Quality-based Multimodal Classification Using Tree-Structured Sparsity

Soheil Bahrampour
Pennsylvania State University
soheil@psu.edu

Asok Ray
Pennsylvania State University
axr2@psu.edu@psu.edu

Nasser M. Nasrabadi
Army Research Laboratory
nasser.m.nasrabadi.civ@mail.mil

Kenneth W. Jenkins
Pennsylvania State University
jenkins@engr.psu.edu

Abstract

Recent studies have demonstrated advantages of information fusion based on sparsity models for multimodal classification. Among several sparsity models, tree-structured sparsity provides a flexible framework for extraction of cross-correlated information from different sources and for enforcing group sparsity at multiple granularities. However, the existing algorithm only solves an approximated version of the cost functional and the resulting solution is not necessarily sparse at group levels. This paper reformulates the tree-structured sparse model for multimodal classification task. An accelerated proximal algorithm is proposed to solve the optimization problem, which is an efficient tool for feature-level fusion among either homogeneous or heterogeneous sources of information. In addition, a (fuzzy-set-theoretic) possibilistic scheme is proposed to weight the available modalities, based on their respective reliability, in a joint optimization problem for finding the sparsity codes. This approach provides a general framework for quality-based fusion that offers added robustness to several sparsity-based multimodal classification algorithms. To demonstrate their efficacy, the proposed methods are evaluated on three different applications – multiview face recognition, multimodal face recognition, and target classification.

1. Introduction

Information fusion using multiple sensors often results in better situation awareness and decision making [7]. While the information from a single sensor is generally localized and can be corrupted, sensor fusion provides a framework to obtain sufficiently local information from different perspectives, which is expected to be more tolerant to the errors of individual sources. Moreover, the cross-correlated information of (possibly heterogeneous) sources can be used for context learning and enhanced machine perception [27].

Fusion algorithms are usually categorized into two levels: feature fusion [22] and classifier fusion [20, 23]. Feature fusion methods combine various features extracted from different sources into a single feature set, which are then used for classification. On the other hand, classifier fusion aggregates decisions from several classifiers, where each classifier is built upon separate sources. While classifier fusion has been well studied, feature fusion is a relatively less-studied problem, mainly due to the incompatibility of feature sets [21]. A naive way of feature fusion is to concatenate features into a longer one [30], which may suffer from the curse of dimensionality. Moreover, the concatenated feature does not contain the cross-correlated information among the sources [22]. However, if above limitations are mitigated, feature fusion can potentially outperform the classifier fusion [12].

Recently, sparse representation has attracted the interest of many researchers, both for reconstructive and discriminative tasks [26]. The underlying assumption is that if a dictionary is constructed with the training samples of all classes, only a few atoms of the dictionary, with the same label as the test data, should contribute to reconstruct the test sample. Feature level fusion using sparse representation has also been recently introduced and is often referred to as "multi-task learning" in which the general goal is to represent samples jointly from several tasks/sources using different sparsity priors [24, 25, 29]. In [18], a joint sparse model is introduced in which multiple observations from the same class are simultaneously represented by a few training samples. In other words, observations of a single scenario from different modalities would generate the same sparsity pattern of representing coefficients, which lies in a low-dimensional subspace. Similarly, modalities are fused with a joint sparsity model in [18] and [24] for target classification and biometric recognition, respectively. Joint sparsity model relies on the fact that *all* the different sources

share the same sparsity patterns at atom level, which is not necessarily true and may limit its applicability.

Another proposed solution is to group the relevant (correlated) tasks together and seek common sparsity within the group only [9] or allowing small collaboration between different groups [16]. A more generalized approach is proposed in [11] for multi-task regression in which different tasks can be grouped in a tree-structured sparsity providing flexibility in fusion of different sources. Although the formulation of tree-structured sparsity proposed in [11] provides a framework to model different sparsity structures among multiple tasks, the proposed optimization algorithm only solves an approximation of the formulation and therefore cannot enforce the desired sparsity within different groups and sparsity can only be achieved within each task, separately. Moreover, in all the discussed multimodal fusion algorithms, including tree-structured sparsity, different modalities are assumed to have equal contributions for classification task. This can significantly limits the performance of fusion algorithms in dealing with occasional perturbation or malfunction of individual modalities. In [24], a quality measure based on the joint sparse representation is introduced to quantify the quality of the data from different modalities. However, this index is measurable only after the sparse codes are obtained. Moreover, it measures the sparsity level of the representing coefficients which does not necessarily reflect the quality of individual modalities.

The major contributions of the paper are as follows: (i) *Reformulation and efficient optimization of the tree-structured sparsity for multimodal classification*: A finite number of separated problems are efficiently solved using the proximal algorithm [8] to provide an exact solution to the tree-structured sparse representation. The proposed learning facilitates feature level fusion of homogeneous/heterogeneous sources at multiple granularities. (ii) *Quality-based fusion*: A (fuzzy-set-theoretic) possibilistic approach [10, 15] is proposed to quantify the quality of different modalities in joint optimization with the reconstruction task. The proposed framework can be integrated with different sparsity priors (*e.g.* joint sparsity or tree-structured sparsity) for quality-based fusion. The proposed method places larger weights on those modalities which have smaller reconstruction errors. (iii) *Improved performance for multimodal classification*: The improved performances and robustness of the proposed algorithms are illustrated on three applications – multiview face recognition, multimodal face recognition, and target classification.

