Low-Rank and Joint Sparse Representations for Multi-modal Recognition

Heng Zhang, *Student Member, IEEE,* Vishal M. Patel, *Senior Member, IEEE,* and Rama Chellappa, *Fellow, IEEE*

***Abstract*—We propose multi-task and multivariate methods for multi-modal recognition based on low-rank and joint sparse representations. Our formulations can be viewed as generalized versions of multivariate low-rank and sparse regression, where sparse and low-rank representations across all modalities are imposed. One of our methods simultaneously couples information within different modalities by enforcing the common low-rank and joint sparse constraints among multi-modal observations. We also modify our formulations by including an occlusion term that is assumed to be sparse. The alternating direction method of multipliers is proposed to efficiently solve the resulting optimization problems. Extensive experiments on three publicly available multi-modal biometrics and object recognition datasets show that our methods compare favorably with other feature-level fusion methods.**

***Index Terms*—multi-modal recognition, feature-level fusion, low-rank representation, joint-sparse representation.**

I. Introduction

Developments in sensing and communication technologies have led to an explosion in the availability of data from multiple sources and modalities. Millions of sensors of different types have been installed in buildings, streets, and airports around the world that are capable of capturing multi-modal information such as light, depth and heat. This has resulted in the development of various multi-sensor fusion methods [1], [2].

The idea of fusing multiple sources or modalities to achieve better performance compared to using a single modality alone is appealing. In particular, multi-modal classification has received a lot of attention where one uses information from various modalities recording the same physical event to achieve improved classification performance. Many practical systems are multi-modal systems. For example, in multi-modal biometrics systems, similarity scores generated by multiple features extracted from face, fingerprints and iris are integrated for identity recognition [3], [4]. One advantage of multi-modal biometrics systems is that they are less vulnerable to spoof attacks.

In recent years, sparse and low-rank representations have been explored in problems such as matrix recovery [5], [6], [7], compressive sensing [8], regression [9], and subspace clustering [10], [11], [12], [13]. In particular, a low-rank and joint sparse representation-based method was proposed in [8] to recover hyperspectral images from a very few number of noisy compressive measurements. A low-rank sparse subspace clustering (LRSSC) method was proposed in [13] that simultaneously enforces low-rank and sparse constraints on the representation matrix for subspace clustering. The trade-off between self-expressiveness property and graph-connectivity was analyzed and LRSSC was shown to take advantage of both low-rank and sparse constraints to yield improved clustering performance.

Motivated by recent developments in joint sparse and low-rank matrix recovery [8], clustering [13] and multi-modal fusion [14], [15], we propose multi-modal feature-level fusion methods by simultaneously enforcing low-rank and joint sparsity constraints across the representations corresponding to multiple modalities. We derive efficient optimization algorithms using the alternating direction method of multipliers (ADMM) to solve the resulting optimization problems. Once the representation coefficients are estimated, the minimum reconstruction rule is used for multi-modal recognition. Figure 1 gives an overview of the proposed method.

This paper makes the following contributions:

1) A general formulation based on low-rank and joint sparse representation is proposed for multi-modal recognition.
2) A modified formulation based on common sparse and low-rank representation is proposed to robustly leverage the correlation and coupling information across the modalities especially when the performance of each modality differs a lot.
3) We evaluate our method on various multi-modal recognition problems such as multi-modal active authentication [16], [17], multi-biometrics recognition [18], and multi-modal object recognition [19].

Earlier versions of this work appeared in [15] and [20]. Joint sparse and low-rank representation as well as common joint sparse and low rank representation-based frameworks and extensive experimental evaluations on the object recognition dataset are extension to [15] and [20].

The rest of the paper is organized as follows. In Section II, we briefly review various multi-modal feature-level fusion algorithms. In Section III, we introduce our formulation based on low-rank and joint sparse representations and present two special cases of the proposed method. In Section IV, we present an extension of our method based on common sparse and low-rank representations. An optimization algorithm based on the ADMM method is presented in Section V. Experimental

Heng Zhang is with the Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742 USA (e-mail: hzhang98@umiacs.umd.edu).

Vishal M. Patel is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 USA (e-mail: vishal.m.patel@rutgers.edu).

Rama Chellappa is with the Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742 USA (e-mail: rama@umiacs.umd.edu).

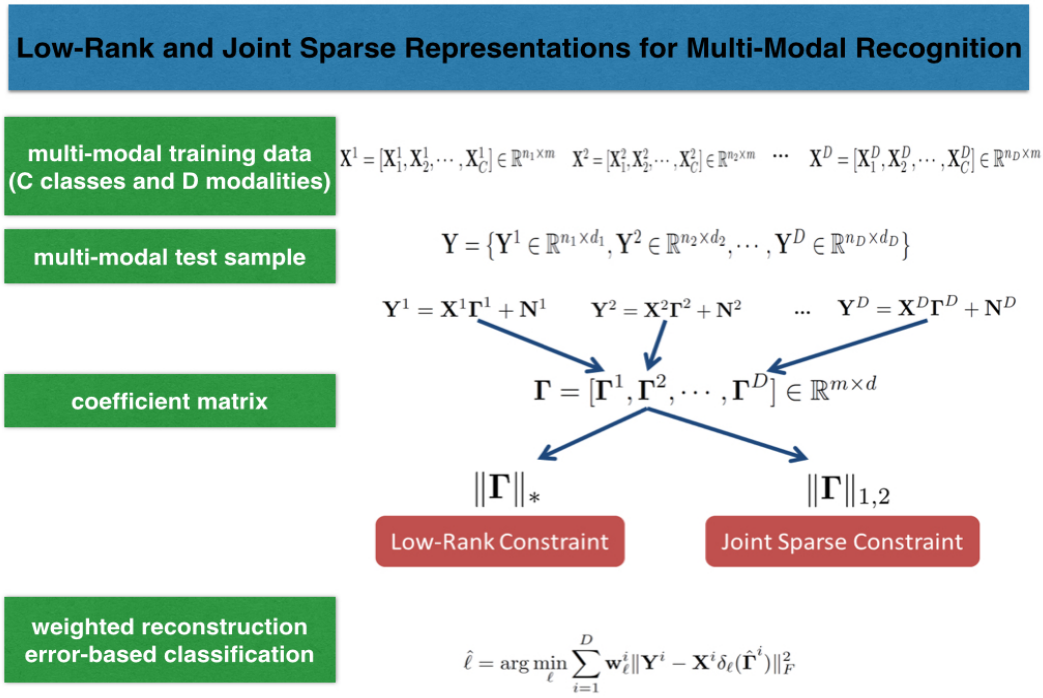

Fig. 1: An overview of the proposed low-rank and joint sparse representation-based multi-modal recognition.

evaluations on three multi-modal datasets are described in Section VI. In Section VII, the complexity of proposed methods is analyzed. Finally, concluding remarks are presented in Section VIII with a brief summary and discussion.

II. Related Work

Data fusion can be achieved at several different levels, which can be broadly classified as sensor-level, feature-level, score-level or decision-level fusion. Since feature-level fusion preserves the raw information, it can be more discriminative and robust than score-level or decision-level fusion. The focus of this paper is on designing new feature-level fusion methods and making comparisons with previous feature-level fusion methods.

Differences in features extracted from different modalities in terms of types and dimensions make the feature-level fusion non trivial. One of the simplest methods for feature-level fusion is feature concatenation [21], [22]. However, feature concatenation often tends to be computationally demanding and inefficient. Multiple Kernel Learning (MKL) has also been used to integrate information from multiple features by learning a weighted combination of appropriate kernels. See [23] for more details on various MKL algorithms.

Recent multi-modal fusion methods based on sparse or low-rank representations of multi-modal data have been shown to produce state-of-the-art results on various multi-modal recognition problems. In [24], a multi-task sparse linear regression model is proposed for image classification. In [25], a joint dynamic sparse representation method was proposed to recognize object viewed from multiple observations (e.g. poses). In [14], a joint sparse representation-based method was proposed for fusing multiple biometrics features. This method is based on multi-task, multivariate Lasso [26]. [15] proposed low-rank representation-based multi-modal recognition methods. In [20] and [15], the idea of enforcing common sparse (low-rank) representation was shown to be robust and more effective especially when the quality of different modality differs a lot.

In [27], a general collaborative sparse-representation framework for multi-sensor classification is proposed. Joint sparsity is enforced within each sensor's multiple observations and is also simultaneously enforced across heterogeneous sensors. Sparse noise and low rank interference signals are considered in their approach. The objective of the resulting optimization is to seek a joint sparse representation while minimizing the sparse error or low rank interference signals. A multi-modal task-driven dictionary learning algorithm with joint sparsity constraint enforced across multiple sources of information is proposed in [28]. In [8], a low-rank and joint sparse representation-based method is proposed for recovering hyperspectral images from a small number of noisy compressive measurements.

Other recent multi-modal feature-level fusion methods include [29] and [30]. In [29], a class consistent multi-modal fusion (CCMM) scheme was proposed which essentially extends the application of binary codes [31] for multi-modal recognition. In [30], harmonic image fusion was proposed to achieve clutter mitigation and speckle noise reduction.

