Multi-View Visual Recognition of Imperfect Testing Data

Qilin Zhang[1]
[1]Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ, USA, 07030
qzhang5@stevens.edu

Gang Hua[1,2]
[2]Microsoft Research Asia
No. 5 Danling Street,Haidian District
Beijing, P.R. China 100080
ganghua@gmail.com

ABSTRACT

A practical yet under-explored problem often encountered by multimedia researchers is the recognition of imperfect testing data, where multiple sensing channels are deployed but interference or transmission distortion corrupts some of them. Typical cases of imperfect testing data include missing features and feature misalignments. To address these challenges, we choose the latent space model and introduce a new similarity learning canonical-correlation analysis (SLCCA) method to capture the semantic consensus between views. The consensus information is preserved by projection matrices learned with modified canonical-correlation analysis (CCA) optimization terms with new, explicit class-similarity constraints. To make it computationally tractable, we propose to combine a practical relaxation and an alternating scheme to solve the optimization problem. Experiments on four challenging multi-view visual recognition datasets demonstrate the efficacy of the proposed method.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords

similarity learning; CCA; missing data; misalignment

1. INTRODUCTION

Thanks to the recent popularity of multi-sensory and multi-spectrum imaging, such as depth-sensing and infrared (IR) sensing cameras, visual recognition with multiple views of information attracts the attention of many multimedia and computer vision researchers. However, the pronounced reliability issues and data integrity concerns of multiple sensors plague the adoption of these systems in real world applications, especially in harsh or adversarial sensing environments.

In particular, this paper addresses two common obstacles in these sensing systems, i.e., the missing feature problem and the feature misalignment problem in testing data. For both problems, the system is typically trained with labeled multi-view data in advance, but the testing data could be imperfect: the secondary view can be completely missing, or both views can be randomly missing, possibly caused by sensor malfunction, predominant interferences or limited transmission bandwidth. The feature misalignment problem could arise when the semantic correspondences between two views (i.e., sensing channels) are corrupted during testing data acquisition or transmission, thus cross-view verification is required to recover these correspondences.

For these problems, the major challenge lies in how to effectively leverage the semantic consensus across all views, so as to facilitate the training of a stronger classifier that works solely on a single view, or a verification algorithm that captures the cross-view similarity information. In particular, we propose to seek a common semantic space to capture the relevance among multiple views, and further incorporate the discriminative information embedded in the class labels.

This is achieved by a new similarity learning canonical-correlation analysis algorithm, namely, the similarity learning canonical-correlation analysis (SLCCA) algorithm, inspired by [34, 35], which constructs projection matrices for a common latent space, where the intra-class and inter-class relationships are preserved. These learned projection matrices are subsequently used to discriminatively transfer all observations from all views to this common semantic space.

After obtaining the projected training samples in the common space, conventional supervised classifiers (e.g., the SVM classifier [5]) or verification algorithm (e.g., the joint Bayesian verification algorithm [6, 18, 19]) can be trained in this space. For the recognition problem, the testing data from the available view is projected onto this space before applying the aforementioned classifier[1]. For the cross-modal verification problem, both modalities are first projected onto this space before learning their pairwise similarity.

The primary contributions of this paper are: (1) we investigate two common data integrity difficulties in multi-view recognition systems (i.e., the missing feature problem and the feature misalignment problem), and propose a unified solution by leveraging the semantic consensus; (2) we propose the new SLCCA algorithm, whose performance enhancements on various recognition/verification asks are validated on four different real world datasets; (3) after formulating the problem as a generic quadratically constrained

[1] An overview of this approach is presented in Fig. 1.

MM'15, October 26–30, 2015, Brisbane, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806224.

quadratic program (QCQP), we relax it and approximate it with an alternating optimization of a linear program and a decomposition procedure, and finally efficiently solve it.

The rest of the paper is organized as follows. Related works are briefly discussed in Section 2. Section 3 formulates the proposed SLCCA algorithm, followed by Section 4 with the solution to it. Experimental datasets and settings are summarized in Section 5. Section 6 presents the experimental results, and Section 7 concludes the paper.

2. RELATED WORK

In this section, related work is summarized, then brief comparisons between these related works and the focused problem of this paper are provided.

Inspired by the human cognition and psychology studies with multi-sensory/uni-sensory experiments in [25] (where human volunteers performing the uni-sensory recognition tasks with prior multi-sensory learning significantly outperforms their counterparts with uni-sensory learning), we make the hypothesis that there should be an underlying common semantic space representing some consensus between multiple views/sensory channels. This hypothesis is capable of justifying the phenomenon in [25], and the remaining challenge is to design at least one method to recover this semantic latent space. Once this latent space is obtained, the missing data resilient multi-view recognition problem and the cross-modal verification problem can be readily solved, following Fig. 1.

In order to construct such a semantic common space, information transfer with preserved discriminative power is crucial. There are plenty of work in the transfer learning literature (e.g., [10, 24, 15]), however, they generally focus on transferring the label information from the source domain to the target domain, instead of preserving the congruent information embedded in the labeled, paired multi-view training samples.

As for preserving the cross-view consensus, the early semi-supervised co-training framework [2] and the later multi-view/multi-task learning work [11, 37, 22, 32, 1] are examples of synergistically modeling multiple views and fusing the information from all views. However, it is worth noting that there are no missing data (missing view or missing observations) involved, and there are no easy extension to address the cross-view verification problem (as defined in Section 1). A recent work [8] addresses both the missing view recognition and domain transfer problems, by imposing additional regularization terms on a multi-view classifier, but without explicitly modeling the discriminative label information.

Another line of research involves methods that also assume a common semantic space model, such as different variants of the CCA algorithms [29, 14, 13, 30, 38]. They either focus on variants of cross-view correlations considering the discriminative label information [13, 30], or incorporate the label information in an ad-hoc way [38]. In [14], the image set recognition requires abundant views, which is not always available in generic multi-view recognition problems. In [29], the second view of multi-dimensional training labels is nevertheless incompatible with generic multi-view recognition problems or missing data recognition problems. In [31], Tommasi *et al.* address a different problem (the bias between datasets) also with a latent space model.

In addition, there are previous papers addressing related problems [7, 12, 28]. They either focus only on the missing view case [7], or the random missing features case [12], or adopt a totally different learning concept such as the multimodal deep learning [28].

