum_{k,j} \sqrt{\sum_{t=1}^{T} w_{kj}^2(t)}
\]

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

\[
\min_{\Omega|w_{kj}(t)=0} \|\Omega\|_{1,2} + \rho \sum_{t=1}^{T} \|W(t)\|_1 \text{s.t. } X(t) = X(t)W(t).
\]

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 \( w_{kj}(t) \) are zero for every \( t \), which implies that \( \|\Omega\| \) is group sparse along the \( t \) dimension. Moreover, minimizing group \( l_1 \)-norm promotes a group-sparse solution. Equation (8) is separable in \( j \), where \( j \) is the column index of \( W \) and can therefore be minimized for every \( j \) and thus re-written as follows,

\[
\min_{\Omega_j} \sum_{k=1}^{n} \sqrt{\sum_{t=1}^{T} w_{kj}^2(t)} + \rho \sum_{k,t} |w_{kj}(t)| \text{s.t. } x_j(t) = X(t)x_j(t) \forall t w_{kj}(t) = 0.
\]

where \( \Omega_j = \{w_j(t)\}_{t=1}^{T} \) and \( \Omega_j = \{\Omega_j\}_{j=1}^{n} \). In order to validate our approach, our goal would be to prove that \( w_{kj} = 0 \) for all \( k \notin S_\alpha \) and \( j \in S_\alpha \), where \( S_\alpha \) is the index of the subspace containing \( x_j(t) \).

1) Robust Group Subspace Recovery:

Similarly to RoSuRe, we propose a robust and non-convex version of the above formulation in Eqn. (8). We assume that the data matrices include non-conforming elements to assume the structure \( X(t) = L(t) + E(t) \), where the columns of \( L(t) \)
reside on their corresponding subspaces and \( E(t) \) is a sparse error matrix. The optimization problem is then rewritten as,

\[
\begin{align*}
\min_{\Omega|w_{ik}(t)=0} & \quad \| \Omega \|_{1,2} + \rho \sum_{t=1}^{T} \| W(t) \|_1 + \lambda \sum_{t=1}^{T} \| E(t) \|_1 \\
\text{s.t.} & \quad X(t) = L(t) + E(t), \quad L(t) = L(t)W(t), \quad \text{(10)}
\end{align*}
\]

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. (9) can be reduced to a two-variable problem by substituting \( L(t) \) with \( X(t) - E(t) \) and using \( L(t) = L(t)W(t) \). Assuming \( T \) modalities, the Lagrangian objective functional now becomes:

\[
L(\Omega, W(t), E(t), Y(t), \mu) = \| \Omega \|_{1,2} + \rho \sum_{t=1}^{T} \| W(t) \|_1 + \\
\lambda \sum_{t=1}^{T} \| E(t) \|_1 + \sum_{t=1}^{T} \langle L(t)W(t) - L(t), Y(t) \rangle + \\
\sum_{t=1}^{T} \frac{\mu_k}{2} \| L_k(t)W(t) - L_k(t) \|_F^2, \tag{11}
\]

where \( Y(t) \) is a matrix of Lagrange multipliers and \( \mu \) is a constant. Let \( W(t) = \mathbf{1} - W(t) \), following the Chambolle-Pock algorithm for non-smooth primal dual algorithms, we update the following update rules for \( W(t) \) and \( E(t) \),

\[
W_{k+1}(t) = \arg \min_{W(t)} \| \Omega \|_{1,2} + \rho \| W(t) \|_1 + \\
\langle L_{k+1}(t)W(t) - L_{k+1}(t), Y_k(t) \rangle + \\
\frac{\mu_k}{2} \| L_{k+1}(t)W(t) - L_{k+1}(t) \|_F^2, \tag{12}
\]

\[
E_{k+1}(t) = \arg \min_{E(t)} \| E(t) \|_1 + \langle L_{k+1}(t)W_k(t), Y_k(t) \rangle + \\
\frac{\mu_k}{2} \| L_{k+1}(t)W_k(t) \|_F^2. \tag{13}
\]

The linearized ADMM in \cite{41} is used to approximate Eqns. (11) and (12) as follows,

\[
W^+_k(t) = \text{prox}_{\rho \| \cdot \|_1} (W_k(t) + \frac{L^T_{k+1}(L_{k+1}W_k(t) - Y_k(t))}{\eta_1}), \quad \tag{14}
\]

\[
W_{k+1}(t) = \gamma \frac{1}{\rho \| \cdot \|_1} (W^+_k(t)) \tag{15}
\]

\[
E_{k+1}(t) = \text{prox}_{\rho \| \cdot \|_1} (E_k(t) + \frac{L_{k+1}W_k(t) - Y_k(t))}{\rho \| \cdot \|_1} W^T_{k+1}), \tag{16}
\]