III. Low-rank and joint sparse representations for multi-modal recognition

Suppose we are given a C-class classification problem with D different modalities. Assume there are m training samples in each modality. For each modality, $i = 1, \cdots, D$, we denote $\mathbf{X}^i = [\mathbf{X}^i_1, \mathbf{X}^i_2, \cdots, \mathbf{X}^i_C]$ as an $n_i \times m$ matrix of training samples containing C sub-matrices $\mathbf{X}^i_j$'s corresponding to C different classes. Each sub-matrix $\mathbf{X}^i_j = [\mathbf{x}^i_{j,1}, \mathbf{x}^i_{j,2}, \cdots, \mathbf{x}^i_{j,m_j}] \in \mathbb{R}^{n_i \times m_j}$ contains a set of training samples from the ith modality corresponding to the jth class. Here, m_j is the number of training samples in class j and n_i is the feature dimension of each sample. As a result, there are in total $m = \sum_{j=1}^{C} m_j$ many samples in $\mathbf{X}^i$. Given a test matrix $\mathbf{Y}$, which consists of D different modalities, $\{\mathbf{Y}^1, \cdots, \mathbf{Y}^D\}$, where each sample $\mathbf{Y}^i$ consists of d_i observations $\mathbf{Y}^i = [\mathbf{y}^i_1, \mathbf{y}^i_2, \cdots, \mathbf{y}^i_{d_i}] \in \mathbb{R}^{n_i \times d_i}$, the objective is to identify the class to which a test sample $\mathbf{Y}$ belongs to.

A. Basic version

In the case when the data is contaminated by random noise, the observations from ith modality can be modeled as follows

$$\mathbf{Y}^i = \mathbf{X}^i\mathbf{\Gamma}^i + \mathbf{N}^i,$$

where $\mathbf{N}^i$ is small dense additive noise. Let $\mathbf{\Gamma} = [\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, \cdots, \mathbf{\Gamma}^D] \in \mathbb{R}^{m \times d}$ be the coefficient matrix formed by concatenating D representation matrices with $d = \sum_{i=1}^{D} d_i$. We wish to solve for the low-rank and joint sparse matrix $\mathbf{\Gamma}$ by solving the following problem

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \frac{1}{2} \sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}^i\|_F^2 + \lambda_1 \|\mathbf{\Gamma}\|_* + \lambda_2 \|\mathbf{\Gamma}\|_{1,2}, \quad (1)$$

where $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$ is the Frobenius norm of $\mathbf{A}$; $\|\mathbf{A}\|_* = \sum_i \sigma_i(A)$ is the sum of the singular values of $\mathbf{A}$ (i.e. the nuclear norm of $\mathbf{A}$); $\|\mathbf{A}\|_{1,2} = \sum_k \|\mathbf{a}^k\|_2$ and $\mathbf{a}^k$ is the kth row vector of the matrix $\mathbf{A}$ (i.e the row sparsity of $\mathbf{A}$); λ_1 and λ_2 are two positive regularization parameters corresponding to low rank constraint and joint sparse constraint, respectively.

Once the coefficient matrix $\hat{\mathbf{\Gamma}}$ is obtained, the class label associated with an observation vector is declared as the one that produces the smallest approximation error

$$\hat{\ell} = \arg\min_{\ell} \sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i \delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_F^2, \quad (2)$$

where $\delta_\ell(\cdot)$ is the matrix indicator function that keeps rows corresponding to the ℓth class and sets all other rows equal to zero.

Ideally, the learned coefficients corresponding to the correct class should exhibit relatively larger values compared to coefficients corresponding to the incorrect classes. In order to exploit this, for a given coefficient vector obtained from the ith modality, we define $\mathbf{w}^i$ as:

$$\mathbf{w}^i = \frac{\lambda_1 (C \frac{max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_*}{\|\hat{\mathbf{\Gamma}}^i\|_*} - 1) + \lambda_2 (C \frac{max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_{1,2}}{\|\hat{\mathbf{\Gamma}}^i\|_{1,2}} - 1)}{(\lambda_1 + \lambda_2)(C-1)}. \quad (3)$$

This weight measures the quality of the learned representation. Representation of high quality will be low-rank ($\max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_*$ close to $\|\hat{\mathbf{\Gamma}}^i\|_*$) and will also be joint sparse ($\max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_{1,2}$ close to $\|\hat{\mathbf{\Gamma}}^i\|_{1,2}$).

The classification rule (2) based on the weighted reconstruction error can be modified as follows

$$\hat{\ell} = \arg\min_{\ell} \sum_{i=1}^{D} \mathbf{w}^i \|\mathbf{Y}^i - \mathbf{X}^i \delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_F^2. \quad (4)$$

Similar ideas have been explored in [32], [14], [15] and [20]. We call the resulting algorithm Multi-modal Recognition via Low-Rank and Joint Sparse (MRLRJS) representation.

B. Robust version

In the case when data is contaminated by noise and occlusion, the observation from the ith modality can be modeled as follows

$$\mathbf{Y}^i = \mathbf{X}^i\mathbf{\Gamma}^i + \mathbf{N}^i + \mathbf{E}^i,$$

where $\mathbf{N}^i$ is a small dense additive noise and $\mathbf{E}^i$ is a matrix of sparse occlusion with arbitrary large magnitude. By taking advantage of the fact that $\mathbf{E}^i$ is sparse, one can simultaneously estimate $\mathbf{\Gamma}^i$ and $\mathbf{E}^i$ by solving the following optimization problem

$$\hat{\mathbf{\Gamma}}, \hat{\mathbf{E}} = \arg\min_{\mathbf{\Gamma}, \mathbf{E}} \frac{1}{2} \sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}^i - \mathbf{E}^i\|_F^2 + \lambda_1 \|\mathbf{\Gamma}\|_* + \lambda_2 \|\mathbf{\Gamma}\|_{1,2} + \lambda_3 \sum_{i=1}^{D} \|\mathbf{E}^i\|_1, \quad (5)$$

where $\mathbf{E} = [\mathbf{E}^1, \mathbf{E}^2, \cdots, \mathbf{E}^D]$ is the spare occlusion matrix and $\|\mathbf{A}\|_1 = \sum_{i,j} |A_{i,j}|$ is the ℓ_1-norm of $\mathbf{A}$. Note that $\mathbf{E}$ is just a compact representation and we solve each $\mathbf{E}^i$ separately since their dimensions can be different. Here, λ_1, λ_2 and λ_3 are positive parameters that control the rank of coefficients, joint sparsity of the coefficients and the sparsity of the occlusion term, respectively.

Once $\mathbf{\Gamma}$ and $\mathbf{E}$ are estimated, the effect of occlusion can be removed by setting $\hat{\mathbf{Y}}^i = \mathbf{Y}^i - \hat{\mathbf{E}}^i$. Finally, one can declare the class label associated with an observation vector as

$$\hat{\ell} = \arg\min_{\ell} \sum_{i=1}^{D} \mathbf{w}^i \|\mathbf{Y}^i - \mathbf{X}^i \delta_\ell(\hat{\mathbf{\Gamma}}^i) - \hat{\mathbf{E}}^i\|_F^2, \quad (6)$$

where $\mathbf{w}^i_\ell$ is defined in (3). We call the resulting algorithm Robust Multi-modal Recognition via Low-Rank and Joint Sparse (RMRLRJS) representation.

C. Two Special Cases

The above formulations take both low-rank and joint sparsity into consideration and the parameters λ_1 and λ_2 control the relative importance between these representations. If λ_1 is set equal to 0, MRLRJS and RMRLRJS are reduced to joint sparse representation-based multi-modal recognition methods proposed in [14]. If λ_2 is set equal to 0, MRLRJS and RMRLRJS reduce to low-rank representation-based multi-modal recognition methods proposed in [15].

Enforcing joint sparsity (row sparsity) ensures that the number of rows that have nonzero norm to be small. Ideally, these nonzero rows correspond to the true class. A matrix which has row sparsity can also be a low-rank matrix (e.g. many rows of the matrix are null vectors). The reason for enforcing the low-rank constraint is that it can explore the underlying structure of the representation matrix especially in the column sense. For the given input multi-modal instance, representations of different modalities are assumed to be correlated, therefore, when these representations are stacked horizontally, the resulting representation matrix is assumed to have a small column rank.

In our experiments, we observed that instances where (1) with $\lambda_1 = 0$ fails are often different from those where (1) with $\lambda_2 = 0$ fails. Hence, combining the two algorithms may lead to a better multi-modal fusion method, since the underlying representation matrix we want to recover is both row-sparse and low-rank simultaneously. Our work is specifically motivated by [9] and [13] where simultaneous ℓ_1-norm and nuclear norm have been studied for general regression and subspace clustering problems, respectively. In contrast, our focus in this paper is specifically on multi-modal recognition problems.

IV. Common low-rank and joint sparse representations for multi-modal recognition

Different from previous formulations, we propose to enforce common sparse and low-rank representations on the coefficients from different modalities. As a result, we are able to exploit the correlations among the different modalities better. In this method, the coefficient matrices corresponding to D different modalities are required to be the same as follows

$$\mathbf{\Gamma} = \mathbf{\Gamma}^1 = \mathbf{\Gamma}^2 = \cdots = \mathbf{\Gamma}^D.$$

In order to make the coefficient matrices match in terms of matrix dimensions, for classifying a multi-modal instance in test phase, the number of samples from each modality has to be the same. With the common representation, imposing low-rank and joint sparse constraints on the concatenated matrix $[\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, \cdots, \mathbf{\Gamma}^D]$ is equivalent to enforcing the constraint on $\mathbf{\Gamma}^1$. Similar ideas have been explored in [33] for image super-resolution and in [15], [20] for multi-modal recognition.