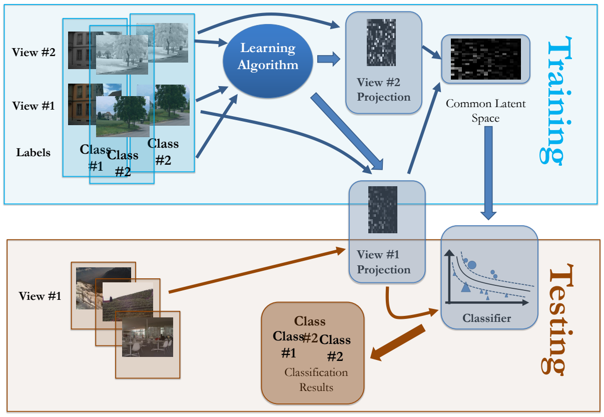


Figure 1: A latent space framework, with the training and testing processes shown in blue and yellow, respectively.

On the contrary, our proposed method directly seeks two similarity preserving projections to reveal the common semantic latent space, thus it is capable of addressing both the systematic and random missing feature problems as well as the cross-view verification problem.

3. FORMULATION

In this section, the recognition problems are first formalized, followed by the overview of the latent space model as well as the proposed SLCCA algorithm.

3.1 Problem Formalization

Let $\mathbf{X}^{(1)} \in \mathcal{R}^{p\times n}$ and $\mathbf{X}^{(2)} \in \mathcal{R}^{q\times n}$ denote the observations from the first and the second view of n training pairs, with each column of the $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ representing a p-dimensional and q-dimensional feature, $\mathbf{X}^{(1)} = [\mathbf{x}_1^{(1)}, \cdots, \mathbf{x}_n^{(1)}]$ and $\mathbf{X}^{(2)} = [\mathbf{x}_1^{(2)}, \cdots, \mathbf{x}_n^{(2)}]$. The training label vector is $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$, $y_i \in \{1, \cdots, c\}$, with c denoting the number of classes. In summary, the multi-view training features and labels are denoted as triplets[2] $\{\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, y_i\}_{i=1}^n$. Unlike classical multi-view problems with testing data in pairs like $\{\tilde{\mathbf{x}}_i^{(1)}, \tilde{\mathbf{x}}_i^{(2)}\}_{i=1}^{n_t}$, three special cases of recognition tasks with incomplete or misaligned n_t testing samples are considered: (1) testing data lacks the 2nd view, $\{\tilde{\mathbf{x}}_i^{(1)}\}_{i=1}^{n_t}$; (2) testing data is a single view observation taken randomly from the 1st or 2nd view, $\{\tilde{\mathbf{x}}_i^{(v_i)}\}_{i=1}^{n_t}$, where random variable v_i satisfies $\Pr(v_i = 1) = \Pr(v_i = 2) = 0.5$; (3) testing data has both views, but misaligned, $\{\tilde{\mathbf{x}}_i^{(1)}\}_{i=1}^{n_t}$ and $\{\tilde{\mathbf{x}}_j^{(2)}\}_{j=1}^{n_t}$, but $\tilde{\mathbf{x}}_k^{(1)}$ and $\tilde{\mathbf{x}}_k^{(2)}$ are not guaranteed to come from the same source for any specific k.

3.2 Latent Space

To address the aforementioned challenges, we propose to construct a semantic latent space where the imperfect testing data can be projected onto. Ideally, the projections retain the discriminative information and assimilate the

[2] Both measurements $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$come from the same source i, but in distinctive view (1) and view (2).

view differences, converting the aforementioned problems into conventional classification/verification tasks as outlined in Fig. 1.

Let $\mathbf{A}^{(1)} \in \mathcal{R}^{n\times d}$ and $\mathbf{A}^{(2)} \in \mathcal{R}^{n\times d}$ denote projection matrices for the 1st and 2nd view, respectively. Kernelization and centralization are applied to all available training data, which leads to the Gram matrices $\mathbf{K}^{(1)}$ and $\mathbf{K}^{(2)}$,

$$\mathbf{K}^{(v)} = [\boldsymbol{\kappa}_1^{(v)}, \cdots, \boldsymbol{\kappa}_n^{(v)}],\ v = 1, 2. \tag{1}$$

with the element at the ith row and jth column being,

$$\mathbf{K}^{(v)}(i,j) = \langle \mathbf{x}_i^{(v)}, \mathbf{x}_j^{(v)} \rangle_{\mathcal{H}},\ i, j = 1, \cdots, n$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product defined in a Reproducing Kernel Hilbert space.

3.3 Similarity Measure

After kernelization, we continue to project these observations onto a latent space, which captures the common semantics with relatively low dimension. This is realized by multiplying the kernel matrices with the projection matrices $\mathbf{A}^{(v)} \in \mathcal{R}^{n\times d}$, $v = 1, 2$. Consider the matrix $\mathbf{R}^{vv'}$ with its (i, j)th element being

$$\begin{aligned}\mathbf{R}^{vv'}(i,j) &= \langle \mathbf{A}^{(v)T}\boldsymbol{\kappa}_i^{(v)}, \mathbf{A}^{(v')T}\boldsymbol{\kappa}_j^{(v')} \rangle \\ &= \boldsymbol{\kappa}_i^{(v)T}\mathbf{A}^{(v)}\mathbf{A}^{(v')T}\boldsymbol{\kappa}_j^{(v')}. \end{aligned} \tag{2}$$

In the normalized latent subspace, the inner product serves as a natural criterion for measuring distances, i.e., larger inner products indicate smaller angles[3] between projected vectors $\mathbf{A}^{(v)T}\boldsymbol{\kappa}_i^{(v)}$ and $\mathbf{A}^{(v')T}\boldsymbol{\kappa}_j^{(v')}$, and vice versa. Inspired by [34, 35], the angles between the projected vectors are utilized to estimate the similarity between the vectors in this paper. Therefore, if the following constraints are imposed,

$$\mathbf{R}^{vv'}(i,j) \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c'_{ij} & \text{if } y_i \neq y_j \end{cases}, \tag{3}$$

where the thresholds c_{ij}, c'_{ij} are properly selected to match the dataset[4], the similarity between the projected vectors $\mathbf{A}^{(v)T}\boldsymbol{\kappa}_i^{(v)}$ and $\mathbf{A}^{(v')T}\boldsymbol{\kappa}_j^{(v')}$ can be regulated with respect to their labels, i.e., keeping pairs close if they share the same class label and simultaneously separating pairs apart if they differ in class labels.