The rest of the paper is organized as follows. In Section 2, after briefly reviewing the joint sparsity model, multimodal tree-structured sparsity is reformulated and solved using the proximal algorithm. Section 3 proposes the quality-based fusion which is followed by comparative studies in Section 4 and conclusions in Section 5.

2. Multimodal sparse representation

This section reviews the joint sparse representation classifiers and reformulates the tree-structured sparsity model [11] as a multimodal classifier. A proximal algorithm is then proposed to solve the associated optimization.

2.1. Joint sparse representation classification

Let $\mathcal{S} \triangleq \{1, \ldots, S\}$ be a finite set of available modalities used for multimodal classification and C be the number of different classes in the dataset. Let the training data consist of $N = \sum_{c=1}^{C} N_c$ training samples from S modalities, where N_c is the number of training samples in the c^{th} class. Let $n^s, s \in \mathcal{S}$, be the dimension of the feature vector for the s^{th} modality and $\boldsymbol{x}_{c,j}^s \in \mathbb{R}^{n^s}$ denote the j^{th} sample of the s^{th} modality belonging to the c^{th} class, where $j \in \{1, \ldots, N_c\}$. In JSRC, S dictionaries $\boldsymbol{X}^s \triangleq [\boldsymbol{X}_1^s \boldsymbol{X}_2^s \ldots \boldsymbol{X}_C^s] \in \mathbb{R}^{n^s \times N}, s \in \mathcal{S}$, are constructed from the (normalized) training samples, where the class-wise sub-dictionary $\boldsymbol{X}_c^s \triangleq \left[\boldsymbol{x}_{c,1}^s, \boldsymbol{x}_{c,2}^s, \ldots, \boldsymbol{x}_{c,N_c}^s\right] \in \mathbb{R}^{n^s \times N_c}$ consists of samples from the c^{th} class and s^{th} modality.

Given the test samples $\boldsymbol{y}^s \in \mathbb{R}^{n^s}, s \in \mathcal{S}$, observed by S different modalities from a single event, the goal is to classify the event. In the sparse representation classification, the key assumption is that a test sample $\boldsymbol{y}^s$ from the c^{th} class lies approximately within the subspace formed by the training samples of the c^{th} class and can be approximated (or reconstructed) from *a few* number of training samples in $\boldsymbol{X}_c^s$ [26]. That is, if the test sample $\boldsymbol{y}^s$ belongs to the c^{th} class, it is represented as:

$$\boldsymbol{y}^s = \boldsymbol{X}^s \boldsymbol{\alpha}^s + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{\alpha}^s \in \mathbb{R}^N$ is a coefficient vector whose entries are mostly 0's except for some of the entries associated with the c^{th} class, i.e., $\boldsymbol{\alpha}^s = \left[\boldsymbol{0}^T, \ldots, \boldsymbol{0}^T, \boldsymbol{\alpha}_c^T, \boldsymbol{0}^T, \ldots, \boldsymbol{0}^T\right]^T$, and $\boldsymbol{e}$ is a small error term due to imperfectness of the samples.

In addition to the above assumption, JSRC enforces collaboration among different modalities to make a joint decision, where the coefficient vectors from different modalities are forced to have the same sparsity pattern. That is, the same training samples from different modalities are used to reconstruct the test data. The coefficient matrix $\boldsymbol{A} = \left[\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^S\right] \in \mathbb{R}^{N \times S}$, where $\boldsymbol{\alpha}^s$ is the sparse coefficient vector for reconstructing $\boldsymbol{y}^s$, is recovered by solving the following ℓ_1/ℓ_q joint optimization problem with $q \geq 1$:

$$\underset{\boldsymbol{A}=[\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^S]}{\operatorname{argmin}} \; f(\boldsymbol{A}) + \lambda \|\boldsymbol{A}\|_{\ell_1/\ell_q}. \tag{2}$$

In Eq. (2), $f(\boldsymbol{A}) \triangleq \frac{1}{2} \sum_{s=1}^{S} \|\boldsymbol{y}^s - \boldsymbol{X}^s \boldsymbol{\alpha}^s\|_{\ell_2}^2$ is the reconstruction error, ℓ_1/ℓ_q norm is defined as $\|\boldsymbol{A}\|_{\ell_1/\ell_q} = \sum_{j=1}^{N} \|\boldsymbol{a}_j\|_{\ell_q}$ in which $\boldsymbol{a}_j$'s are row vectors of $\boldsymbol{A}$, and $\lambda > 0$ is a regularization parameter. The number q is usually set to 2 and thus the second term in the cost function

is refereed as ℓ_1/ℓ_2 penalty term. The above optimization problem encourages sharing of patterns across related observations so that the solution $\boldsymbol{A}$ has a common support at the column level [18], which can be obtained by using different optimization algorithms (*e.g.* alternating direction method of multipliers [28]). The proximal algorithm is used in this paper [19].

Let $\delta_c(\alpha) \in \mathbb{R}^N$ be a vector indication function in which the rows corresponding to c^{th} class are retained and the rest are set to zeros. The test data is classified using the class-specific reconstruction errors as:

$$c^* = \underset{c}{\operatorname{argmin}} \sum_{s=1}^{S} \|\boldsymbol{y}^s - \boldsymbol{X}^s \delta_c(\boldsymbol{\alpha}^{s*})\|_{\ell_2}^2 \tag{3}$$

where $\boldsymbol{\alpha}^{s*}$'s are optimal solutions of Eq. (2).