A. Robust Formulation

When we consider the common representation, we assume that the ith modality's observations are of the following form

$$\mathbf{Y}^i = \mathbf{X}^i\mathbf{\Gamma} + \mathbf{N}^i + \mathbf{E}^i.$$

With this model, the robust common low-rank and joint sparse representation-based multi-modal recognition (RMRLRJS-C) problem can be formulated as

$$\hat{\mathbf{\Gamma}}, \hat{\mathbf{E}} = \arg\min_{\mathbf{\Gamma},\mathbf{E}} \frac{1}{2}\sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma} - \mathbf{E}^i\|_F^2 + \lambda_1\|\mathbf{\Gamma}\|_* + \lambda_2\|\mathbf{\Gamma}\|_{1,2} + \lambda_3\sum_{i=1}^{D}\|\mathbf{E}^i\|_1. \quad (7)$$

For the case of single modality ($D = 1$), the common representation-based formulation is the same as the general formulation proposed in Section III-B. For the case of multiple modalities, the common representation-based formulation requires the number of test samples of each modality to be the same. This requirement makes the common representation-based formulation less general.

It is easy to note that the model based on the common representation is equivalent to vertically stacking features corresponding to each modality together and applying the optimization algorithm with $D = 1$. After solving for $\hat{\mathbf{\Gamma}}$ and $\hat{\mathbf{E}}$, the following minimum reconstruction error rule can be used to classify the multi-modal data as

$$\hat{\ell} = \arg\min_{\ell} \sum_{i=1}^{D} \mathbf{w}\|\mathbf{Y}^i - \mathbf{X}^i\delta_\ell(\hat{\mathbf{\Gamma}}) - \hat{\mathbf{E}}^i\|_F^2, \quad (8)$$

where $\mathbf{w}$ is defined as

$$\mathbf{w} = \frac{\lambda_1(C\frac{\max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}})\|_*}{\|\hat{\mathbf{\Gamma}}\|_*} - 1) + \lambda_2(C\frac{\max_\ell \|\delta_\ell(\hat{\mathbf{\Gamma}})\|_{1,2}}{\|\hat{\mathbf{\Gamma}}\|_{1,2}} - 1)}{(\lambda_1 + \lambda_2)(C-1)}. \quad (9)$$

Similar to the formulation (5), eliminating the sparse error term $\mathbf{E}^i$ leads to the basic version which is denoted as common low-rank and joint sparse representation-based multi-modal recognition (MRLRJS-C).

B. Two Special Cases

Similarly, if λ_1 is set equal to 0, MRLRJS-C and RMRLRJS-C reduce to the common sparse representation-based multi-modal recognition methods proposed in [20]; if λ_2 is set equal to 0, MRLRJS-C and RMRLRJS-C reduce to the common low-rank representation-based multi-modal recognition methods proposed in [15].

V. Optimization

In this section, we propose an approach based on the ADMM method [34] to solve the resulting optimization problems. Due to the similarity of these problems, we only provide details on the optimization of (5). In ADMM, appropriate auxiliary variables are introduced into the optimization program, the constraints are augmented into the objective function and the Lagrangian is iteratively minimized with respect to the primal variables and maximized with respect to the Lagrange multipliers.

A. Optimization of RMRLRJS

Problem (5) can be reformulated by introducing the auxiliary variables as follows

$$\arg\min_{\mathbf{\Gamma},\mathbf{E},\mathbf{V},\mathbf{U},\mathbf{Z}} \frac{1}{2}\sum_{i=1}^{D} \|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}^i - \mathbf{E}^i\|_F^2 + \lambda_1\|\mathbf{V}\|_* + \lambda_2\|\mathbf{Z}\|_{1,2} + \lambda_3\sum_{i=1}^{D}\|\mathbf{E}^i\|_1 \quad (10)$$

$$\text{s.t. } \mathbf{\Gamma} = \mathbf{V}, \mathbf{\Gamma} = \mathbf{Z}.$$

Note that similar to $\mathbf{\Gamma}$, we denote $\mathbf{V} = [\mathbf{V}^1, \mathbf{V}^2, \cdots, \mathbf{V}^D] \in \mathbb{R}^{m\times d}$, $\mathbf{Z} = [\mathbf{Z}^1, \mathbf{Z}^2, \cdots, \mathbf{Z}^D] \in \mathbb{R}^{m\times d}$.

Equation (10) can be solved using the Augmented Lagrangian Method (ALM) [34]. The augmented Lagrangian function $f_{\alpha_V,\alpha_Z}(\mathbf{\Gamma}, \mathbf{E}, \mathbf{V}, \mathbf{Z}; \mathbf{A}_V, \mathbf{A}_Z)$ is defined as

$$\begin{aligned}\arg\min_{\mathbf{\Gamma},\mathbf{E},\mathbf{V},\mathbf{Z}} &\frac{1}{2}\sum_{i=1}^{D}\|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}^i - \mathbf{E}^i\|_F^2 + \lambda_3\sum_{i=1}^{D}\|\mathbf{E}^i\|_1 \\ &+ \lambda_1\|\mathbf{V}\|_* + \langle \mathbf{A}_V, \mathbf{\Gamma} - \mathbf{V}\rangle + \frac{\alpha_V}{2}\|\mathbf{\Gamma} - \mathbf{V}\|_F^2 \\ &+ \lambda_2\|\mathbf{Z}\|_{1,2} + \langle \mathbf{A}_Z, \mathbf{\Gamma} - \mathbf{Z}\rangle + \frac{\alpha_Z}{2}\|\mathbf{\Gamma} - \mathbf{Z}\|_F^2,\end{aligned} \tag{11}$$

where $\mathbf{A}_V$ and $\mathbf{A}_Z$ are the multipliers of the linear constrains, α_V and α_Z are the positive parameters, $\langle \mathbf{A}, \mathbf{B}\rangle$ denotes $tr(\mathbf{A}^T\mathbf{B})$. We denote $\mathbf{A}_V = [\mathbf{A}_V^1, \mathbf{A}_V^2, \cdots, \mathbf{A}_V^D] \in \mathbb{R}^{m\times d}$ and $\mathbf{A}_Z = [\mathbf{A}_Z^1, \mathbf{A}_Z^2, \cdots, \mathbf{A}_Z^D] \in \mathbb{R}^{m\times d}$.

In the ALM algorithm, f_{α_V,α_Z} is solved iteratively with respect to $\mathbf{\Gamma}, \mathbf{E}, \mathbf{V}$ and $\mathbf{Z}$ jointly while keeping $\mathbf{A}_V$ and $\mathbf{A}_Z$ fixed and then updating $\mathbf{A}_V$ and $\mathbf{A}_Z$ keeping the remaining variables fixed.

1) Update step for $\mathbf{\Gamma}$*:* Obtain $\mathbf{\Gamma}_{t+1}$ by minimizing f_{α_V,α_Z} with respect to $\mathbf{\Gamma}$. This can be done by taking the first-order derivative of $f_{\alpha_V,\alpha_Z,\alpha_U}$ and setting it equal to zero. Furthermore, as the first term of f_{α_V,α_Z} is a sum of convex functions associated with sub-matrices $\mathbf{\Gamma}^i$, one can find $\mathbf{\Gamma}_{t+1}^i (i = 1, \cdots, D)$ by solving the following linear system

$$\begin{aligned}({\mathbf{X}^i}^T\mathbf{X}^i + (\alpha_V + \alpha_Z)\mathbf{I})\mathbf{\Gamma}_{t+1}^i &= {\mathbf{X}^i}^T(\mathbf{Y}^i - \mathbf{E}_t^i) \\ &+ \alpha_V\mathbf{V}_t^i - \mathbf{A}_{V,t}^i + \alpha_Z\mathbf{Z}_t^i - \mathbf{A}_{Z,t}^i,\end{aligned} \tag{12}$$

where $\mathbf{I}$ is $m \times m$ identity matrix and $\mathbf{E}_t^i$, $\mathbf{V}_t^i$, $\mathbf{Z}_t^i$, $\mathbf{A}_{V,t}^i$ and $\mathbf{A}_{Z,t}^i$ are submatrices of $\mathbf{E}_t, \mathbf{V}_t, \mathbf{Z}_t$, $\mathbf{A}_{V,t}$ and $\mathbf{A}_{Z,t}$, respectively. When m is is not very large, one can simply apply matrix inversion to obtain $\mathbf{\Gamma}_{t+1}^i$ from (12). For large values of m, gradient-based methods should be employed to obtain $\mathbf{\Gamma}_{t+1}^i$.

2) Update step for $\mathbf{E}$*:* In order to update $\mathbf{E}_{t+1}^i (i = 1, \cdots, D)$, one needs to solve the following ℓ_1 minimization problem

$$\min \frac{1}{2}\|\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}_{t+1}^i - \mathbf{E}_{t+1}^i\|_F^2 + \lambda_3\|\mathbf{E}^i\|_1, \tag{13}$$

whose solution is given by [35]

$$\mathbf{E}_{t+1}^i = \mathcal{S}\left(\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}_{t+1}^i, \lambda_3\right),$$

where $\mathcal{S}(a, b) = sgn(a)(|a| - b)$ for $|a| \geq b$ and zero otherwise.

3) Update step for $\mathbf{V}$*:* The subproblem for updating $\mathbf{V}$ has the following form

$$\min \frac{1}{2}\|\mathbf{\Gamma}_{t+1} + \alpha_V^{-1}\mathbf{A}_{V,t} - \mathbf{V}\|_F^2 + \frac{\lambda_1}{\alpha_V}\|\mathbf{V}\|_*. \tag{14}$$

Solution to this optimization problem is obtained by shrinking the singular values of $\mathbf{\Gamma}_{t+1} + \alpha_V^{-1}\mathbf{A}_{V,t}$ [36], [37]. As a result, we obtain the following update for $\mathbf{V}$

$$\mathbf{V}_{t+1} = \mathbf{F}\mathcal{S}(\mathbf{\Sigma}, \frac{\lambda_1}{\alpha_V})\mathbf{B}^T,$$

where $\mathbf{F}\mathbf{\Sigma}\mathbf{B}^T$ is the Singular Value Decomposition (SVD) of $\mathbf{\Gamma}_{t+1} + \alpha_V^{-1}\mathbf{A}_{V,t}$. Same $\mathcal{S}(a, b)$ is applied as above.