3.4 Similarity Learning CCA

For notational simplicity, we define the aggregated projection matrices $\mathbf{G} = \left[\mathbf{A}^{(1)T}, \mathbf{A}^{(2)T}\right]^T$. The optimization target of the SLCCA algorithm is,

$$\max_{\mathbf{G}}\ \mathrm{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{G}\mathbf{G}^T\mathbf{L}_1^T\right) \tag{4}$$

$$\text{s.t. } \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{G}\mathbf{G}^T\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} \text{ if } y_i = y_j \\ \leq c_{ij} \text{ otherwise} \end{cases} \tag{5}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{G}\mathbf{G}^T\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} \text{ if } y_i = y_j \\ \leq c_{ij} \text{ otherwise} \end{cases} \tag{6}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{G}\mathbf{G}^T\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} \text{ if } y_i = y_j \\ \leq c_{ij} \text{ otherwise} \end{cases} \tag{7}$$

$$\mathbf{G}^T\boldsymbol{\Gamma}(\mathbf{K})\mathbf{G} = \mathbf{I} \tag{8}$$

where $\mathbf{L}_1 \overset{\text{def}}{=} [\mathbf{I}_n, 0_n]$, $\mathbf{L}_2 \overset{\text{def}}{=} [0_n, \mathbf{I}_n]$, and

$$\boldsymbol{\Gamma}(\mathbf{K}) = \begin{bmatrix} \mathbf{K}^{(1)}\mathbf{K}^{(1)T} + \lambda\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}^{(2)}\mathbf{K}^{(2)T} + \lambda\mathbf{I} \end{bmatrix}, \tag{9}$$

and λ is a small constant number added to eliminate numerical issues.

Note that Eq. (4) and Eq. (8) are equivalent to the standard CCA [13] target, i.e., maximizing the correlations, because in Eq. (4),

$$\begin{aligned} &\mathrm{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{G}\mathbf{G}^T\mathbf{L}_1^T\right) && (10) \\ =&\mathrm{tr}\left(\mathbf{G}^T\mathbf{L}_1^T\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{G}\right) && (11) \\ =&\mathrm{tr}\left(\mathbf{A}^{(1)T}\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{A}^{(2)}\right) && (12) \\ =&\sum_{i=1}^{d}\boldsymbol{\alpha}_i^{(1)T}\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\boldsymbol{\alpha}_i^{(2)} && (13) \\ =&\sum_{i=1}^{d}\boldsymbol{\alpha}_i^{(1)T}\boldsymbol{\Sigma}^{(1,2)}\boldsymbol{\alpha}_i^{(2)}, && (14) \end{aligned}$$

where $\boldsymbol{\alpha}_i^{(1)}$ and $\boldsymbol{\alpha}_i^{(2)}$ in Eq. (13)–(14) denote the ith column of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, respectively. Therefore, $\boldsymbol{\alpha}_i^{(1)T}\mathbf{K}^{(1)}$ and $\boldsymbol{\alpha}_i^{(2)T}\mathbf{K}^{(2)}$ in Eq. (13) denote the ith pair of canonical variables. $\boldsymbol{\Sigma}^{(1,2)} = \mathbf{K}^{(1)}\mathbf{K}^{(2)T}$ denotes the cross-covariance matrix between the two views. Note similar to the standard CCA, normalization of the variables is also required. Plugging Eq. (9) into Eq. (8) leads to

$$\begin{aligned} &\mathbf{G}^T\boldsymbol{\Gamma}(\mathbf{K})\mathbf{G} \\ &= \begin{bmatrix}\mathbf{A}^{(1)T} & \mathbf{A}^{(2)T}\end{bmatrix}\boldsymbol{\Gamma}(\mathbf{K})\begin{bmatrix}\mathbf{A}^{(1)} \\ \mathbf{A}^{(2)}\end{bmatrix} \\ &= \mathbf{A}^{(1)T}\left(\mathbf{K}^{(1)}\mathbf{K}^{(1)T} + \lambda\mathbf{I}\right)\mathbf{A}^{(1)} \\ &+ \mathbf{A}^{(2)T}\left(\mathbf{K}^{(2)}\mathbf{K}^{(2)T} + \lambda\mathbf{I}\right)\mathbf{A}^{(2)} = \mathbf{I}, \end{aligned} \tag{15}$$

which is a relaxed version of the following two constraints Eq. (16)–(17) (ignoring the scaling constant),

$$\mathbf{A}^{(1)T}\left(\mathbf{K}^{(1)}\mathbf{K}^{(1)T} + \lambda\mathbf{I}\right)\mathbf{A}^{(1)} = \mathbf{I}, \tag{16}$$

$$\mathbf{A}^{(2)T}\left(\mathbf{K}^{(2)}\mathbf{K}^{(2)T} + \lambda\mathbf{I}\right)\mathbf{A}^{(2)} = \mathbf{I}. \tag{17}$$

Eq. (16)–(17) are the standard CCA orthogonal constraints in matrix equations form. For example, Eq. (16) is equiva-

[3] Assuming unit vector norms.

[4] The choices of c_{ij}, c'_{ij} are subtle, which need to strike a balance between the feasibility of the optimization and the effects of similarity learning. Details are provided in the following Section 5.

lent to

$$\boldsymbol{\alpha}_i^{(1)T}\hat{\boldsymbol{\Sigma}}^{(1,1)}\boldsymbol{\alpha}_i^{(1)} = 1,\ i = 1, \cdots, d \tag{18}$$

$$\boldsymbol{\alpha}_i^{(1)T}\hat{\boldsymbol{\Sigma}}^{(1,1)}\boldsymbol{\alpha}_j^{(1)} = 0,\ i, j = 1, \cdots, d, i \neq j, \tag{19}$$

where $\boldsymbol{\alpha}_i^{(1)}$ and $\boldsymbol{\alpha}_j^{(1)}$ denote the ith and jth column of $\mathbf{A}^{(1)}$, respectively, and $\hat{\boldsymbol{\Sigma}}^{(1,1)} = \mathbf{K}^{(1)}\mathbf{K}^{(1)T} + \lambda\mathbf{I}$ is the regularized covariance matrices in the first view.

The compact format of Eq. (15) is preferable to Eq. (16)–(17) due to it is easier to optimize with respect to $\mathbf{G}$ using a LP solver. Empirically, Eq. (15) gives highly similar results to that produced by Eq. (16)–(17).

In addition to the standard CCA target Eq. (4) and constraint Eq. (8), additional constraints of Eq. (5)–(7) are included to impose the cross-view similarity and intra-view similarity conditions. Plugging the equations $\mathbf{A}^{(1)} = \mathbf{L}_1\mathbf{G}$ and $\mathbf{A}^{(2)} = \mathbf{L}_2\mathbf{G}$ into Eq. (5)–(7) leads to the aforementioned similarity preserving constraints detailed in Eq. (3).

Unfortunately, the optimization problem in Eq. (4)–(8) is a generic quadratically constrained quadratic program (QCQP) problem and it is NP-hard [23, 26]. In order to approximate the solution with reasonable computational complexity, in the following Section 4 we first adopt relaxations to the problem, and subsequently develop an efficient iterative solution to it.

4. RELAXATION AND SOLUTION

To alleviate the computational difficulty of optimizing the objective function in Eq. (4)–(8), we apply the SDP relaxation techniques presented in [23] and propose the approximate solutions, accordingly.