2.2. Multimodal tree-structured sparse representation classification

As discussed in Section 1, although different sources are correlated, the joint sparsity assumption of JSRC may be too restrictive for some applications in which not all the different modalities are equally correlated and stronger correlations between some groups of the modalities may exist.

Tree-structured sparsity model provides a flexible framework to enforce prior knowledge in grouping different modalities by encoding them in a tree, where each leaf node represents an individual modality and each internal node represents a grouping of its child nodes. This arrangement allows modalities to be grouped at multiple granularity [11]. Adopting the definition in [8], a tree-structured groups of modalities $\mathcal{G} \subseteq \left(2^{\mathcal{S}} \setminus \emptyset\right)$ is defined as a collection of subsets of the set of modalities $\mathcal{S}$ such that $\bigcup_{g\in\mathcal{G}} g = \mathcal{S}$ and $\forall g, \tilde{g} \in \mathcal{G}, (g \cap \tilde{g} \neq \emptyset) \Rightarrow ((g \subseteq \tilde{g}) \vee (\tilde{g} \subseteq g))$. It is assumed here that $\mathcal{G}$ is ordered according to relation $\preccurlyeq$ which is defined as $(g \preccurlyeq \tilde{g}) \Rightarrow ((g \subseteq \tilde{g}) \vee (g \cap \tilde{g} = \emptyset))$. If the prior knowledge about grouping of modalities is not available, then hierarchical clustering algorithms could be used to find the tree structure [11].

Given a tree-structured collection $\mathcal{G}$ of groups, the proposed multimodal tree-structured sparse representation classification (MTSRC) is formulated as:

$$\underset{\boldsymbol{A}=[\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^S]}{\operatorname{argmin}} \quad f(\boldsymbol{A}) + \lambda\, \Omega\left(\boldsymbol{A}\right) \tag{4}$$

where $f(\boldsymbol{A})$ is defined the same as in Eq. (2), and the tree-structured sparse model is defined as:

$$\Omega\left(\boldsymbol{A}\right) \triangleq \sum_{j=1}^{N} \sum_{g\in\mathcal{G}} \omega_g \|\boldsymbol{a}_{jg}\|_{\ell_2}. \tag{5}$$

In Eq. (5), ω_g is a positive weight for group g and $\boldsymbol{a}_{jg}$ is a $(1 \times S)$ row vector whose coordinates are equal to the j^{th} row of $\boldsymbol{A}$ for indices in the group g, and 0 otherwise.

The above optimization problem allows sharing of patterns across related groups of modalities. Thus the optimal solution $\boldsymbol{A}^*$ has a common support at the group level and the resulting sparsity is dependant on the relative weights ω_g of different groups [11]. Having obtained $\boldsymbol{A}^*$, the test samples are classified using (3). The tree-structured sparsity provides a flexible framework to enforce different sparsity priors. For example, if $\mathcal{G}$ consists of only one group, containing all modalities, then (4) reduces to that of JSRC in (2). In another example where $\mathcal{G}$ consists of only singleton sets of individual modalities, no sparsity pattern is sought among different modalities and the optimization (4) reduces to S separate ℓ_1 optimization problems.

2.3. Optimization algorithm

As discussed in Section 1, the optimization procedure proposed in [11] for tree-structured sparsity only solves an approximated version of the optimization problem (4) and, therefore, does not necessarily results in a solution with desired group sparsity. In this section, an accelerated proximal gradient method [19] is used to solve (4) in which the optimal solution is obtained by solving a finite number of tractable optimization problems without approximating the cost function. Let the initial value of $\boldsymbol{A}$, which can be chosen as an arbitrary sparse vector, be zero. Then, at k^{th} iteration, the proposed accelerated proximal gradient is as follows [19]:

$$\begin{aligned} \boldsymbol{B}^{k+1} &= \hat{\boldsymbol{A}}^k + \rho^k \left(\hat{\boldsymbol{A}}^k - \hat{\boldsymbol{A}}^{k-1}\right) \\ \hat{\boldsymbol{A}}^{k+1} &= \operatorname{prox}_{\lambda t^k \Omega}\left(\boldsymbol{B}^{k+1} - t^k \nabla f\left(\boldsymbol{B}^{\boldsymbol{k+1}}\right)\right) \end{aligned} \tag{6}$$

where t^k is the step size at time step k which can be set as a constant or be updated using a line search algorithm [19]; $\hat{\boldsymbol{A}}^k$ is the estimation of the optimal solution $\boldsymbol{A}$ at time step k; the extrapolating parameter ρ^k could be chosen as $\frac{k}{k+3}$; and the associated proximal optimization problem is defined as:

$$\operatorname{prox}_{\beta\Omega}\left(\boldsymbol{V}\right) = \underset{\boldsymbol{U}\in\mathbb{R}^{N\times S}}{\operatorname{argmin}}\ \Omega\left(\boldsymbol{U}\right) + \frac{1}{2\beta}\|\boldsymbol{U} - \boldsymbol{V}\|_F^2, \tag{7}$$

where $\|.\|_F$ is the Frobenius norm . Using Eq. (5), the proximal optimization problem is reformulated as:

$$\begin{aligned} &\operatorname{prox}_{\beta\Omega}\left(\boldsymbol{V}\right) = \\ &\underset{\boldsymbol{U}\in\mathbb{R}^{N\times S}}{\operatorname{argmin}} \sum_{j=1}^{N} \left(\sum_{g\in\mathcal{G}} \omega_g \|\boldsymbol{u}_{jg}\|_{\ell_2} + \frac{1}{2\beta}\|\boldsymbol{u}_j - \boldsymbol{v}_j\|_{\ell_2}^2 \right) \end{aligned} \tag{8}$$

where $\boldsymbol{u}_{jg}$ is defined similar to $\boldsymbol{a}_{jg}$ in Eq. (5); and $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the j^{th} rows of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Consequently, the solution of (8) is obtained by N separate optimizations on S-dimensional vectors. Since the groups are defined to be ordered, each of the optimization problems can be solved