4) Update step for $\mathbf{Z}$*:* In order to update $\mathbf{Z}$, we need to solve the following optimization problem

$$\min \frac{1}{2}\|\mathbf{\Gamma}_{t+1} + \alpha_Z^{-1}\mathbf{A}_{Z,t} - \mathbf{Z}\|_F^2 + \frac{\lambda_2}{\alpha_Z}\|\mathbf{Z}\|_{1,2}. \tag{15}$$

Due to the separable structure of (15), it can be solved by minimizing it with respect to each row of $\mathbf{Z}$ separately. Following the method used in [14], we let $\gamma_{i,t+1}$, $\mathbf{a}_{Z_{i,t}}$ and $\mathbf{z}_{i,t+1}$ be the ith row of matrices $\mathbf{\Gamma}_{t+1}$, $\mathbf{A}_{Z,t}$ and $\mathbf{Z}_{t+1}$ respectively. Then for each row, we solve the following subproblem

$$\mathbf{z}_{i,t+1} = \min \frac{1}{2}\|\mathbf{p} - \mathbf{z}\|_2^2 + \frac{\lambda_2}{\alpha_Z}\|\mathbf{z}\|_2, \tag{16}$$

where $\mathbf{p} = \gamma_{i,t+1} + \mathbf{a}_{Z_{i,t}}\alpha_Z^{-1}$. The closed form solution of (16) is given as follows

$$\mathbf{z}_{i,t+1} = \left(\mathbf{1} - \frac{\lambda_2}{\alpha_\mathbf{Z}\|\mathbf{p}\|_2}\right)_+\mathbf{p},$$

where $(\mathbf{v})_+$ is the vector with entries receiving values $\max(v_i, 0)$.

5) Update steps for $\mathbf{A}_V$ *and* $\mathbf{A}_Z$*:* Finally, the Lagrange multipliers are updated as

$$\mathbf{A}_{V,t+1} = \mathbf{A}_{V,t} + \alpha_V(\mathbf{\Gamma}_{t+1} - \mathbf{V}_{t+1}), \tag{17}$$

$$\mathbf{A}_{Z,t+1} = \mathbf{A}_{Z,t} + \alpha_Z(\mathbf{\Gamma}_{t+1} - \mathbf{Z}_{t+1}). \tag{18}$$

The proposed ADMM algorithm for solving the RMRLRJS problem is summarized in Algorithm 1. Note that the optimization problem is not convex and there does not exist any guarantee for the Algorithm 1 to converge. The convergence issue of ADMM is still not fully understood and remains an open research problem. Yet, ADMM works well in practice. For our proposed methods, the termination condition is either when the difference of the cost function errors is below some threshold or the maximum number of iteration is reached.

B. Optimization of RMRLRJS-C

The RMRLRJS-C problem (7) can be optimized in a similar way using the ADMM method. However, there are a few key differences in solving the subproblems. In particular, $\mathbf{\Gamma}$ is not separated into D different parts and $\mathbf{\Gamma}$ can be updated as

$$\begin{aligned}\mathbf{\Gamma}_{t+1} = &(\sum_{i=1}^{D}{\mathbf{X}^i}^T\mathbf{X}^i + (\alpha_V + \alpha_Z)\mathbf{I})^{-1}(\sum_{i=1}^{D}{\mathbf{X}^i}^T(\mathbf{Y}^i - \mathbf{E}^i) \\ &+ \alpha_V\mathbf{V}_t^i - \mathbf{A}_{V,t}^i + \alpha_Z\mathbf{Z}_t^i - \mathbf{A}_{Z,t}^i).\end{aligned} \tag{19}$$

After solving $\hat{\mathbf{E}}_i (i = 1, \cdots, D)$ and $\hat{\mathbf{\Gamma}}$, the class label can by obtained by using (8) and (9).

VI. Experimental Results

In this section, we evaluate the proposed algorithms on three publicly available multi-modal recognition datasets, namely the WVU multi-modal biometrics dataset [18], the UMDAA-01 multi-modal active authentication dataset [16], [17] and the multi-modal object recognition [19]. We compare the proposed method with various feature-level fusion methods including multiple kernel learning based multi-modal fusion

Fig. 2: Sample fingerprint and iris images from the WVU dataset.

Algorithm 1: Robust Multi-modal Recognition via Low-Rank and Joint Sparse Representation (RMRLRJS) using ADMM.

Input: Multi-modal training samples $\{\mathbf{X}_i\}_{i=1}^D$, multi-modal test sample $\{\mathbf{Y}_i\}_{i=1}^D$, $\lambda_1, \lambda_2, \lambda_3$, α_V and α_Z

Initialization:

$\mathbf{\Gamma}_0, \mathbf{V}_0, \mathbf{Z}_0, \mathbf{A}_{V,0}, \mathbf{A}_{Z,0}$ are initialized to be zero matrices.

While not converged do

1. Update $\mathbf{\Gamma}$: $\mathbf{\Gamma}_{t+1} = [\mathbf{\Gamma}_{t+1}^1, \cdots, \mathbf{\Gamma}_{t+1}^D]$, where

$$\mathbf{\Gamma}_{t+1}^i = ({\mathbf{X}^i}^T\mathbf{X}^i + (\alpha_V + \alpha_Z)\mathbf{I})^{-1}({\mathbf{X}^i}^T(\mathbf{Y}^i - \mathbf{E}_t^i) + \alpha_V\mathbf{V}_t^i - \mathbf{A}_{V,t}^i + \alpha_Z\mathbf{Z}_t^i - \mathbf{A}_{Z,t}^i)$$

2. Update $\mathbf{E}$: $\mathbf{E}_{t+1} = [\mathbf{E}_{t+1}^i, \cdots, \mathbf{E}_{t+1}^D]$, where

$$\mathbf{E}_{t+1}^i = \mathcal{S}\left(\mathbf{Y}^i - \mathbf{X}^i\mathbf{\Gamma}_{t+1}^i, \lambda_3\right)$$

3. Update $\mathbf{V}$:

$$\mathbf{V}_{t+1} = \mathbf{F}\mathcal{L}_{\frac{\lambda_1}{\alpha_V}}(\mathbf{\Sigma})\,\mathbf{B}^T$$

4. Update $\mathbf{Z}$:

$$\mathbf{z}_{i,t+1} = \left(\mathbf{1} - \frac{\lambda_2}{\alpha_\mathbf{Z}\|\mathbf{p}\|_2}\right)_+ \mathbf{p}$$

5. Update $\mathbf{A}_V$ and $\mathbf{A}_Z$:

$$\mathbf{A}_{V,t+1} = \mathbf{A}_{V,t} + \alpha_V(\mathbf{\Gamma}_{t+1} - \mathbf{V}_{t+1})$$

$$\mathbf{A}_{Z,t+1} = \mathbf{A}_{Z,t} + \alpha_Z(\mathbf{\Gamma}_{t+1} - \mathbf{Z}_{t+1})$$

Classification:

Let $\hat{\mathbf{E}}^i = \mathbf{E}_{t+1}^i (i = 1, \cdots, D)$ and $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}_{t+1}$,

1. Compute weight $\mathbf{w}^i$:

$$\mathbf{w}^i = \frac{\lambda_1(C\frac{max_\ell\|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_*}{\|\hat{\mathbf{\Gamma}}^i\|_*} - 1) + \lambda_2(C\frac{max_\ell\|\delta_\ell(\hat{\mathbf{\Gamma}}^i)\|_{1,2}}{\|\hat{\mathbf{\Gamma}}^i\|_{1,2}} - 1)}{(\lambda_1 + \lambda_2)(C-1)}$$

2. Assign the class label with minimum error:

$$\hat{\ell} = \arg\min_\ell \sum_{i=1}^D \mathbf{w}^i\|\mathbf{Y}^i - \mathbf{X}^i\delta_\ell(\hat{\mathbf{\Gamma}}^i) - \hat{\mathbf{E}}^i\|_F^2$$

Output: class label $\hat{\ell}$

method (MKL) [38], joint sparse representation-based multi-modal fusion methods (SMBR-WE and SMBR-E) [14], common sparse representation-based multi-modal fusion methods (MCSR and RMCSR) [20], low-rank representation-based multi-modal fusion methods (MLRR, RMLRR, MCLRR and RMCLRR) [15] and the class consistent multi-modal fusion (CCMM) [29].

The proposed methods can have up to six parameters during the optimization procedure. To efficiently tune these parameters, we adopt the following strategy: solve for appropriate parameters for joint sparse representation-based optimization and low-rank representation-based optimization separately and then weigh these parameters to control their relative contributions to final recognition. For example, in order to tune the parameters in Algorithm 1, we first consider the sparsity constraint only by letting λ_1 be 0 and obtain "optimal" λ_2 and λ_{3_s}, α_Z and α_{U_s} through grid search. Then, we consider the low-rank constraint only and obtain λ_1 and λ_{3_r}, α_V and α_{U_r}. Finally, we introduce a parameter $r(0 \leq r \leq 1)$ to control the relative contribution and the final parameters used are $r\lambda_1$, $(1-r)\lambda_2$ and $r\lambda_{3_r} + (1-r)\lambda_{3_s}$, $r\alpha_V$, $(1-r)\alpha_Z$ and $r\alpha_{U_r} + (1-r)\alpha_{U_s}$.