As a standard step of SDP relaxation, an extra positive semidefinite matrix $\mathbf{M}_G$ relevant to the original optimization variable $\mathbf{G}$ need to be defined. The definition of $\mathbf{M}_G$ determines how the original constraint in Eq. (8) is relaxed. As later shown in Eq. (26), we choose a conservative definition of $\mathbf{M}_G$ with dimension[5] of $2n \times 2n$. Ideally,

$$\mathbf{M}_G = \mathbf{G}\mathbf{G}^T = \begin{bmatrix} \mathbf{A}^{(1)}\mathbf{A}^{(1)T} & \mathbf{A}^{(1)}\mathbf{A}^{(2)T} \\ \mathbf{A}^{(2)}\mathbf{A}^{(1)T} & \mathbf{A}^{(2)}\mathbf{A}^{(2)T} \end{bmatrix}. \tag{20}$$

However, if Eq. (20) holds exactly, the optimization is identical to (i.e., as difficult as) the original QCQP problem. Hence in stead of requiring $\mathbf{M}_G = GG^T$, we relax it to $\mathbf{M}_G - GG^T \succeq 0$ (following [23]), and using the Schur's complement, this constraint is equivalent to

$$\begin{bmatrix} \mathbf{M}_G & \mathbf{G} \\ \mathbf{G}^T & \mathbf{I}_d \end{bmatrix} \succeq 0. \tag{21}$$

With Eq. (21) defined, the original constraint in Eq. (8) can be readily relaxed to Eq. (26), where only the trace[6] of $\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_G$ is fixed instead of a series of constraints on each element of the $\mathbf{G}^T\boldsymbol{\Gamma}(\mathbf{K})\mathbf{G}$. Empirically, a large number of constraints drastically increase the computational burden and even make the optimization infeasible, and this relaxation reduces the number of constraints[7], which makes the practical solution of Eq. (22)–(28) possible.

$$\max_{\mathbf{G},\mathbf{M}_G}\ \mathrm{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_1^T\right) \tag{22}$$

$$\text{s.t. } \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{23}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{24}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{25}$$

$$\mathrm{tr}\left(\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_G\right) = d \tag{26}$$

$$\mathbf{M}_G = \mathbf{M}_G^T \tag{27}$$

$$\begin{bmatrix} \mathbf{M}_G & \mathbf{G} \\ \mathbf{G}^T & \mathbf{I}_d \end{bmatrix} \succeq 0, \tag{28}$$

Note that the problem in Eq. (22)–(28) takes the standard form of an SDP, hence it is convex and theoretically solvable with the standard interior point methods. However, in common visual recognition scenarios, the total dimension of variables in $\mathbf{M}_G$ and $\mathbf{G}$ are too large for the off-the-shelf SDP solvers, and it is known that a large scale SDP problem is computational prohibitive [23]. Therefore we further relax the problem in Eq. (22)–(28) to the following alternating optimization where a computational efficient linear program is adopted instead.

$$\max_{\mathbf{M}_G}\ \mathrm{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_1^T\right) + \mu\mathrm{tr}\left(\mathbf{G}^T\mathbf{M}_G\mathbf{G}\right) \tag{29}$$

$$\text{s.t. } \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{30}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{31}$$

$$\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \tag{32}$$

$$\mathrm{tr}\left(\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_G\right) = d \tag{33}$$

$$\mathbf{M}_G = \mathbf{M}_G^T,\ \mathbf{G}^T\mathbf{M}_G\mathbf{G} = \mathbf{I} \tag{34}$$

and

$$\max_{\mathbf{G}}\ \mathrm{tr}\left(\mathbf{G}^T\mathbf{M}_G\mathbf{G}\right) \tag{35}$$

$$\text{s.t. } \mathbf{G}^T\mathbf{G} = \mathbf{I} \tag{36}$$

By alternating optimizing with respect to $\mathbf{M}_G$ and $\mathbf{G}$ using Eq. (29)–(34) and Eq. (35)–(36), respectively, the aggregated projection matrix $\mathbf{G}$ can be obtained. As for the selection of parameters and initialization, these details are given in the following Section 5. Whilst this alternating process could not necessarily find the global optimum $\mathbf{G}$ specified in Eq. (4)–(8), it is guaranteed to find at least a strong local minimum such that no further perturbations increase Eq. (29) and Eq. (35).

5. DATASETS AND SETTINGS

The experiments are conducted on four different multi-view datasets, i.e., the RGBD Object dataset [16], the NYU Depth V1 Indoor Scenes dataset [27], the multi-spectral

[5] Alternative relaxations are presented in the appendix.

[6] It is also the trace of $\mathbf{G}^T\boldsymbol{\Gamma}(\mathbf{K})\mathbf{G}$.

[7] From d^2, the total number of element in matrix $\mathbf{G}^T\boldsymbol{\Gamma}(\mathbf{K})\mathbf{G}$ down to 1.

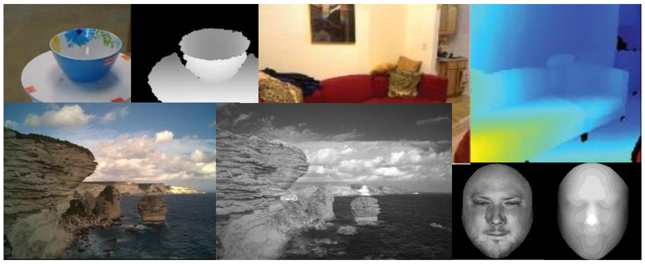


Figure 2: Examples of multi-view image pairs. Left to right, top to bottom, [16] RGB/depth, [27] RGB/depth; [4] RGB/infrared, [36] grayscale/depth.

scene dataset [4] and Binghamton University 3D Facial Expression dataset [36]. Examples are shown in Fig. 2.

For the missing data recognition task, two types of missing schemes are considered. First the systematic missing scheme is included, where both views are present in the training phase, but the second view is nonexistent in the testing phase. The alternative type is the random missing scenario, where both views are present in the training phase, but the testing data contains only a single view (either the first view or the second view), chosen randomly according to a Bernoulli distribution with probability $p = 0.5$.

With all four datasets, we test our proposed method against competing ones in both missing data scenarios, i.e., the complete missing of second view case and the random missing case. In addition, with the Binghamton University 3D Facial Expression dataset, we also conducted the cross-view identity verification task, whose target is to determine whether the "depth-face" and "image-face" pairs come from the same person. Regarding the off-the-shelf recognition algorithms, the LIBSVM [5] is adopted for classification while the joint Bayesian verification algorithm [6, 18, 19] is adopted for verification.