where \( \text{prox}_{\beta}(A_{i,j}(t)) = A_{i,j}(t) \ast \max\{\left(\frac{\sum_{t=1}^{T} A_{i,j}(t)^2 - \beta}{0}\right) / \sqrt{\sum_{t=1}^{T} A_{i,j}(t)^2} \} \) and \( \gamma(B_{i,j}) = \text{sign}(B_{i,j}) \ast \max\{|B_{i,j}| - \tau, 0\} \). The Lagrange multipliers are updated as follows,

\[
Y_{k+1}(t) = Y_k(t) + \mu_k(L_{k+1}(t)W_{k+1}(t) - L_{k+1}(t)) \tag{17}
\]

\[
\mu_k = \epsilon \mu_k. \tag{18}
\]

A summary of the resulting algorithm is given in Algorithm 1.

Algorithm 1 Robust Group Sparse Clustering

Initialize: \( X(t), \rho, \lambda, \epsilon, \eta_1 \) and \( \eta_2 \)

while not converged do

\[
L_{k+1}(t) = X(t) - E_k(t) \tag{19}
\]

\[
W^+_k(t) = \text{prox}_{\rho \| \cdot \|_1} (W_k(t) + \frac{L^T_{k+1}(L_{k+1}W_k(t) - Y_k(t))}{\eta_1}), \tag{20}
\]

\[
W_{k+1}(t) = \gamma \frac{1}{\rho \| \cdot \|_1} (W^+_k(t)) \tag{21}
\]

\[
E_{k+1}(t) = \text{prox}_{\rho \| \cdot \|_1} (E_k(t) + \frac{L_{k+1}W_k(t) - Y_k(t))}{\rho \| \cdot \|_1} W^T_{k+1}), \tag{22}
\]

\[
Y_{k+1}(t) = Y_k(t) + \mu_k(L_{k+1}(t)W_{k+1}(t) - L_{k+1}(t)) \tag{23}
\]

\[
\mu_k = \epsilon \mu_k \tag{24}
\]

end while

2) Theoretical Discussion: Let \( X(t) \) represent the dataset with unit length data from the \( t \)th modality for every \( t \in [T] \). Moreover, \( S^1(t) = S^1(t) \cup S^2(t) \cup \ldots S^T(t) \) is the union of subspaces of the underlying structure, where \( S^t(t) \) denotes the \( t \)th subspace of the \( t \)th modality. We seek the partitioning \( X^1, X^2, \ldots, X^P \) of \([n]\) observations as elaborated in Section II. In the following, we state the theorem that will support our approach. We study the case for \( \rho=0 \) for the sake of clarity, and compare to the work in \cite{37} for the deterministic model; in which the orientation of the subspaces as well as the distribution of the points in each subspace is deterministic. The same setting was considered by Elhamifar et. al in \cite{38} and \cite{39}. 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 \cite{37}. 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 \( \{S^t(t)\}_{t=1}^T \) and points \( X(t) \) obey the group subspace detection if and only if it holds that for all \( i \), the optimal solution to Eqn. (9) has nonzero entries only when the corresponding columns of \( X(t) \) are in the same subspace as \( x_i(t) \).

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

Definition 3. We take \( P_{-j} = \{x_i(t) | \sum_{t=1}^{T} \xi^2(t) \leq 1, \xi \neq j \} \), where \( q \) belongs to the same subspace as \( j \).

Theorem 1. Let \( X(t) \) represent the dataset with unit length data for every \( t \in [T] \). Suppose that Eqn.(9) has a feasible solution \( \Omega \) where \( w_{ij}(t) = 0 \) 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 \( \max_i \cos^2 \theta(t) < \min_j v^2(P_{-j}) \), then the subspace detection property holds.
The theorem basically guarantees that the 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 range-finder, magnetometer, and low-frequency RF antenna. The same dataset was used in [40]. 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 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 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. 1(a) and 1(b) respectively.

### III. 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 range-finder, magnetometer, and low-frequency RF antenna. The same dataset was used in [40]. 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 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 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. 1(a) and 1(b) respectively.

Acoustic sensors have been analyzed in various applications related to automatic transportation systems [41] [42]. Mel Frequency Cepstral Coefficients [43] 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 downsampled 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 filterbank 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 [44] [19] [20]. 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.

| 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              | 20             |
| Hyundai Tucson   | 50              | 20             |
| Uhaul Truck      | 50              | 32             |

Table I

The dataset description
IV. EXPERIMENTAL RESULTS

A. RoSuRe-Based Fusion

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. (4), $W(t)$, 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 [45]. 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 Eq. (4). The sparse coefficient matrix $W(t)$ \{t = 1 for magnetometer, t = 2 for audio \} is thus computed using PCA-based representation in lieu of $X(t)$. The affinity matrix is then calculated and the spectral clustering classification technique explained in II.B.1 is utilized to cluster the subspaces. The sparse coefficient matrices for magnetic and acoustic sensors are respectively illustrated in Fig. 2(a) and 2(b). The block-diagonal structure can be clearly seen from either of the matrices.