The parameters of each methods being compared are tuned in order to provide the best recognition results by fusing all the modalities. We also apply the same parameters and show the recognition result using single modality alone. In this way, the results reported using single modality may not be the best because the parameters are tuned for the fusion of all the modalities. As discussed before, when using a single modality for recognition ($D = 1$), the common representation-based methods and the corresponding general methods are the same, but the results are different. This is because the parameters used are tuned for the fusion of all the modalities and they are not necessarily the same for two kinds of methods.

A. WVU multi-modal biometrics dataset

The WVU biometrics dataset is a comprehensive collection of different biometric modalities such as fingerprint, iris, palmprint, hand geometry, and voice from subjects of different age, gender, and ethnicity. It is a challenging dataset as many of these samples are corrupted with blur, occlusion, and sensor noise. Following the experimental settings described in [14], we chose four fingerprint modalities and two iris modalities on a subset of 219 subjects having data in all these modalities. Figure 2 shows some sample fingerprint and iris images from this dataset.

1) Preprocessing and Feature Extraction: We applied the same preprocessing and feature extraction methods used in [14]. In particular, fingerprint images were enhanced using the filtering methods described in [39]. After detecting the core point [40], Gabor features were extracted around the core point and a feature vector of dimension 3600 was obtained for each fingerprint image. The iris images were segmented using the method proposed in [41] and the publicly available code

described in [42] was applied to create 25×240 iris templates. A Gabor feature of dimension 6000 was generated for every iris image.

2) Experiment Setup, Results and Analysis: The data instances (one instance includes six samples corresponding to six modalities) were randomly divided into four training instances per class and the remaining instances were used for testing. As a result, 876 instances were used for training and 519 instances were used for testing. The recognition result was averaged over five runs and we report the mean and standard deviation of rank one recognition accuracy. The rank one recognition results comparing the proposed methods with other feature-level multi-modal fusion methods are shown in Table I and Table II for each modality alone and the case of fused modalities, respectively. RMRLRJS-C shows the best recognition performance and the corresponding parameters λ_1, λ_2, λ_3, $r\alpha_V$, α_Z are set equal to 0.0004, 0.0006, 0.0007, 0.0004, 0.006 respectively.

From the results shown in Table I and Table II, we make the following observations. (1) All the considered methods achieve better recognition accuracy when fusing multiple modalities than using a single modality for recognition. (2) Robust formulations that include the sparse error term in the optimization step can lead to improved recognition results. (3) Compared to applying the low-rank constraint or joint sparsity constraint alone, the proposed methods that enforce both low-rank and joint sparse constraints perform better. (4) Common representation-based methods RMRLRJS-C perform slightly better than their corresponding methods without applying common representation constraints.

For the first proposed formulation (MRLRJS and RMRLRJS), the representation we seek is $\mathbf{\Gamma} = [\mathbf{\Gamma}^1, \mathbf{\Gamma}^2, \cdots, \mathbf{\Gamma}^D] \in \mathbb{R}^{m \times d}$. The advantage of this formulation is that the information from each modality is preserved in the representation matrix; the disadvantage is that a single modality may determine the low-rank and joint sparse property of the representation matrix, thus determining the overall performance. For example, if the representation of a certain modality is not low-rank or joint sparse, this modality can still determine the overall low-rank and joint sparse property of the overall representation matrix and as a result, we may get a poor performance.

For the second proposed formulation (MRLRJS-C and RMRLRJS-C), the representation is the same for all the D modalities, i.e. $\mathbf{\Gamma} = \mathbf{\Gamma}^1 = \mathbf{\Gamma}^2 = \cdots = \mathbf{\Gamma}^D$. The advantage of this formulation is that it satisfies the low-rank and joint sparse constraint more easily and it is more robust as each modality contributes partially to the same representation and no modality can determine the overall representation alone; the disadvantage is that it loses some discriminative information since only a single representation is enforced for all modalities.

Therefore, for this dataset in which the performance of each modalities is at the same level, both the proposed methods work. However, due to the advantage and disadvantage of common representation (second formulation), RMRLRJS-C works only slightly better than RMRLRJS.

B. UMDAA-01 Active Authentication Dataset

The UMD mobile active authentication dataset [16], [17] is a bimodal dataset consisting of face images and screen touch data collected from 50 users while they were playing with a data collection App on iPhone5s. The goal of the active authentication research is to study possible physiological and behavioral traits to continuously authenticate users while using mobile devices without interrupting users' interaction with the devices. Despite the fact that a face image is a strong biometric for verifying the user's identity, touch gestures, the way users swipe their fingers on the touchscreen of the mobile device, has shown to be a promising behavioral biometric trait for authentication. More details can be found in [43] and [17].

While users played with the App, their touch data sensed by the screen and face images acquired by the front-facing camera were simultaneously captured. The users were asked to interact with the App in 3 different sessions with different ambient conditions, namely in a well-lit room, in a dim-lit room, and in a room with natural daytime illumination. The goal was to simulate real-world scenarios to study how ambient changes can influence the performance of authenticating users using face and touch gestures. During data collection, users were free to use the phone in either orientation mode and hold the phone in any position of their choice.

Since face data were collected in an unconstrained manner, many faces exhibit different poses, rotations and extreme illuminations. In particular, partial faces are common in this dataset. Figures 3 shows sample face images from this dataset. Each row shows images from a particular ambient condition and each column shows a randomly selected user. For screen touch data, user-intra variation is large because of the unconstrained data acquisition strategy. The trajectories of some randomly selected raw touch swipes are shown in Figure 4.

1) Preprocessing and Feature Extraction: We used the preprocessing and feature extraction steps for the face images suggested in [15]. In particular, the landmarks of face images were detected using the tree-based landmarks detector [44], then face images were then cropped and aligned based on the landmarks' locations by applying the method in [45]. Illumination normalization [46] was applied to the cropped face images. Finally the face images were rescaled to dimension $192 \times 168 \times 3$ and converted to grayscale images. After preprocessing, we down sampled the preprocessed face images to 24 by 21 and simply used the whole image as a feature vector of dimension 504.

Every touch swipe $\mathbf{S}$ was encoded as a sequence of vectors

$$\mathbf{s}_i = (x_i, y_i, t_i, A_i, o_i^{ph}),$$

$i \in \{1, \cdots, N_c\}$ where x_i, y_i are the location points, t_i is the time stamp, A_i is the area occluded by the finger and o_i^{ph} is the orientation of the phone (e.g. landscape or portrait). Given these touch data, we extracted a 27-dimensional feature vector for every single swipe by using the method described in [47].

2) Experimental Setup, Results and Analysis: In order to evaluate the proposed multi-modal fusion methods, we sampled a subset from this dataset. For each user in each of the three sessions, thirty face images and thirty touch swipes

Methods	Finger 1	Finger 2	Finger 3	Finger 4	Iris 1	Iris 2
CCMM	67.8 ± 1.2	86.9 ± 1.1	69.4 ± 1.9	89.3 ± 1.6	60.5 ± 1.7	61.2 ± 0.9
SMBR-WE	68.1 ± 1.1	88.4 ± 1.2	69.2 ± 1.5	87.5 ± 1.5	60.0 ± 1.5	62.1 ± 0.4
SMBR-E	67.1 ± 1.0	87.9 ± 0.8	67.4 ± 1.9	86.9 ± 1.5	62.5 ± 1.2	64.3 ± 1.0
MCSR	70.3 ± 1.0	90.1 ± 0.8	69.2 ± 2.3	89.5 ± 1.4	62.6 ± 1.8	64.6 ± 1.0
RMCSR	69.8 ± 1.4	89.4 ± 1.0	69.2 ± 2.3	89.2 ± 1.1	**70.5 ± 1.1**	**71.7 ± 0.5**
MLRR	70.0 ± 1.8	90.0 ± 0.9	68.3 ± 1.8	89.6 ± 1.4	59.0 ± 1.8	60.1 ± 0.8
RMLRR	**70.4 ± 1.5**	89.8 ± 1.0	68.8 ± 2.1	89.9 ± 1.9	63.0 ± 1.4	65.2 ± 0.6
MCLRR	68.5 ± 1.9	88.8 ± 1.2	67.5 ± 1.5	88.5 ± 1.6	56.5 ± 1.4	58.8 ± 0.6
RMCLRR	68.5 ± 1.5	88.3 ± 1.1	67.0 ± 1.6	87.9 ± 1.7	58.7 ± 1.0	60.1 ± 0.6
MRLRJS	69.7 ± 1.1	89.7 ± 1.3	70.6 ± 1.6	90.4 ± 0.6	59.6 ± 1.0	61.0 ± 0.4
RMRLRJS	68.6 ± 1.3	89.3 ± 1.1	69.0 ± 2.0	89.0 ± 1.4	63.5 ± 1.1	64.6 ± 1.0
MRLRJS-C	69.5 ± 0.9	90.0 ± 1.0	70.1 ± 1.6	90.4 ± 0.5	59.1 ± 0.8	60.6 ± 0.5
RMRLRJS-C	70.1 ± 1.8	**90.1 ± 0.3**	**71.2 ± 1.3**	**90.5 ± 0.1**	69.5 ± 1.3	69.8 ± 0.6

TABLE I: Rank one recognition accuracy (in %) for WVU biometric multi-modal dataset for individual modality.