While evaluating the performance, we adopted five competing algorithms, namely, SVM/Raw, SVM-2K*, KCCA, RGCCA, and DCCA. With both missing data recognition tasks, the baseline is the direct application of the SVM. For the complete missing of second view scenario, one SVM is learned only on the first view training data, and tested on the first view testing data. For the random missing scenario, two SVM are separately learned on two views, and the testing data is processed by the corespondent SVM. For the cross-modal verification task with the Binghamton University 3D Facial Expression dataset, the "Raw" approach directly trains a joint Bayesian verification algorithm based on the similar/dis-similar cross-view pairs without any latent space projections.

Termed as the "SVM-2K*", this approach is a naive extension of "SVM-2K" [11] to the missing data scenario of multi-view learning, in which we train with both views, but test with only the first view classifier or corresponding single view classifier, in the "missing 2nd view" and "missing randomly" scenarios, respectively. The "KCCA" approach also adopts the latent space framework in Fig. 1, however, it maximizes the nonlinear correlation between the first view and the second view so as to construct a pair of projection vectors iteratively using the standard kernelized CCA [13]. After obtaining the latent space, subsequent SVM or joint Bayesian verification algorithm is applied in this latent space. "RGCCA" and "DCCA" are two extended versions of the "KCCA", they are different from "KCCA" only in replacing the standard kernelized CCA [13] with the algorithms proposed in [30] and [38], respectively.

In the following experiments, the RBF kernel is applied to both the KCCA algorithm and the SVM algorithms. The bandwidth parameter in the RBF kernel, the parameters in the SVM-2K* algorithm, λ and μ in the proposed approach are all determined by a 4-fold cross validation. The data dependent thresholds c_{ij} in Eq. (30)–(32) are determined as follows. First, the proposed algorithm is initialized with the projection matrix $\mathbf{G}_0$ learned by the KCCA algorithm [8], and $\mathbf{M}_{G0} = \mathbf{G}_0\mathbf{G}_0^T$. Then the initial original values c_{0ij} are computed using the Eq. (30)–(32), respectively (e.g., with Eq. (30), $c_{0ij} = \mathrm{tr}(\boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_{G0}\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)})$). After obtaining the c_{0ij} values, a contraction/expansion process is applied as follows. $c_{ij} = (1+\epsilon)c_{0ij}$ if $y_i = y_j$, and $c_{ij} = (1-\epsilon)c_{0ij}$ otherwise, where the scalar $\epsilon \in [0 : 0.05 : 0.5]$ is determined by a cross validation.

6. EXPERIMENTAL RESULTS

In this section, the proposed method is tested against competing algorithms on four real world datasets with two missing data recognition tasks and one cross-modal verification task.

6.1 UW RGBD Dataset

The RGBD object dataset [16, 3] consists of paired RGB images and depth maps of multiple object instances from 51 categories. Without loss of generality, we assign the RGB images as the first view, and the depth maps as the second view. We focus on the instance level object recognition within each of the categories, and demonstrate that the proposed SLCCA paired with the discriminative latent space model can facilitate better recognition based on the single view data, or randomly interlaced single view data. We follow [16] to construct the "leave-sequence-out" training/testing split[9], and [3] to extract the HMP-based features from both views.

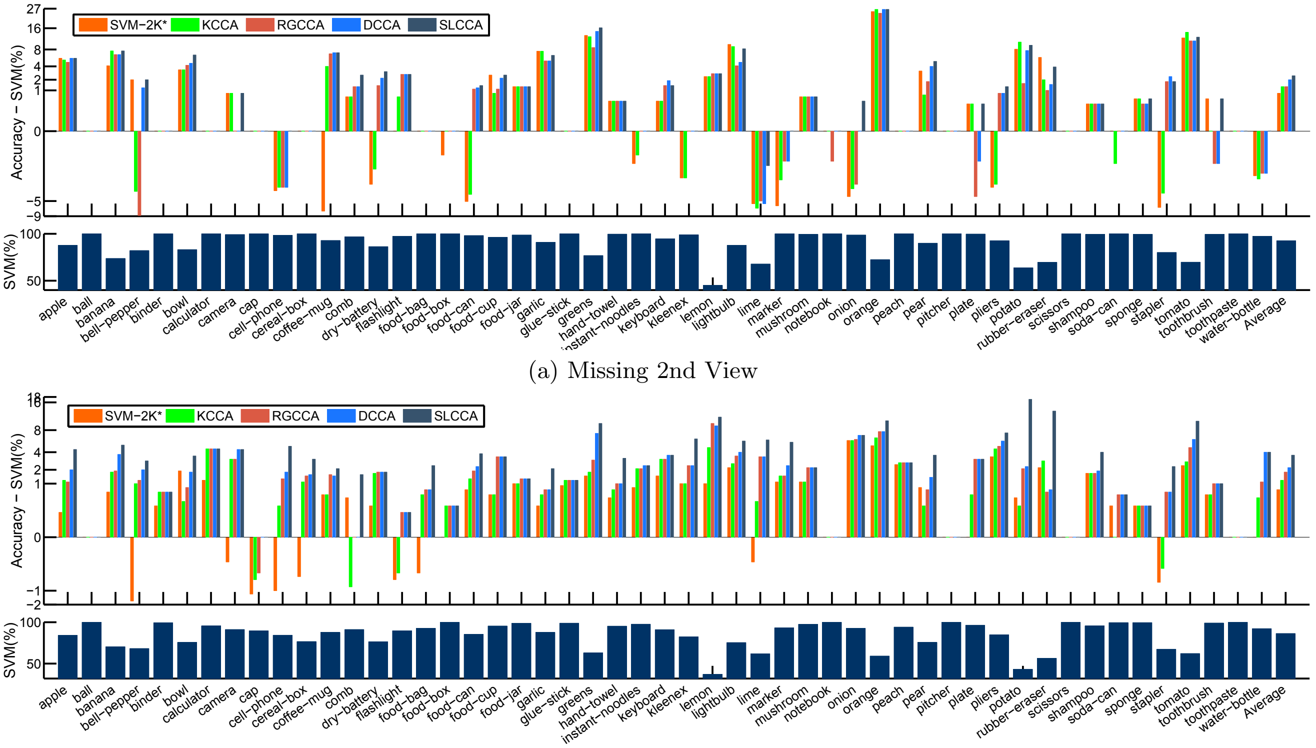
The recognition results under the "missing 2nd view" scenario and "missing randomly" scenario are illustrated in Fig. 3 and also summarized in Table 1. Due to the drastic change of recognition difficulties among the categories[10], the recognition accuracies fluctuate significantly. To alleviate the difficulty of comparison, the SVM is selected as a baseline with its accuracies plotted at the bottom in both the "Missing 2nd View" and "Missing Randomly" cases. The recognition accuracies of the remaining algorithms are first subtracted by the SVM baseline and the differences are situated at the top in both cases.