Algorithm 1 Algorithm to solve the proximal optimization step (Eq. (8)) of the accelerated proximal gradient method (Eq. (6)) corresponding to the MTSRC optimization problem (Eq. (4))

Input: $V \in \mathbb{R}^{N\times S}$, ordered set of groups $\mathcal{G}$, weights ω_g for each group $g \in \mathcal{G}$, and scaler β.
Output: $U \in \mathbb{R}^{N\times S}$

```
for j = 1, ..., N do
    Let η^1 = ··· = η^{|G|} = 0 and u_j = v_j.
    for g = g_1, g_2, ··· ∈ G do
        u_j = v_j − Σ_{h≠g} η^h.
        η^g = { u_jg / ||u_jg||_{ℓ2},   if ||u_jg||_{ℓ2} > βω_g
              { u_jg,                   if ||u_jg||_{ℓ2} ≤ βω_g
    end for
    u_j = v_j − Σ_{g∈G} η^h.
end for
```

in a single iteration using the dual form [8]. Therefore, the proximal step of the tree-structured sparsity can be solved with the same computational cost as that of joint sparsity. Algorithm 1, which is a direct extension of the optimization algorithm in [8], solves the proximal step. It should be noted that the computational complexity of the optimization algorithm grows linearly as the number of training samples increases. One can potentially learn the dictionaries with (significantly) fewer number of atoms using dictionary learning algorithms [13]. This paper uses the Sparse Modeling Software [8] to solve the proximal step.

3. Weighted scheme for quality-based fusion

In most of the sparsity-based multimodal classification algorithms, including JSRC and MTSRC, it is inherently assumed that available modalities contribute equally. This may significantly limit the performance in dealing with occasional perturbation or malfunction of individual sources. Ideally, a fusion scheme should *adaptively* weight the modalities based on their reliabilities. In [24], a quality measure is introduced for JSRC, where a sparsity concentration index is calculated to quantify the quality of modalities. The main limitation of this approach, however, is that the index is obtained only after the sparse codes are calculated and a weak modality may hurt the performances of other modalities due to the enforced sparsity priors. Moreover, the index is defined based on the sparsity levels and does not necessarily reflect the quality of each modalities. This paper proposes to find the adaptive quality of each modality and sparse codes jointly in a single optimization problem.

Let μ^s be the quality weight for the s^{th} modality. A weighted scheme for multimodal reconstruction, with similar structure to Eq. (4), is proposed as follows:

$$\underset{\boldsymbol{A}=[\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^S],\mu^s}{\operatorname{argmin}} \sum_{s=1}^{S} \frac{(\mu^s)^m}{2} \|\boldsymbol{y}^s - \boldsymbol{X}^s\boldsymbol{\alpha}^s\|_{\ell_2}^2 + \Psi(\boldsymbol{A}), \quad (9)$$

with the constraint $\mu^s \geq 0,\ \forall s \in \mathcal{S}$, where the exponent $m \in (1,\infty)$ is a fuzzifier parameter, similar to formulation of fuzzy c-means clustering [3]; and $\Psi(\boldsymbol{A})$ enforces desired sparsity priors within the modalities (*e.g.* ℓ_1/ℓ_2 constraint in JSRC or tree-structured sparsity prior of MTSRC).

Another constraint on μ^s is necessary to avoid a degenerate solution of Eq. (9). A constraint such as $\sum_{s=1}^{S}\mu^s = 1$ is apparently feasible; however, since $m > 1$ in Eq, (9), the larger weight of a modality compared to those of other modalities may effectively increase this weight close to 1 while forcing the rest of the weights toward 0. To alleviate this problem, a "possibility"-like constraint, similar to the possibilistic fuzzy c-means clustering [1, 15], is proposed to allow the weights of different modalities to be specified independently. The proposed composite optimization problem to achieve quality-based multimodal fusion is posed as:

$$\underset{\boldsymbol{A}=[\boldsymbol{\alpha}^1,\ldots,\boldsymbol{\alpha}^S],\mu^s}{\operatorname{argmin}} \left(\sum_{s=1}^{S} \frac{(\mu^s)^m}{2} \|\boldsymbol{y}^s - \boldsymbol{X}^s\boldsymbol{\alpha}^s\|_{\ell_2}^2 + \Psi(\boldsymbol{A}) + \sum_{s=1}^{S} \lambda_{\mu^s}(1-\mu^s)^m \right), \mu^s \geq 0, \forall s \in \mathcal{S} \quad (10)$$

where λ_{μ^s} are the regularization parameters for $s \in \mathcal{S}$. After finding optimal $(\mu^s)^*$ and $\boldsymbol{A}^*$, the test samples are classified using the weighted reconstruction errors, i.e.,

$$c^* = \underset{c}{\operatorname{argmin}} \sum_{s=1}^{S} (\mu^s)^m \|\boldsymbol{y}^s - \boldsymbol{X}^s\delta_c(\boldsymbol{\alpha}^{s*})\|_{\ell_2}^2. \quad (11)$$