![Sparse coefficient matrices](image)

(a) The sparse coefficient matrix for magnetometer data  
(b) The sparse coefficient matrix for audio data

(c) The overall sparse coefficient matrix.

Fig. 2. The sparse coefficient matrices for ROSURE.

| TABLE II | THE CLUSTERING PERFORMANCE FOR DIFFERENT CLUSTERING METHODS |
|----------|-------------------------------------------------------------|
|          | RoSuRe | kmeans | 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% |

2) Bi-Modality RoSuRe-Driven Fusion (Multi-modal RoSuRe):

Interpreting the subspace-based affinities based on $W(t)$ as a layered set of networks, we proceed to carry out what amounts to modality fusion. The two sparse matrices $W_{audio}$ and $W_{magnetic}$ are added to produce one sparse matrix for both modalities, $W_{Total}$, thereby improving performance. By doing so, we reinforce the contribution of similar representations that exist in both modalities as justified in II.B.2. The overall sparse matrix, $W_{Total}$ is displayed in Fig. 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 $W_{Total}$. For clustering by $W_{Total}$, we applied the same spectral clustering approach that we previously demonstrated in II.B.1. As a result, the classification accuracy improved to 98.29% as highlighted in Table 2.

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 [46]. 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 [47] [48]. Hierarchical clustering is a technique which aims to build a hierarchy of clusters [49].

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. 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 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%.

B. Using Robust Group Subspace Recovery (RoGSuRe)

1) Fusing the data modalities with RoGSuRe:
In this subsection, we use the Robust Group Subspace Recov-
ery 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 III.
We start by extracting the principal components of the data
conventional to each sensor [45] to serve as representatives
of the data in the principal component space (in some sense
denised). The sparse coefficient matrix $W(t)$ is computed
by solving Eqn.(9) through Algorithm 1. Next, we threshold
$W(t)$ by its median value. The sparse coefficient matrices for
magnetic and acoustic sensors are respectively illustrated in
Fig. 3(a) and 3(b). The block-diagonal structure can be clearly
seen from either of the matrices.

\[ W_{total} = W(1) \cdot W(2) \]  

(16)

The overall sparse matrix, $W_{total}$ is displayed in Fig.
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 $W_{total}$. Using $W_{total}$, in clustering
proceeded in the same way as previously pointed out in II.B.1.
As a result, the classification accuracy improved to 98.9% as
highlighted in Table 3.

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 $l_2$
لن 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 $W_{total}$ 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
3. From the results, it is clear that the Robust Group Subspace
Recovery technique for the fused data remarkably outperforms
RoSuRe.

![Fig. 3. The sparse coefficient matrices for ROGSuRE.](image)

In the following, we will compare the performance of
RoGSuRe to some existing and known multimodal subspace
clustering technique such as LTMSC [31], kMLRR [32],
kMSSC [32], MSSC [32] and MLRR [32]. 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. 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 modal-
ities, 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

| TABLE III |

| Learning | Validation |
|----------|------------|
| RoGSuRe  | 99.14%     |
| RoSuRe   | 98.29%     |
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 IV | Performance for Different Multi-modal Subspace Clustering Methods |
|----------|---------------------------------------------------------------|
| Method   | Accuracy           |
| LTMSC    | 75.83%            |
| KMLRR    | 77.71%            |
| KMSSC    | 79.14%            |
| MLRR     | 89.14%            |
| MSSC     | 98.29%            |
| RoSuRe   | 98.29%            |
| RoGSuRe  | 99.14%            |

Fig. 4. The confusion matrices for different multi-modal subspace clustering methods.

V. Conclusion

In this paper, 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.

Appendix A

Proof of Theorem 1

Proof. We recall that we only study for the case $\rho = 0$ in Eqn. (9). We denote by $z_j(t)$ the dual vector associated with the constraints in Eqn. (9). In the following, we construct a solution $\mathbf{w}$ and a dual variable $z_j(t)$ satisfying the optimality conditions of Eqn. (9) and the subspace detection property in Definition 1. For this purpose, we introduce the following optimization problem,

$$\min_{\varphi_j(t)} \sum_{l=1}^{T} \left( \sum_{t=1}^{T} \bar{w}_{l,j}^2(t) \right) \text{s.t. } x_j(t) = \mathbf{X}_{-j}(t)\tilde{w}_j(t)$$