Methods	4 Fingerprints	2 Irises	All Modalities
MKL	86.2 ± 1.2	76.8 ± 2.5	89.8 ±0.9
CCMM	**98.9**± 0.5	82.9 ± 1.4	99.6 ±0.2
SMBR-WE	97.9 ± 0.4	76.5 ± 1.6	98.7 ±0.2
SMBR-E	97.6 ± 0.6	78.2 ± 1.2	98.6 ±0.5
MCSR	95.6 ± 0.4	78.3 ± 0.2	98.2 ±0.4
RMCSR	96.1 ± 0.6	85.3 ± 1.9	99.4 ±0.5
MLRR	98.7 ± 0.6	74.0 ± 0.9	98.9 ±0.4
RMLRR	98.7 ± 0.5	78.2 ± 1.2	99.1 ±0.4
MCLRR	96.0 ± 0.4	74.9 ± 1.7	98.6 ±0.5
RMCLRR	96.5 ± 0.2	77.0 ± 1.6	99.4 ±0.5
MRLRJS	98.5 ± 0.7	75.9 ± 0.9	99.0 ± 0.2
RMRLRJS	98.2 ± 0.5	78.6 ± 1.7	99.2 ± 0.1
MRLRJS-C	96.0 ± 0.6	76.2 ± 2.12	99.0 ± 0.7
RMRLRJS-C	96.6 ± 0.2	**85.6 ± 1.7**	**99.8 ± 0.1**

TABLE II: Rank one recognition accuracy (in %) for the WVU multi-modal biometric dataset for fusion of different modalities.

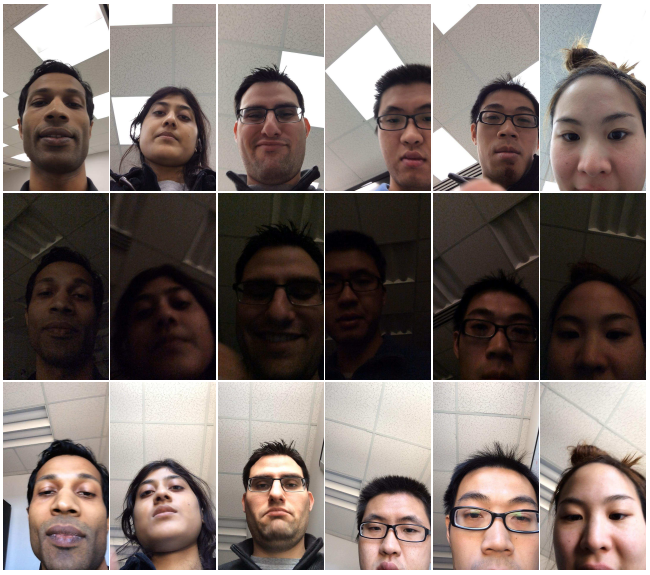

Fig. 3: Examples of face images in Active Authentication dataset. Each row shows face images collected from a mobile device in a particular ambient condition. Images in each column correspond to the same individual. It can be seen that images from different ambient conditions show very different characteristics.

were randomly selected and the resulting subset consists of 4500 face images and 4500 touch swipes corresponding to fifty users across three sessions. We selected 10, 15 and 20 instances for each user to form the training data, and used the remaining data for testing. In total, there are 500, 750, and 1000 instances for training and 4000, 3750 and 3500 instances for testing. Each instance contains a 504-dimensional feature vector for the face image and a 27-dimensional feature vector for screen touch gestures. By randomly splitting data for training and testing, we repeated each experiment ten times and report the mean and standard deviation of the rank one recognition accuracy.

The reason why we chose a small fraction of data for training is because in active authentication, the matching algorithm is supposed to work on mobile devices nearly in real-time. Our algorithm calculates the representation (either sparse or low rank or both) using the training samples, thus more training samples means high-dimensional representation and high computational cost, which should be tuned carefully in order to achieve a balance between performance and speed.

The experimental results comparing our proposed methods with the other fusion methods are shown in Table III, Table IV, and Table V respectively, corresponding to 10, 15 and 20 training instances for each user. MRLRJS-C shows the best recognition performance and the corresponding parameters λ_1, λ_2, $r\alpha_V$, α_Z are set equal to 0.0014, 0.0001, 0.45, 0.001, respectively.

Methods	Face	Touch	Face & Touch
MKL	72.58 ± 1.08	**36.02 ± 0.49**	75.13 ±2.22
CCMM	76.87 ± 1.18	33.54 ± 1.71	79.25 ±1.39
SMBR-WE	75.37 ± 1.13	30.40 ± 1.59	66.69 ±0.78
SMBR-E	73.05 ± 1.29	27.72 ± 1.50	64.49 ±1.61
MCSR	78.23 ± 0.98	28.44 ± 1.27	78.50 ±0.87
RMCSR	78.38 ± 0.87	27.72 ± 1.50	78.44 ±0.87
MLRR	76.04 ± 0.92	21.95 ± 1.41	69.24 ±0.85
RMLRR	75.94 ± 1.16	21.88 ± 1.35	69.21 ±1.17
MCLRR	75.49 ± 1.03	22.02 ± 1.37	78.58 ±1.21
RMCLRR	72.72 ± 1.49	21.88 ± 1.34	77.93 ±1.35
MRLRJS	77.36 ± 1.19	31.09 ± 1.61	68.96 ±0.86
RMRLRJS	77.15 ± 0.98	28.82 ± 1.64	63.74 ±1.04
MRLRJS-C	**80.28 ± 1.01**	23.85 ± 1.57	**81.94 ± 1.09**
RMRLRJS-C	78.77 ± 1.05	24.95 ± 1.56	81.15 ± 1.05

TABLE III: Rank one recognition accuracy (in %) for different fusion methods using ten samples from each user for training.

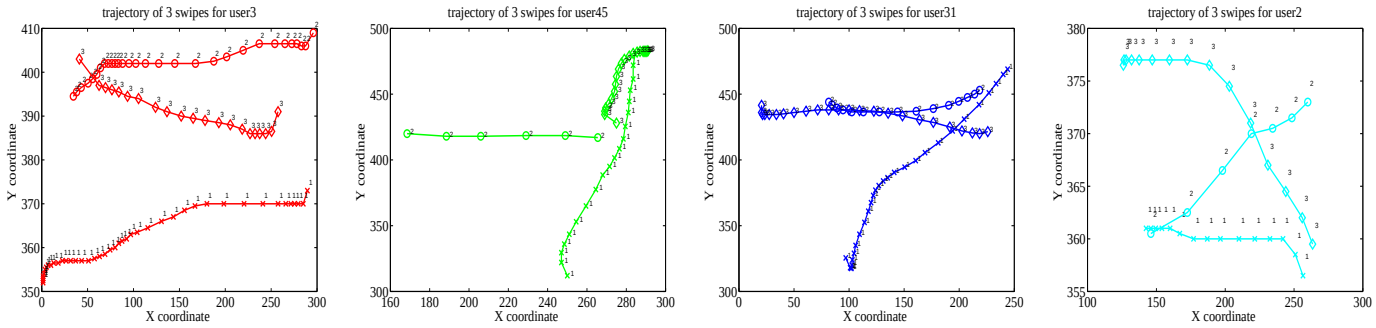

Fig. 4: Touch data corresponds to four different individuals performing the same task. The figure is best viewed in color and 200% zoom in. It is interesting to see that even for the same task touch data of different users show significant differences. This is a clue that touch data might be useful to authenticate different users.

Methods	Face	Touch	Face & Touch
MKL	77.23 ± 0.57	**39.19 ± 1.25**	80.80 ±1.22
CCMM	79.78 ± 0.61	37.27 ± 1.11	83.16 ±1.03
SMBR-WE	81.44 ± 0.49	32.42 ± 1.13	74.31 ±1.10
SMBR-E	79.12 ± 0.61	30.18 ± 1.22	71.90 ±1.36
MCSR	83.71 ± 0.47	29.79 ± 1.14	84.95 ±0.49
RMCSR	83.96 ± 0.45	29.93 ± 1.14	85.02 ±0.43
MLRR	81.04 ± 0.60	23.26 ± 1.57	75.82 ±1.06
RMLRR	81.19 ± 0.63	23.27 ± 1.69	76.28 ±1.06
MCLRR	80.60 ± 0.52	23.26 ± 1.58	83.68 ±0.53
RMCLRR	79.19 ± 0.72	23.27 ± 1.65	83.75 ±0.66
MRLRJS	83.09 ± 0.61	32.64 ± 1.18	76.08 ±1.02
RMRLRJS	81.54 ± 0.63	31.21 ± 1.34	71.28 ±0.99
MRLRJS-C	**85.47 ± 0.54**	24.78 ± 1.43	**87.45 ± 0.58**
RMRLRJS-C	84.44 ± 0.38	25.94 ± 1.28	87.26 ± 0.46

TABLE IV: Rank one recognition accuracy (in %) for different fusion methods using fifteen samples from each user for training.

Methods	Face	Touch	Face & Touch
MKL	78.36 ± 0.94	**41.48 ± 0.56**	82.20 ±0.61
CCMM	83.29 ± 0.71	40.15 ± 1.03	87.54 ±0.72
SMBR-WE	85.83 ± 0.66	32.71 ± 0.99	74.64 ±0.85
SMBR-E	87.47 ± 0.66	28.61 ± 1.45	74.88 ±1.00
MCSR	87.06 ± 0.64	29.07 ± 1.07	88.49 ±0.95
RMCSR	87.11 ± 0.71	29.08 ± 1.16	88.48 ±0.56
MLRR	87.67 ± 0.70	23.35 ± 0.99	78.94 ±0.78
RMLRR	88.02 ± 0.82	23.52 ± 1.07	79.65 ±0.86
MCLRR	87.44 ± 0.73	23.41 ± 1.10	89.33 ±0.61
RMCLRR	86.69 ± 0.85	23.61 ± 1.11	89.60 ±0.85
MRLRJS	86.30 ± 0.74	33.97 ± 1.13	80.66 ± 0.86
RMRLRJS	85.64 ± 0.78	32.01 ± 1.19	75.80 ± 0.88
MRLRJS-C	**88.58 ± 0.60**	26.78 ± 1.17	90.42 ± 0.54
RMRLRJS-C	87.57 ± 0.68	26.64 ± 1.11	**90.45 ± 0.62**

TABLE V: Rank one recognition accuracy (in %) for different fusion methods using twenty samples from each user for training.