The optimization problem in (10) is not jointly convex in $\boldsymbol{A}$ and μ^s. A sub-optimal solution can be obtained by alternating between the updates of $\boldsymbol{A}$ and $\{\mu^s\}$, minimizing over one variable while keeping the other variable fixed. The accelerated proximal gradient algorithm in (6) with $f(\boldsymbol{A}) = \frac{1}{2}\sum_{s=1}^{S}(\mu^s)^m\|\boldsymbol{y}^s - \boldsymbol{X}^s\boldsymbol{\alpha}^s\|_{\ell_2}^2$ is used to optimize over $\boldsymbol{A}$ and optimal $\{\mu^s\}$ at each iteration of the alternative optimization is found in a closed form [15] as:

$$\mu^s = \frac{1}{1 + \left(\frac{\|\boldsymbol{y}^s - \boldsymbol{X}^s\boldsymbol{\alpha}^s\|_{\ell_2}^2}{\lambda_{\mu^s}}\right)^{\frac{1}{m-1}}}, s \in \mathcal{S}. \quad (12)$$

The regularization parameters λ_{μ^s} need to be chosen based on the desired bandwidth of the possibility distribution for each modality. In this paper, optimization over $\boldsymbol{A}$ is first performed without weighting scheme to find an initial estimate of $\boldsymbol{A}$. Also the following definition, similar to possibilistic clustering [1], is used to determine and fix λ_{μ^s}:

$$\lambda_{\mu^s} = \|\boldsymbol{y}^s - \boldsymbol{X}^s\boldsymbol{\alpha}^s\|_{\ell_2}^2, \quad (13)$$

resulting all μ^s to be 0.5 initially. It is observed that only a few iterations is required for the algorithm to converge. In this paper, the number of alternations and the fuzzifier m are set to be 10 and 2, respectively. Algorithm 2 summarizes the proposed quality-based multimodal fusion method. As discussed, this scheme can be used with different sparsity priors as long as optimal $\boldsymbol{A}$ can be found efficiently [19].

Algorithm 2 Quality-based multimodal fusion.

Input: Initial coefficient matrix $\boldsymbol{A}$, dictionary $\boldsymbol{X}^s$ and test sample $\boldsymbol{y}^s$ of the s^{th} modality, $s \in \{1, \cdots, S\}$, and number of iterations T.
Output: Coefficient matrix $\boldsymbol{A}$ and weights μ^s as a solution to Eq.(9).
1: Set λ_{μ^s} using Eq. (13)
2: **for** $k = 1, \ldots, T$ **do**
3: Update weights μ^s using Eq. (12)
4: Update $\boldsymbol{A}$ by solving Eq. (9) with fixed μ^s.
5: **end for**

4. Results and discussion

In this section we present the results for the proposed MTSRC and weighted scheme in three different applications: multiview face recognition, multimodal face recognition, and multimodal target classification. For MTSRC, the group weights ω_g are selected using a combination of aprior information/assumption and cross validation. We assumed equal weights for the same sized groups which reduces the number of tuning parameters. The relative weights between the groups with different sizes, however, are not fixed and are selected using cross-validation in a finite set $\{10^{-5}, 10^{-4}, ..., 10^{5}\}$. Hierarchical clustering can also be used to tune the weights [11]. It is observed that bigger groups are usually assigned with bigger weights in the studied applications. Thus MTSRC intuitively enforces collaboration among all the modalities and yet provides flexibility (compared to JSRC) by allowing collaborations to be formed at lower granularities as well. It is observed in all the studied applications that MTSRC performs similarly when the weights are varied without being reordered.

For the weighted scheme, JSRC and MTSRC are equipped with the modality weights and the resulting algorithms are denoted as JSRC-W and MTSRC-W, respectively. The performance of the proposed feature-level fusion algorithms is compared with that of several state-of-the-art decision-level and feature-level fusion algorithms. For decision-level fusion, outputs of the independent classifiers, each trained on separate modality, are aggregated by adding the corresponding probability outputs of each modality, which is denoted as *Sum*. For this purpose, SVM [4], sparse representation classifier (SRC) [26], and sparse logistic regression(SLR) [14] are used. The proposed methods are also evaluated using existing feature-level fusion methods that include holistic sparse representation classifier (HSRC) [30], JSRC [18], joint dynamic sparse representation classifier (JDSRC) [30] and relaxed collaborative representation (RCR) [29].

4.1. Multiview face recognition

In this experiment, we evaluate the performance of the proposed MTSRC for multiview face recognition using UMIST face database which consists of 564 cropped images of 20 individuals with mixed race and gender [6]. Poses of each individual are sorted from profile to frontal

Figure 1: Different poses for one individual in the UMIST database that is divided into four different view-ranges shown by four rows.