where $\alpha$ indicates the subspace of the $j^{th}$ datapoint, $\mathbf{X}_{-j}(t)$ is every point in $\mathbf{X}(t)$ from the $\alpha$ subspace except $x_j(t)$ and $\bar{w}_{l,j}(t)$ denotes the $l^{th}$ element in $\bar{w}_j(t)$. With an abuse of notation, we also take $\tilde{w}_j(t)$ to be the optimal solution for Eqn. (17). Moreover, we denote its corresponding optimal dual vector with the smallest $l_2$ norm by $\tilde{z}_j(t)$. Hence, $\tilde{z}_j(t) \in \text{col}\{\mathbf{X}_{-j}(t)\}$, where col represent the column space, since otherwise, the projection of $\tilde{z}_j(t)$ onto $\text{col}\{\mathbf{X}_{-j}(t)\}$ serves as another dual vector with strictly smaller $l_2$ norm. Then, the optimality condition of Eqn. (17) yields $((\mathbf{X}_{-j}(t))^T\tilde{z}_j(t))^T\tilde{z}_j(t) \in \partial \sum_{l=1}^{T} \sqrt{\sum_{t=1}^{T} \bar{w}_{l,j}^2(t)}$ where $\partial$ denote the sub-differential set such that,

$$\left( (\mathbf{X}_{-j}(t))^T\tilde{z}_j(t) \right)_t = \frac{\bar{w}_{l,j}(t)}{\sqrt{\sum_{t=1}^{T} \bar{w}_{l,j}(t)}} \text{ if } \sum_{t=1}^{T} \bar{w}_{l,j}(t) \neq 0,$$

$$\sqrt{\sum_{t=1}^{T} ((\mathbf{X}_{-j}(t))^T\tilde{z}_j(t))^2} \leq 1 \text{ if } \sum_{t=1}^{T} \bar{w}_{l,j}(t) = 0.$$

Now, we construct $\varphi_j(t)$ by appropriately appending zero entries to $\tilde{w}_j(t)$ 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 $z_j(t) = \tilde{z}_j(t)$. In the following, we prove that $\varphi_j(t)$ and $z_j(t)$ satisfy the optimality condition of Eqn. (9), hence being an optimal solution with subspace detection property. The optimality condition for Eqn. (9) can be written as:
solution for Eqn. (9) when subspace detection property holds. Then, the condition in Eqn. (20) is satisfied and therefore, from Eqn. (18), we conclude that
\[
\sum_{t=1}^{T} ((X(t))^{T}z_{j}(t))^{2} \leq 1 \text{ if } \sum_{t} w_{k,j}^{2}(t) = 0.
\]

It is simple to check that the conditions in Eqn. (19) for \( k \in S_{j} \) are satisfied by the definition in Eqn. (18). Also, note that for \( k \notin S_{j} \), we have \( w_{k,j}(t) = 0 \). Therefore, in order to prove that \( \forall k \notin S_{j} \), we have
\[
\sum_{t=1}^{T} (X_{k}(t), z_{j}(t))^{2} < 1.
\]

To proceed, we observe that
\[
\sum_{t=1}^{T} (X_{k}(t), z_{j}(t))^{2} = \sum_{t=1}^{T} \left( \frac{\theta(t)}{\|x_{k}(t)\|} \right) \leq \max_{t} \cos^{2}(\theta(t)) \sum_{t=1}^{T} \|z_{j}(t)\|^{2}.
\]

The polar of a convex body \( C \) is given by,
\[
C^{o} = \{ x \in \mathbb{R}^{n} : \|C^{T}y\| \leq 1 \forall y \in C \}
\]

Therefore, we observe that the polar \( P_{o,j}^{o} \) of \( P_{o,j} \) is given by:
\[
P_{o,j}^{o} = \left\{ y(t)_{t=1}^{T} \right\} = \left\{ \left( x_{q}(t)_{t=1}^{T} y(t) \right)^{2} \leq 1 \right\}
\]

As a result, from Eqn. (18), we conclude that \( \{z_{j}(t)\}_{t=1}^{T} \in P_{o,j}^{o} \). 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,
\[
\sum_{t=1}^{T} \|z_{j}(t)\|^{2} \leq R^{2}(P_{o,j}^{o})
\]

For a symmetric convex body \( P \), the following relationship between the inradius of \( P \) and circumradius of its polar \( P^{o} \) holds [50]:
\[
r(P)R(P^{o}) = 1
\]

Therefore,
\[
\sum_{t=1}^{T} \|z_{j}(t)\|^{2} \leq \frac{1}{r^{2}(P_{o,j}^{o})}
\]

In summary, Eqns. (22) and (26) imply that it suffices to verify that \( \forall k \notin S_{j} \), we have,
\[
\max_{t} \cos^{2}(\theta(t)) < \min_{t} r^{2}(P_{o,j}^{o})
\]

Then, the condition in Eqn. (20) is satisfied and \( w_{j}(t) \) is a solution for Eqn. (9) when \( \rho = 0 \) which implies that the subspace detection property holds.

\[\square\]