With the case of missing the second view (Fig. 3(a)), the SVM-2K* variant offers no significant improvements against the baseline SVM approach, while the KCCA and RGCCA both offer marginally better recognition results. The DCCA approach is better with an average recognition accuracy of 94.6%, but not as high as our proposed SLCCA approach, which achieves 95.1% accuracy. Again with the random missing scheme, in Fig. 3(b), although the recognition ac-

[8] The KCCA algorithm is needed both for a reasonable initialization and also for faster practical convergence.

[9] This is a fixed split scheme, hence neither standard deviation nor t-test is applicable.

[10] For example, it is much harder to distinguish three tomatoes than distinctive brands of cellphones.



(a) Missing 2nd View

(b) Missing Randomly

Figure 3: Multi-View RGBD Object Instance Recognition from all 51 categories and average results (in percentage). For both case (a) Missing 2nd View and case (b) Missing 2nd View, the accuracies of SVM baseline are situated at the bottom, while the differences of subtracting SVM accuracies from the corresponding, remaining algorithms are situated at the top. The Average accuracies across all 51 categories for both cases are situated furthest to the right.

curacy varies a lot across categories, the proposed approach offers obvious performance boost against competing algorithms, achieves an average accuracy of 90.0%.

Table 1: Multi-View RGBD Object Instance Recognition, average results across all 51 categories.

Missing	SVM	SVM-2K*	KCCA	RG-CCA	DCCA	SLCCA
2nd View	92.6	93.4	93.9	93.9	94.6	95.1
Randomly	86.4	87.1	87.6	88.2	88.6	90.0

6.2 NYU Indoor Scenes Dataset

Collected by a custom made Kinect sensor, the NYU-Depth-V1-Indoor Scene Dataset [27] consists of RGB-depth image pairs of 7 classes of indoor scenes. This multi-channel capturing scheme enables us to examine a wide range of missing-data recognition scenarios. First, two types of view assignments are considered. As shown in Table 2, *L+D* denotes the case where the luminance channel (i.e., the grayscale image, which is obtained by converting from the RGB channels) feature is assigned as the first view, while the depth channel as the second view, and *RGB+D* denotes the case where the first view consists of features extracted independently from all the R,G and B channels and then concatenated, while the depth channel feature is regarded as the second view.

For the missing data profile, we consider both the systematic missing scheme and the random missing scheme detailed in Section 5, which are denoted as “Missing 2nd View” and “Missing Randomly” in Table 2, respectively. We consider two types of features: both the lower dimensional GIST [21] features and the higher dimensional spatial pyramid bag-of-feature representation [17]. All these features are firstly extracted independently from the respective channel (luminance, R, G, B or depth channel). For the spatial pyramid bag-of-feature representation, we densely extract SIFT descriptors from 40 × 40 patches with a stride of 10 pixels, and use two dictionary of size (200 and 800) for the k-means clustering. We denote these two feature extraction schemes as SP200 and SP800 in Table 2, respectively.

The experimental data are randomly split into 50 folds of training data and testing data with comparable number of samples (following [27]). The proposed and the competing algorithms are tested and the mean and the standard deviation of the recognition accuracies (in percentage) are summarized in Table 2, with three different kinds of feature matching schemes, two types of missing data scenarios and two sets of view assignments.

It can be observed that the higher dimensional and more sophisticated spatial pyramid bag-of-feature representation is capable of achieving higher recognition accuracies than the lower dimensional GIST features. Also, the random missing scheme generally poses a more challenging problem than the systematic missing scheme, due to the phenomenon that the first view (grayscale or RGB images) is intrinsically more informative than the second view (depth maps) in determining

Table 2: NYU Depth V1 Indoor Scenes

Settings	GIST, Missing 2nd View		GIST, Missing Randomly	
Views	L+D	RGB+D	L+D	RGB+D
SVM	59.61±3.42	61.33±3.32	42.17±4.72	43.43±5.87
SVM-2K*	58.26±3.71	60.92±3.80	42.47±4.33	43.51±4.82
KCCA	59.33±6.92	61.58±6.28	43.43±5.44	43.67±4.32
RGCCA	58.97±5.83	62.15±4.85	44.43±4.52	45.22±5.11
DCCA	60.37±4.23	62.98±4.37	44.88±6.10	45.52±4.08
SLCCA	60.45±6.89	62.87±5.92	46.22±5.12	47.08±4.71
Settings	SP200, Missing 2nd View		SP200, Missing Randomly	
Views	L+D	RGB+D	L+D	RGB+D
SVM	63.24±3.61	65.01±4.05	45.59±4.12	46.60±3.88
SVM-2K*	60.37±4.25	61.23±4.93	44.52±5.25	46.76±4.65
KCCA	63.37±5.29	64.22±5.41	46.02±4.30	46.68±5.31
RGCCA	63.04±4.83	63.85±5.37	46.22±5.31	46.70±5.27
DCCA	65.90±4.76	65.22±5.69	47.02±6.20	46.98±4.82
SLCCA	66.02±7.69	66.29±6.49	48.98±5.13	49.92±4.98
Settings	SP800, Missing 2nd View		SP800, Missing Randomly	
Views	L+D	RGB+D	L+D	RGB+D
SVM	64.63±3.13	65.72±3.75	46.80±3.37	48.59±4.62
SVM-2K*	59.73±4.32	61.82±4.49	45.47±4.38	47.65±5.73
KCCA	65.14±5.72	65.56±6.23	48.20±3.31	48.60±3.28
RGCCA	65.09±5.39	66.16±5.43	48.60±4.69	48.88±5.88
DCCA	65.76±5.07	66.44±5.02	48.92±5.34	49.06±6.21
SLCCA	66.79±5.94	67.26±6.54	51.46±4.33	52.02±4.27

the scene types, especially the highly clutter indoor scenes in this dataset.

Comparing the results from these algorithms, we note that the variant of SVM-2K* fails to achieve significant performance advantage over the baseline SVM algorithm. We would like to point out that SVM-2K [11] was formulated as a multi-view learning algorithm without considering the missing data case, and the simple extension to the missing data case may not satisfy the assumption of the "prediction consistency" regularization. The competing KCCA, RGCCA and DCCA algorithms provide various degrees of performance enhancements but not as significant as the proposed SLCCA approach. Conducting the *t*-tests for the 12 settings in Table 2, the proposed SLCCA approach significantly outperforms the baseline SVM algorithm 8 and 2 times with confidence level 0.95 and 0.90, respectively.