From the results shown in Tables III, IV and V, we make the following observations: (1) Face modality works much better than touch modality. (2) As we increase the number of training samples, we observe consistent performance for each fusion method. (3) Methods without enforcing common representation fails to generate better performance for face and touch modality than for a single modality alone. On the contrary, methods that enforce the common representation (MCSR, RMCSR, MCLRR, RMCLRR, MRLRJS-C, RMRLRJS-C) successfully fuse two modalities.

In this dataset, faces (strong modality) as physical biometrics are more robust and reliable while screen touch gestures (weak modality), as a kind of behavioral biometric, exhibit more variations and can change more easily. The performance of face modality and touch modality differs a lot. For fusion methods enforcing the common representation, it is more robust as each modality contributes partially to the same representation and no modality can determine the overall representation alone. Therefore, it can successfully fuse two modalities even in the presence of weak modality. However, for fusion methods that do not enforce the common representation, the weak modality can significantly influence the quality of the overall representation and lead to worse performance when fusing two modalities compared to using the face modality alone.

C. Pascal-Sentence Dataset

Pascal-Sentence dataset is a bi-modal dataset consisting of two modalities, i. e image and sentences describing the image [19]. The images are chosen from the PASCAL VOC 2008 Challenge, which is a benchmark dataset for object recognition and detection. Thousand images were randomly selected from twenty classes. Each image was annotated with five sentences using Amazon's Mechanical Turk. Samples images and the corresponding sentences from this dataset are shown in Figure 5.

1) Preprocessing and Feature Extraction: We followed the same feature extraction method as described in [29]. Specifically, the image features are collections of responses from a variety of detectors, image classifiers and scene classifiers. The semantic features were constructed by using word-net semantic with a dictionary of 1200 words. The details of feature extraction for both modalities are described in [48]. These low-level features were then converted to binary codes using the methods described in [31]. The binary codes were then used to evaluate the performance of various feature-level fusion methods.

2) Experiment Setup, Results and Analysis: Following the experimental setup in [29], we randomly chose 500 samples

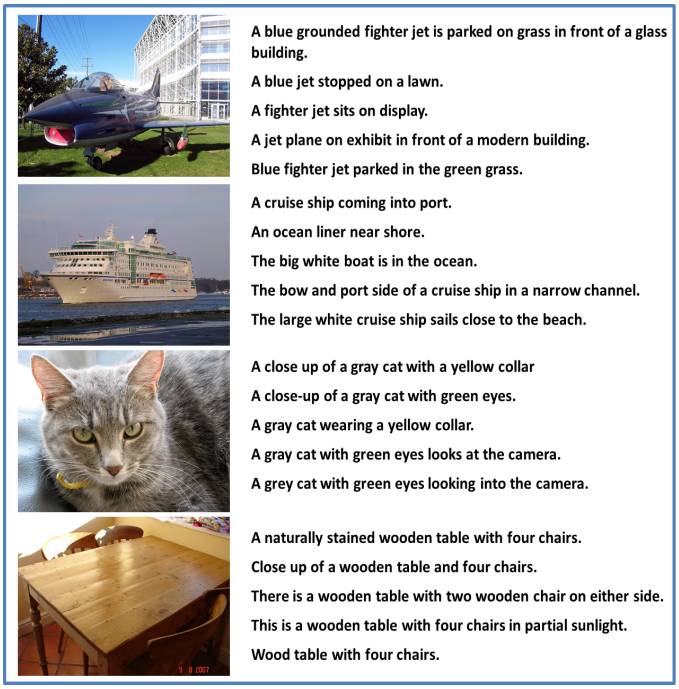

Fig. 5: Sample images and corresponding sentences from the Pascal-Sentence dataset.

for training and kept the remaining 500 samples for testing. We repeated this process five times and report the final accuracy in terms of mean and standard deviation (std) in TableVI. Note that the results of the other methods are directly copied from [29] which essentially follows the same protocol but does not report the std values. RMRLRJS and MRLRJS show the best recognition performance and the corresponding parameters λ_1, λ_2, λ_3, $r\alpha_V$, α_Z, are set equal to 0.5, 1, 0.5, 0.5, 1 respectively.

Methods	Intensity Features	Semantic Features	Fusion
MKL	67.2	64.4	76
CCMM	66.2	63.2	77.2
SMBR	66.2	69.6	75.4
MRLRJS	$\mathbf{75.5 \pm 0.2}$	$\mathbf{77.7 \pm 0.1}$	$\mathbf{82.7 \pm 0.3}$
RMRLRJS	$\mathbf{75.5 \pm 0.2}$	$\mathbf{77.7 \pm 0.1}$	$\mathbf{82.7 \pm 0.3}$
MRLRJS-C	75.0 ± 0.2	74.6 ± 0.5	81.1 ± 0.6
RMRLRJS-C	75.0 ± 0.2	74.6 ± 0.5	81.1 ± 0.6

TABLE VI: Classification accuracy (in %) for the Pascal-Sentence dataset.

From the results shown in Tables VI, we make the following observations: (1) The performance of each modality is about the same. (2) The robust version of each formulations (RMRLRJS, EMRLRJS-C) does not yield better performance than their corresponding basic version (MRLRJS, MRLRJS-C). (3) Enforcing the common representation does not yield better performance.

In this dataset, since the performance of each modality is similar, both formulations perform comparably. The proposed formulation that enforces the common representation does not show better results because we get a more robust representation at the cost of losing (discriminative) information. Also, the robust version of each formulation does not show significant performance improvement because of the fact that the original low-level features were converted into binary codes which are already robust to sparse errors.

D. Low-Rank versus Joint Sparsity

To study the relative contributions of low-rank constraint and joint sparse constraint, we vary the parameter r from 0 to 1 in the increments of 0.1 and plot the mean rank one recognition accuracy for RMRLRJS-C. When $r = 0$, our method reduces to RMCSR and when $r = 1$ the proposed method reduces to RMCLRR. Figure 6 shows the performance change of RMRLRJS-C under different values of r. This figure clearly illustrates the advantage of enforcing low-rank and joint sparsity constraints together over enforcing just low-rank or joint sparsity constraint alone.

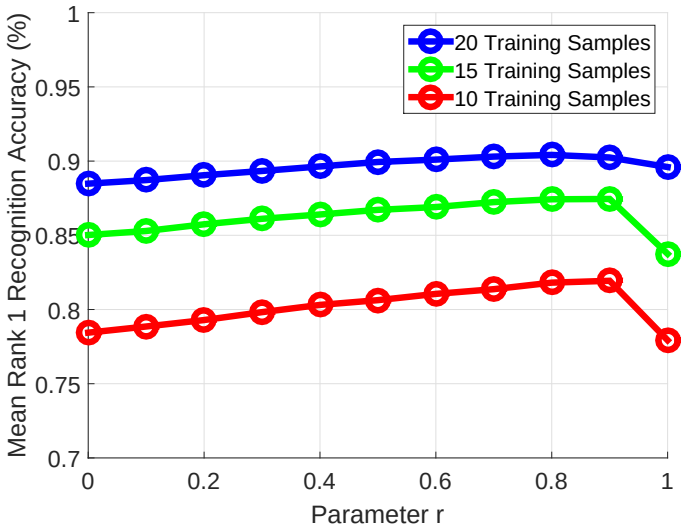

Fig. 6: Mean rank one recognition accuracy versus the relative contribution of low-rank and joint sparsity constraint. (Results on UMDAA-01 dataset)

E. Weighted vs Non-Weighted Classification

We applied the weighted reconstruction error to assign a given test instance after solving the (common) low-rank and joint sparse representation. To empirically compare these two classification strategies, we applied non-weighted classification using the same representation obtained by the proposed methods on the three datasets and report the recognition. As shown in Table VII, the weighted classification rule provides no worse results than those obtained by non-weighted classification.

Dataset	Non-Weighted	Weighted
WVU	99.80	99.80
UMDAA-01	89.51	90.45
Pascal-Sentence	81.48	82.72

TABLE VII: Rank one recognition accuracy (in %) for weighted and non-weighted classification on three datasets.

VII. Complexity Analysis

To analyze the computational complexity of the proposed methods, we look at each step in the algorithm. For simplicity, we assume the number of modalities is D, the number of classes is C, the dimension of the feature vector from different modality is n, the number of training samples is m, the number of iterations is k and the number of observations from different modality in one test sample is p. D and p are usually much smaller than C, m and n. k depends on how quickly the algorithm can converge.

In general, the complexity of matrix multiplication is $\mathcal{O}(n^3)$ and the complexity of matrix addition is $\mathcal{O}(n^2)$ for two $n \times n$ matrices. The complexity of matrix inversion and singular value decomposition is $\mathcal{O}(n^3)$ for an $n \times n$ matrix. For the proposed algorithm, in every iteration, the complexity of computing $\mathbf{\Gamma}$ and $\mathbf{E}$ is $\mathcal{O}(mnpD)$. Note that the matrix inversion part can be computed in advance since it is fixed. Computing $\mathbf{U}$ requires thresholding each element and its complexity is $\mathcal{O}(npD)$. Computing $\mathbf{V}$ involves singular value decomposition, singular value shrinking and matrix multiplication and their complexity is $\mathcal{O}(m^2pD)$. The complexity of computing $\mathbf{Z}$ is $\mathcal{O}(mpD)$. The complexity of computing $\mathbf{A}_V$, $\mathbf{A}_Z$, $\mathbf{A}_U$ is also $\mathcal{O}(mpD)$. Therefore, computing the coefficient matrix through k iterations requires the computations in the order of $\mathcal{O}(k(mnpD + m^2pD))$. For classifying the test sample, we need to compute the weights and reconstruction error, leading to $\mathcal{O}(mnpCD)$ complexity. Note that, the overall complexity of the proposed algorithms is the same as its special cases, even though more variables are introduced and computed.