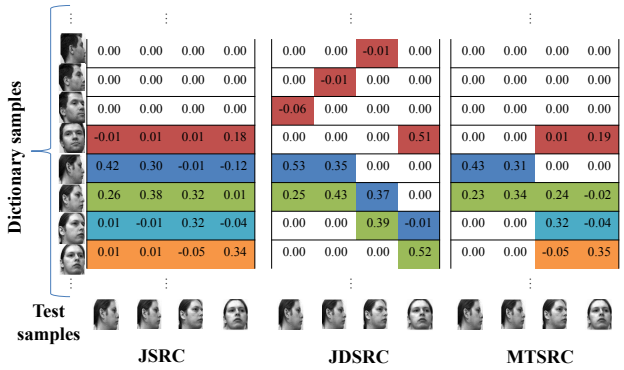

Figure 2: Representation coefficients generated by JSRC, JDSRC, and MTSRC on test observations that correspond to 4 different views on the UMIST database. JSRC enforces joint sparsity among all views and requires the same sparsity pattern at atom level. JDSRC allows contributions from different training samples to approximate a set of test observations and requires the same sparsity pattern at class level. MTSRC enforces joint sparsity only when relevant.

views and then divided into S view-ranges with equal number of images in each view-range. This facilitates multiview face recognition using UMIST database. The performance of the algorithms are tested using different values of view-ranges $S \in \{2, 3, 4\}$. It should be noted that the environment is unconstrained and captured faces may have pose variations within each view-range. Different poses for one of the individual in the UMIST database is shown in Fig. 1 where the images are divided into four view-ranges, shown in four rows. Due to limited number of training samples, a single dictionary is constructed by randomly selecting one (normalized) image per view for all the individuals in the database which is shared among different view-ranges. The rest of the images are used as the test data.

As observations from closeby views are more likely to share similar poses and be correlated, the tree structured sparsity of MTSRC is chosen to enforce group sparsity within related views. For this purpose, the tree-structured sets of groups using 2, 3, or 4 view-ranges are selected to be $\{\{1\}, \{2\}, \{1, 2\}\}$, $\{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$, $\{\{1\}, \{2\}, \{1, 2\}, \{3\}, \{4\}, \{3, 4\}, \{1, 2, 3, 4\}\}$, respectively. Fig. 2 shows the representation coefficients generated by JSRC, JDSRC, and MTSRC on a test scenario

Table 1: Multiview classification results obtained on the UMIST database.

	2 Views	3 Views	4 Views	Avg.
HSRC [30]	84.37	97.80	99.91	94.03
JSRC [18]	87.77	99.51	99.91	95.73
JDSRC [30]	86.52	98.96	99.91	95.13
MTSRC	**88.42**	**99.63**	**100.00**	**96.02**

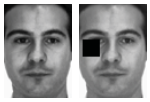

Figure 3: A test image and its occlusion in the AR dataset.

Table 2: Correct classification rates obtained using individual modalities in AR database. Modalities include left and right periocular, nose, mouth, and face.

	L periocular	R periocular	Nose	Mouth	Face
SVM	71.00	74.00	44.00	44.29	86.86
SRC	79.29	78.29	63.43	64.14	93.71

Table 3: Multimodal classification results obtained on the AR database.

Method	CCR	Method	CCR
SVM-Sum [24]	92.57	SLR-Sum [24]	77.9
JDSRC [30]	93.14	RCR [29]	94.00
JSRC [18]	95.57	JSRC-W	96.43
MTSRC	97.14	MTSRC-W	97.14

where four different view-ranges are used. As expected, JSRC enforces joint sparsity among all different views at atom level, which may be too restrictive due to relatively less correlation between frontal and profile views. JDSRC relaxes joint sparsity constraint at atom level and allows contributions from different training samples to approximate one set of the test observations but still requires joint sparsity pattern at class level. As shown, proposed MTSRC enforces joint sparsity within relevant views and also among all modalities and has the most flexible structure for multimodal classification. The results of 10-fold cross validation using different sparsity priors are summarized in Table 1. As seen, the MTSRC method achieves the best performance. Since the quality of different view-ranges are similar, JSRC-W and MTSRC-W result in similar performances as those of JSRC and MTSRC, respectively, and therefore are omitted here.

4.2. Multimodal face recognition: AR database

In this set of experiments, the performance of the proposed algorithms are evaluated on the AR database (Figure 3) which consists of faces under different poses, illumination and expression conditions, captured in two sessions [17]. A set of 100 users are used, each consisting of

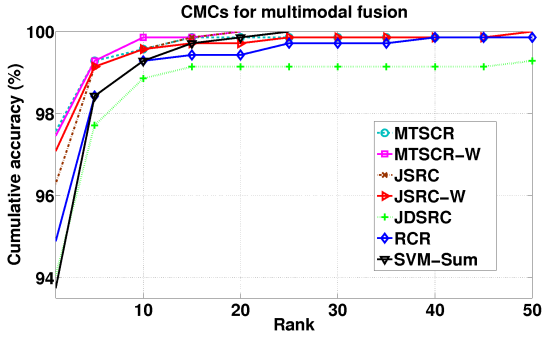

Figure 4: CMCs obtained using multimodal classification algorithms on AR database with random occlusion.

seven images from the first session as the training samples and seven images from the second session as test test samples. A small randomly selected portion of training samples, 30 out of 700, is used as validation set for optimizing the design parameters of classifiers. Fusion is taken on five modalities, similar to setup in [24], including left and right periocular, nose, and mouth regions, as well as the whole face. After resizing the images, intensity values are used as features for all modalities. Results of using individual modalities for classification using SVM and SRC algorithms are shown in Table 2. As expected, the whole face is the strongest modality in the sense of solving the identification task followed by left and right eyes. For MTSRC, the groups are chosen to be $\mathcal{G} = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\} = \{\{1\}, \{2\}, \{1, 2\}, \{3\}, \{4\}, \{5\}, \{1, 2, 3, 4, 5\}\}$, where 1 and 2 represents left and right periocular and $3, 4, 5$ represents other modalities. Weights are selected using a similar approach discussed in Section 4.3 to encourage group sparsity between all modalities and also joint representation for left and right periocular in lower granularity. The performances of the proposed algorithms are compared with several competitive methods as shown in Table 3. Fig. 4 shows the corresponding cumulative matched score curves (CMC). CMC is a performance measure, similar to ROC, which is originally proposed for biometric recognition systems [5]. As shown, the tree-structured sparsity based algorithms achieve the best performances with correct classification rate (CCR) of 97.14%.