6.3 Multi-Spectral Scene Dataset

With these experimental results on the multi-spectral scene dataset [4], we demonstrate the efficacy of our proposed approach in missing-data resilient scene recognition, with both the systematic missing scheme ("Missing 2nd View") and the random missing scheme ("Missing Randomly"). The multi-spectral scene dataset [4] consists of 477 paired RGB and near-infrared (IR) images from 9 different scene categories.

In line with [4], the RGB images are first converted to the **LAB** color space, subsequently the GIST [21] features are extracted independently on each of the L, A, B and IR channels. Two view assignments are considered, i.e., *L+I* denotes GIST features extracted from the luminance channel and the IR channel are assigned as the first view and the second view, respectively; and *LAB+I* denotes GIST features extracted independently from the L, A and B channels are concatenated to form the first view, while those extracted from the IR channel are considered as the second view. Following [4], we construct 50 random training/testing splits, and the mean and the standard deviation of the recognition accuracies (in percentage) are reported in Table 3.

From Table 3, despite the large standard deviations (due to the large intra-class variations [4]), we note that the *LAB+I* assignments are generally better than the *L+I*, which is consistent with the *LAB+I* assignment being more informative. Again, the random missing scheme is generally more challenging than the systematic missing scheme, also due to the limited discriminative power in the second view. With the competing algorithms, neither the KCCA nor the SVM-2K* algorithm achieves a significant advantage over the SVM baseline. The more flexible RGCCA and DCCA algorithms marginally outperform the baseline SVM algorithm. As anticipated, the proposed SLCCA achieves the highest accuracies and it significantly outpeforms the SVM baseline in all four cases of Table 3 in *t*-tests at confidence level 0.99.

Table 3: Multi-Spectral Scenes

Settings	Missing 2nd View		Missing Randomly	
Views	LAB+I	L+I	LAB+I	L+I
SVM	67.78±5.25	61.82±3.77	42.11±5.56	38.02±4.69
SVM-2K*	67.17±5.58	61.82±4.59	42.80±5.42	38.22±4.30
KCCA	67.87±4.76	62.32±4.81	44.21±4.89	39.08±5.61
RGCCA	68.59±3.94	62.22±3.96	46.02±3.36	40.73±5.28
DCCA	70.51±2.37	65.66±4.57	46.60±5.40	41.22±4.61
SLCCA	71.22±3.26	66.12±4.68	49.18±4.92	43.64±5.21

6.4 Binghamton 3D Facial Expression dataset

For the experiment on this Binghamton 3D Facial Expression dataset [36], we demonstrate the efficacy of our proposed approach in both the missing data recognition problems and the cross-modal verification problem. We show that by collegially modeling the discriminative model based on two views, various visual recognition tasks (expression, gender, and race recognition) with missing data could be improved. In addition, we demonstrate the effectiveness of the proposed approach in recover the semantic correspondences with cross-modal identity verification (determining whether a pair of image-face and depth-face come from the same person).

The Binghamton dataset [36] consists of 3D face models as well as face images of 100 subjects (with both genders and a variety of ethnic/racial ancestries, i.e., White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino) of 7 facial expressions (happiness, disgust, fear, angry, surprise, sadness and neutral). We extract all the frontal face images and the frontal face depth maps from the 3D point clouds, and feed them to the Eigen-PEP model [6, 18, 19] to extract features independently on the grayscale images and on the depth maps, respectively. Unlike 3D face features such as [33, 20] that require manual labelling of facial landmarks, our features are automatically generated without the need of these facial landmark labels.

Following conventions, we assign the image features as the first view and the depth features as the second view. For statistical stability, 50 random training/testing splits are constructed, with equal number of subjects as training and testing samples. While constructing these splits, we adopt an exclusive-identity protocol, i.e., we make sure that any individual appearing in the training set never simultaneously appears in the testing set, and vice versa. The classification algorithm is SVM using the LIBSVM [5], and the target is to recognize the expression, gender, ethnic/racial ancestries of the testing samples. The target of the verification is to verify whether a facial image and a facial depth map come from the same individual (regardless of expression variations), and the algorithm is based on the joint Bayesian verification algorithm in [6, 18, 19], where 10000 pairs of similar and dissimilar pairs are sampled and fed to the training process.

Table 4: Binghamton 3D Facial Expression

Settings	Recognition:Missing 2nd View			
Algorithms	SVM	KCCA	DCCA	SLCCA
expression	72.2±4.1	73.1±3.7	74.5±4.1	75.8±3.2
gender	92.1±3.2	89.4±4.1	92.6±5.3	92.8±3.8
race	72.0±3.9	74.1±4.2	75.2±5.2	78.1±4.3
Settings	Recognition:Missing Randomly			
expression	69.1±3.9	69.5±3.3	71.9±5.9	73.8±4.2
gender	87.4±4.2	88.6±3.5	89.5±4.4	89.6±3.5
race	64.0±4.2	66.2±3.7	66.5±5.3	68.4±4.3
Settings	Cross-modal Verification			
Algorithms	Raw	KCCA	DCCA	SLCCA
Accuracy	81.1±4.2	82.3±3.9	82.6±4.4	86.4±5.4

The recognition/verification accuracies (in percentage) are summarized in Table 4. The missing data recognition results are shown in the upper part of Table 4, with the systematic missing scheme followed by the random missing scheme. In both scenarios, we carried out recognition based on the expression, gender and race labels. In both the expression and race recognition tasks, the proposed algorithm outperforms the baseline SVM algorithm significantly in *t*-tests at confidence level 0.99. The gender recognition task is different in that only two classes (male and female) are present with potentially larger intra-class variations, which could limit the effectiveness of discriminative similarity learning algorithms.

In summary, the proposed algorithm achieves the highest verification accuracy in the cross-modal identity verification task, and it significantly outperforms the Raw method (detailed in Section 5) in the *t*-test at confidence level 0.99.

6.5 Comparison of Algorithms

KCCA finds the latent space without explicitly resorting to the labels; DCCA first encodes an extra view based on training labels, and subsequently incorporates the label information by maximizing correlations between existing views and the encoded view; unlike the previous two-step way, the proposed SLCCA explicitly incorporates similarity constraints inside the optimization target, which leads to the best performance benchmarks in the aforementioned experiments.

7. CONCLUSION

We have proposed a recognition framework casting the multi-view information into a single semantic latent space and developed the SLCCA algorithm to construct such a latent space for the missing data recognition / verification problem. The proposed algorithm explicitly preserves the intra-class and inter-class relationships, and it is implemented with relaxation and alternating optimization. Experiments with four different datasets demonstrate that the proposed method outperforms the competing ones.

8. ACKNOWLEDGMENTS

Research reported in this publication was partly supported by the National Institute Of Nursing Research of the National Institutes of Health under Award Number R01NR015371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work is also partly supported by US National Science Foundation Grant IIS 1350763, Gang Hua's start-up funds and Qilin Zhang's Innovation & Entrepreneurship Doctoral Fellowship from Stevens Institute of Technology.