VIII. Conclusion

We proposed joint sparsity and low-rank representation-based methods for multi-modal recognition. The second formulation further enforces the common representation across all the modalities in order to get a more robust representation. Previously proposed joint sparsity or low-rank representation-based multi-modal recognition methods are special cases of the proposed formulations. Efficient algorithms based on ADMM are derived to solve the proposed problems.

From experimental results, we can conclude that: (1) for datasets, such as the WVU dataset and the Pascal-Sentence dataset, in which the performance of each modality is about the same, there is no guarantee that enforcing the common representation (MRLRJS-C and RMRLRJS-C) may always show better results; (2) for datasets, such as UMDAA-01 dataset, in which the performance of each modality differs a lot, enforcing the common representation (MRLRJS and RMRLRJS) will successfully fuse all the modalities and perform much better than the general formulation (MRLRJS and RMRLRJS) which fail to fuse strong and weak modalities together.

Acknowledgment

This work was supported by DARPA Active Authentication Project under cooperative agreement FA8750-13-2-0279 and by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

References

[1] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[2] P. K. Varshney, "Multisensor data fusion," *Electronics and Communication Engineering Journal*, vol. 9, no. 6, pp. 245–253, 1997.

[3] A. Ross and A. K. Jain, "Multimodal biometrics: an overview," in *European Signal Processing Conference*, 2004, pp. 1221–1224.

[4] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*. Springer, 2006.

[5] S. Becker, J. Bobin, and E. J. Candès, "Nesta: A fast and accurate first-order method for sparse recovery," *SIAM J. Img. Sci.*, vol. 4, pp. 1–39, Jan 2011.

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[7] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 2080–2088.

[8] M. Golbabaee and P. Vandergheynst, "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2741–2744.

[9] E. Richard, P.-A. Savalle, and N. Vayatis, "Estimation of simultaneously sparse and low rank matrices," in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ser. ICML'12, 2012, pp. 51–58.

[10] 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, Nov 2013.

[11] 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, 2013.

[12] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognition Letters*, 2013.

[13] Y.-X. Wang, H. Xu, and C. Leng, "Provable subspace clustering: When lrr meets ssc," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 64–72.

[14] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, Jan 2014.

[15] H. Zhang, V. M. Patel, and R. Chellappa, "Robust multimodal recognition via multitask multivariate low-rank representations," in *IEEE International Conference on Automatic Face and Gesture Recognition*, vol. 1, May 2015, pp. 1–8.

[16] M. E. Fathy, V. M. Patel, and R. Chellappa, "Face-based active authentication on mobile devices," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2015, pp. 1687–1691.

[17] H. Zhang, V. M. Patel, M. Fathy, and R. Chellappa, "Touch gesture-based active user authentication using dictionaries," in *IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 207–214.

[18] S. S. S. Crihalmeanu, A. Ross, and L. Hornak, "A protocol for multibiometric data acquisition, storage and dissemination," WVU, Lane Department of Computer Science and Electrical Engineering, Tech. Rep., 2007.

[19] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 15–29.

[20] H. Zhang, V. M. Patel, and R. Chellappa, "Multitask multivariate common sparse representations for robust multimodal biometrics recognition," in *IEEE International Conference on Image Processing*, Sept 2015, pp. 202–206.

[21] A. Rattani, D. Kisku, M. Bicego, and M. Tistarelli, "Feature level fusion of face and fingerprint biometrics," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007, pp. 1–6.

[22] X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," in *International Conference on Pattern Recognition*, 2006, pp. 529–532.

[23] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, Jul. 2011.

[24] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3493–3500.
[25] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multiobservation visual recognition via joint dynamic sparse representation," in *International Conference on Computer Vision*, 2011, pp. 595–602.
[26] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, pp. 49–67, Feb 2006.
[27] M. Dao, N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Collaborative multi-sensor classification via sparsity-based representation," *IEEE Transactions on Signal Processing*, vol. 64, no. 9, pp. 2400–2415, May 2016.
[28] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.
[29] A. Shrivastava, M. Rastegari, S. Shekhar, R. Chellappa, and L. S. Davis, "Class consistent multi-modal fusion with binary features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
[30] J. S. Turek, J. Sulam, M. Elad, and I. Yavneh, "Fusion of ultrasound harmonic imaging with clutter removal using sparse signal separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 793–797.
[31] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, ser. ECCV'12, 2012.
[32] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.
[33] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, November 2010.
[34] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jan 2011.
[35] M. Elad, *Sparse and Redundant Representations: From theory to applications in Signal and Image processing*. Springer, 2010.
[36] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, Mar 2010.
[37] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, May 2011.
[38] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.
[39] S. Chikkerur, C. Wu, and V. Govindaraju, "A systematic approach for feature extraction in fingerprint images," in *Biometric Authentication, First International Conference, ICBA 2004, Hong Kong, China, July 15-17, 2004, Proceedings*, 2004, pp. 344–350.
[40] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Filterbank-based fingerprint matching," *Trans. Img. Proc.*, vol. 9, no. 5, pp. 846–859, May 2000.
[41] S. Pundlik, D. Woodard, and S. Birchfield, "Non-ideal iris segmentation using graph cuts," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, June 2008, pp. 1–6.
[42] P. K. Libor Masek, "Matlab source code for a biometric identification system based on iris patterns," The School of Computer Science and Software Engineering, The University of Western Australia, 2003.
[43] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 136–148, Jan 2013.
[44] D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12, 2012, pp. 2879–2886.
[45] V. Štruc and N. Pavešić, "The complete gabor-fisher classifier for robust face recognition," *EURASIP Advances in Signal Processing*, vol. 2010, p. 26, 2010.
[46] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Proceedings of the 3rd International Conference on Analysis and Modeling of Faces and Gestures*, ser. AMFG'07, 2007, pp. 168–182.
[47] H. Zhang, V. Patel, M. Fathy, and R. Chellappa, "Touch gesture-based active user authentication using dictionaries," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, Jan 2015, pp. 207–214.
[48] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ser. ECCV'10, 2010.

Heng Zhang received the B.S. degrees in electrical engineering (Hons.) from Northwestern Polytechnical University (NWPU), Xi'an, Shaanxi, China in 2011 and the Ph.D. degree in electrical engineering from the University of Maryland College Park, MD, USA, in 2017. He is currently a Member of Technical Staff at VMware Inc. Prior to joining VMware, he was a research assistant at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His research interests include signal processing, computer vision, machine learning with application to biometrics and imaging. He is a member of IEEE.

Vishal M. Patel received the B.S. degrees in electrical engineering and applied mathematics (Hons.) and the M.S. degree in applied mathematics from North Carolina State University, Raleigh, NC, USA, in 2004 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland College Park, MD, USA, in 2010. He is currently an A. Walter Tyson Assistant Professor in the Department of Electrical and Computer Engineering (ECE) at Rutgers University. Prior to joining Rutgers University, he was a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). His current research interests include signal processing, computer vision, and pattern recognition with applications in biometrics and imaging. He has received a number of awards including the 2016 ONR Young Investigator Award, the 2016 Jimmy Lin Award for Invention, A. Walter Tyson Assistant Professorship Award, the Best Paper Award at IEEE BTAS 2015, and Best Poster Awards at BTAS 2015 and 2016. He is an Associate Editor of the IEEE Signal Processing Magazine, IEEE Biometrics Compendium, and serves on the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He is a member of Eta Kappa Nu, Pi Mu Epsilon, and Phi Beta Kappa.

Rama Chellappa is a Distinguished University Professor, a Minta Martin Professor in Engineering and Chair of the Department of Electrical and Computer Engineering at the University of Maryland, College Park, MD. He received the B.E. (Hons.) degree in Electronics and Communication Engineering from the University of Madras, India and the M.E. (with Distinction) degree from the Indian Institute of Science, Bangalore, India. He received the M.S.E.E. and Ph.D. Degrees in Electrical Engineering from Purdue University, West Lafayette, IN. At UMD, he is an affiliate Professor of Computer Science Department, Applied Mathematics and Scientific Computing Program, member of the Center for Automation Research and a Permanent Member of the Institute for Advanced Computer Studies. His current research interests span many areas in image processing, computer vision and machine learning. Prof. Chellappa is a recipient of an NSF Presidential Young Investigator Award and four IBM Faculty Development Awards. He received two paper awards and the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). He is a recipient of the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. Recently, he received the inaugural Leadership Award from the IEEE Biometrics Council. At UMD, he has received numerous college and university level recognitions for research, teaching, innovation and mentoring of undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. He received the Distinguished Alumni Award from the Indian Institute of Science in 2016. Prof. Chellappa served as the EIC of IEEE Transactions on Pattern Analysis and Machine Intelligence, as the Co-EIC of Graphical Models and Image Processing, as the Associate Editor of four IEEE Transactions, as a Co-Guest Editor of many special issues, and is currently on the Editorial Board of SIAM Jl. of Imaging Science and Image and Vision Computing. He has also served as the General and Technical Program Chair/Co-Chair for several IEEE international and national conferences and workshops. He is a Golden Core Member of the IEEE Computer Society, served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the President of IEEE Biometrics Council. He is a Fellow of IEEE, IAPR, OSA, AAAS, ACM, and AAAI and holds six patents.