To compare the robustness of different algorithms, each test images is occluded by randomly chosen block (See Fig. 3). Fig. 5 shows the CMC's generated when the size of occluding blocks are 15. As seen, the proposed tree-structured algorithms are the top performing algorithms. Fig. 6 compares CCR of different algorithms with block occlusion of increasing sizes. MTSRC-W has the most robust performance. Also it is observed that the weighted scheme significantly improves the performance of the JSRC.

Since the proposed weighted scheme is solved using alternating minimization, a set of experiments are performed to test its performance sensitivity to the different initializa-

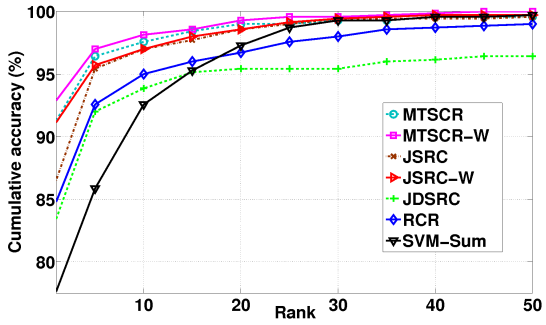

Figure 5: CMCs obtained using different multimodal classification algorithm on AR database with random occlusion.

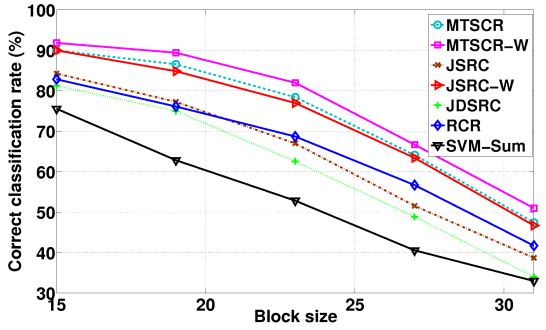

Figure 6: Correct classification rates of multimodal classification algorithms with block occlusion of different sizes.

tion of the modalities weights μ^s and regularization parameters λ_{μ^s}. In each experiment, all initial weights are set to a different value. Also all λ_{μ^s} are set to a different value, instead of being set by Eq. (13). The sparse coefficients A are initialized to zero in all the experiments. We observed similar results for relatively large variations in initial weights and regularization parameters. The performance of the weighted scheme degrades if the regularization parameters are set to be too small. On the other hand, its performance approaches that of the non-weighted scheme for large values of the regularization parameters, as expected by observing cost function (10). It is also observed that setting the regularization parameters by Eq. (13) with all the weights being initialized to $1/S$ persistently works well.

4.3. Multimodal target classification

This section presents the results of target classification on a field dataset consisting of measurements obtained from a passive infrared (PIR) and three seismic sensors of an unattended ground sensor system as discussed in [2]. Symbolic dynamic filtering is used for feature extraction from time-series data [2]. The subset of data used here consists of two days data. Day 1 includes 47 human targets and 35 animal-led-by-human targets while the corresponding numbers for Day 2 are 32 and 34, respectively. A two-way cross-validation is used to assess the performance of the classification algorithms, i.e. Day 1 data is used for training and Day 2 is used as test data and vice versa.

For MTSRC, the tree-structured set of groups is selected to be $\mathcal{G} = \{g_1, g_2, g_3, g_4, g_5, g_6\} = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}, \{4\}, \{1, 2, 3, 4\}\}$ where 1, 2 and 3 refer to the seismic channels and 4 refers to the PIR channel. Table 4 summarizes the average human detection rate (HDR), human false alarm rate (HFAR), and misclassification rates (MR) obtained using different multimodal classification algorithms. As seen, the proposed JSRC-W and MTSRC-W algorithms resulted in the best HDR and HFAR and, consequently the best overall performance. Moreover, if the different modalities are weighted equally, the MTSRC achieves the best performance.

Table 4: Results of target classification obtained by different multimodal classification algorithms by fusing 1 PIR and 3 seismic sensors data. HDR: Human Detection Rate, HFAR: Human False Alarm Rate, M: Misclassification.

	HDR	HFAR	MR
SVM-Sum	0.94	0.15	10.61%
HSRC	0.96	0.10	6.76%
JDSRC	0.97	0.09	8.11%
RCR	0.94	0.12	8.78%
JSRC	1.00	0.12	5.41%
JSRC-W	1.00	0.07	3.38%
MTSRC	1.00	0.09	4.05%
MTSRC-W	1.00	0.07	3.38%

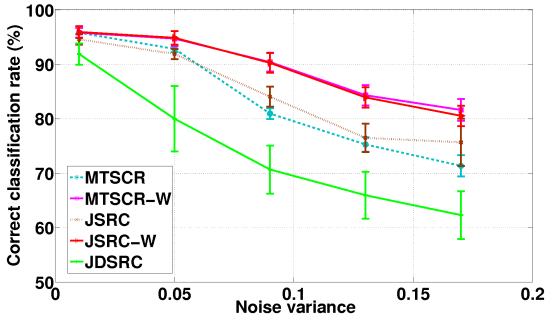

Figure 7: Correct classification rates of multimodal classification algorithms in dealing with random noise of varying variance that is added to the seismic 1 channel.

To evaluate the robustness of the proposed algorithms, random noise with varying variance is injected to the test samples of the seismic sensor 1 measurements. Fig. 7 shows the CCR obtained using different methods. The proposed weighted scheme has the most robust performance in dealing with noise. It is also seen that JSRC algorithm performs better than MTSRC as the level of noise increases. One possible reason is that in MTSRC the assumption of strong correlation between the three seismic channels are not valid with large value of noises. Fig. 8 shows averaged modality weights generated by the MTSRC-W. As expected, weight

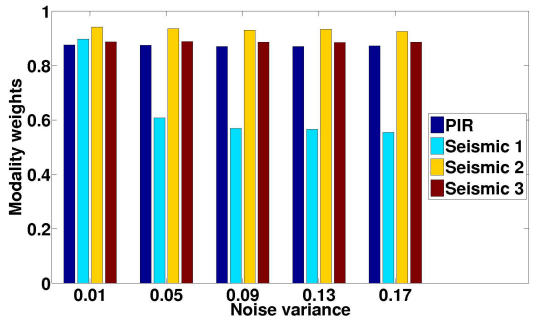

Figure 8: Averaged weights for each modality obtained by MTSRC-W algorithm when test samples from second modality, seismic sensor 1, are perturbed with random noise. As the noise level increases, the weight for second modality decreases while the corresponding weights for other unperturbed modalities remains almost constant.

of the perturbed modality decreases as noise level increases.

5. Conclusions

This paper presents the reformulation of tree-structured sparsity models for the purpose of multimodal classification and proposes an accelerated proximal gradient method to solve this class of problems. The tree-structured sparsity allows extraction of cross-correlated information among multiple modalities at different granularities. The paper also presents a possibilistic weighting scheme to jointly represent and quantify multimodal test samples by using several sparsity priors. This formulation provides a framework for robust fusion of available sources based on their respective reliability. The results show that the proposed algorithms achieve the state-of-the-art performances on the field data of three applications: multiview face recognition, multimodal face recognition, and multimodal target classification.

References

[1] S. Bahrampour, B. Moshiri, and K. Salahshoor. Weighted and constrained possibilistic c-means clustering for online fault detection and isolation. *Applied Intelligence*, 35(2):269–284, 2011. 4

[2] S. Bahrampour, A. Ray, S. Sarkar, T. Damarla, and N. M. Nasrabadi. Performance comparison of feature extraction algorithms for target detection and classification. *Pattern Recognition Letters*, 34(16):2126–2134, 2013. 7

[3] J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984. 4

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 5

[5] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. The relation between the roc curve and the cmc. In *Auto ID*, 2005. 6

[6] D. Graham and N. M. Allinson. Face recognition: From theory to applications. *NATO ASI Series F, Computer and Systems Sciences*, 163:446–456, 1999. 5

[7] D. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1), 1997. 1

[8] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, 2010. 2, 3, 4

[9] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 2

[10] A. Kendall. *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley, 1986. 2

[11] S. Kim and E. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010. 2, 3, 5

[12] A. Klausne, A. Tengg, and B. Rinner. Vehicle Classification on Multi-Sensor Smart Cameras Using Feature- and Decision-Fusion. In *ICDSC*, 2007. 1

[13] S. Kong and D. Wang. A Brief Summary of Dictionary Learning Based Approach for Classification. *arxivId:1205.6544*, 2012. 4

[14] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE TPAMI*, 27(6):957–968, 2005. 5

[15] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE TFS*, 1(2):98–110, 1993. 2, 4

[16] A. Kumar and H. Daume III. Learning task grouping and overlap in multi-task learning. *arXiv:1206.6417*, 2012. 2

[17] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998. 6

[18] N. Nguyen, N. Nasrabadi, and T. Tran. Robust multi-sensor classification via joint sparse representation. In *FUSION*, 2011. 1, 3, 5, 6

[19] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, pages 1–96, 2013. 3, 4

[20] S. Pirooz Azad, S. Bahrampour, B. Moshiri, and K. Salahshoor. New fusion architectures for performance enhancement of a pca-based fault diagnosis and isolation system. In *SAFEPROCESS*, 2009. 1

[21] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli. Feature Level Fusion of Face and Fingerprint Biometrics. In *BTAS*, 2007. 1

[22] A. Ross and R. Govindarajan. Feature Level Fusion Using Hand and Face Biometrics. In *SPIE BTHI*, 2005. 1

[23] D. Ruta and B. Gabrys. An overview of classifier fusion methods. *CIS*, 7(1):1–10, 2000. 1

[24] S. Shekhar, V. Patel, N. Nasrabadi, and R. Chellappa. Joint sparse representation for robust multimodal biometrics recognition. *IEEE TPAMI*, PP(99):1–1, 2013. 1, 2, 4, 6

[25] U. Srinivas, H. Mousavi, C. Jeon, V. Monga, A. Hattel, and B. Jayarao. SHIRC: A simultaneous sparsity model for histopathological image representation and classification. In *ISBI*, 2013. 1

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009. 1, 2, 5

[27] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang. Sensor Fusion Using Dempster-Shafer Theory. In *IMTC*, 2002. 1

[28] J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SISC*, 33(1):250–278, 2011. 3

[29] M. Yang, L. Zhang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *CVPR*, 2012. 1, 5, 6

[30] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang. Multi-observation visual recognition via joint dynamic sparse representation. In *ICCV*, 2011. 1, 5, 6