9. REFERENCES

[1] A. Argyriou, T. Evgeniou, M. Pontil, A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. In *Machine Learning*. press, 2007.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[3] L. Bo, X. Ren, and D. Fox. Unsupervised feature learning for rgb-d based object recognition. In *Experimental Robotics*, pages 387–402. Springer, 2013.

[4] M. Brown and S. Susstrunk. Multi-spectral sift for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 177–184. IEEE, 2011.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.

[7] J. Chen, X. Liu, and S. Lyu. Boosting with side information. In *Computer Vision–ACCV 2012*, pages 563–577. Springer, 2013.

[8] L. Chen, W. Li, and D. Xu. Recognizing rgb images by learning from rgb-d data. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1418–1425. IEEE, 2014.

[9] J. Currie and D. I. Wilson. Opti: lowering the barrier between open source optimizers and the industrial matlab user. *Foundations of computer-aided process operations, Savannah, Georgia, USA*, pages 8–11, 2012.

[10] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[11] J. Farquhar, D. Hardoon, H. Meng, J. S. Shawe-taylor, and S. Szedmak. Two view learning: Svm-2k, theory and practice. In *Advances in neural information processing systems*, pages 355–362, 2005.

[12] A. Globerson and S. Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360. ACM, 2006.

[13] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[14] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine*

Intelligence, IEEE Transactions on, 29(6):1005–1018, 2007.

[15] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.

[16] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[18] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE, 2013.

[19] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 793–800. IEEE, 2013.

[20] A. Maalej, B. B. Amor, M. Daoudi, A. Srivastava, and S. Berretti. Shape analysis of local facial patches for 3d facial expression recognition. *Pattern Recognition*, 44(8):1581–1589, 2011.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[22] Z. Qi, M. Yang, Z. M. Zhang, and Z. Zhang. Mining noisy tagging from multi-label space. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1925–1929. ACM, 2012.

[23] A. Qualizza, P. Belotti, and F. Margot. Linear programming relaxations of quadratically constrained quadratic programs. In *Mixed Integer Nonlinear Programming*, pages 407–426. Springer, 2012.

[24] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pages 213–226. Springer, 2010.

[25] L. Shams, D. R. Wozny, R. Kim, and A. Seitz. Influences of multisensory experience on subsequent unisensory processing. *Frontiers in psychology*, 2, 2011.

[26] H. D. Sherali and B. M. Fraticelli. Enhancing rlt relaxations via a new class of semidefinite cuts. *Journal of Global Optimization*, 22(1-4):233–261, 2002.

[27] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 601–608. IEEE, 2011.

[28] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[29] L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):194–200, 2011.

[30] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

[31] T. Tommasi, N. Quadrianto, B. Caputo, and C. H. Lampert. Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In *Computer Vision–ACCV 2012*, pages 1–15. Springer, 2013.

[32] V. Vapnik, A. Vashist, and N. Pavlovitch. Learning using hidden information (learning with teacher). In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 3188–3195. IEEE, 2009.

[33] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1399–1406. IEEE, 2006.

[34] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[35] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512, 2002.

[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[37] D. Zhang, J. He, Y. Liu, L. Si, and R. D. Lawrence. Multi-view transfer learning with a large margin approach. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1208–1216, 2011.

[38] Q. Zhang, G. Hua, W. Liu, Z. Liu, and Z. Zhang. Can visual recognition benefit from auxiliary information in training? In *Computer Vision – ACCV 2014*, volume 9003 of *Lecture Notes in Computer Science*, pages 65–80. Springer International Publishing, 2015.

APPENDIX

A. ALTERNATIVE RELAXATIONS

In this appendix, two alternative relaxations to the one presented in Section 4 are first presented, and followed by brief comparisons.

Indeed, the relaxations proposed in Section 4 are not tight; abundant works have been devoted to the QCQP relaxations [23, 26] and many tighter relaxations exist, but they are computational prohibitive while solving the practical scale optimization problem in multimedia recognition. These alternative relaxations are tighter approximations[11], but they

[11] In the sense of enforcing Eq. (8).

are not chosen for computational reasons and numerical issues.

A.1 Alternative Relaxation 1

In stead of Eq. (26) in Section 4, this alternative relaxation defines d distinctive $\mathbf{M}_{Gj}$, where $j = 1, 2, \cdots, d$ to make the constraints in Eq. (39) possible. Ideally,

$$\mathbf{M}_{Gj} = \mathbf{G}(:, j)\mathbf{G}(:, j)^T, \; j = 1, \cdots, d, \tag{37}$$

where $\mathbf{G}(:, j), j = 1, \cdots, d$ denotes the jth column of matrix $\mathbf{G}$. However, the SDP relaxation only enforces Eq. (40) instead of Eq. (37). Therefore, the optimization target is,

$$\max_{\mathbf{G}, \mathbf{M}_{G1}, \cdots, \mathbf{M}_{Gd}} \sum_{j=1}^{d} \operatorname{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{M}_{Gj}\mathbf{L}_1^T\right) \tag{38}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_{Gj}\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_{Gj}\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{M}_{Gj}\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \operatorname{tr}(\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_{Gj}) = 1, \qquad j = 1, \cdots, d \qquad (39) \\
& \mathbf{M}_{Gj} = \mathbf{M}_{Gj}^T, \qquad j = 1, \cdots, d \\
& \begin{bmatrix} \mathbf{M}_{Gj} & \mathbf{G}(:, j) \\ \mathbf{G}(:, j)^T & 1 \end{bmatrix} \succeq 0, \qquad j = 1, \cdots, d \qquad (40)
\end{aligned}$$

The specific definition in Eq. (37) leads to Eq. (39), where the d^2 scalar equations in Eq. (8) are reduced to d scalar equations in Eq. (39). Compared with Eq. (26), the number of constraints d are still to large to practically enforce. Worse still, there are d times more elements in $\mathbf{M}_{Gj}$, $j = 1, \cdots, d$ of Eq. (37) than the $\mathbf{M}_G$ in Eq. (20).

A.2 Alternative Relaxation 2

Similar to Eq. (37), the ideal value of $\mathbf{M}_G$ is $\operatorname{vec}(\mathbf{G})\operatorname{vec}(\mathbf{G})^T$, and it is similarly relaxed and enforced in Eq.(43). Therefore, the optimization target is:

$$\max_{\operatorname{vec}(\mathbf{G}), \mathbf{M}_G} \sum_{j=1}^{d} \operatorname{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{M}_G(jj)\mathbf{L}_1^T\right) \tag{41}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G(jj)\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G(jj)\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \sum_{j=1}^{d} \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{M}_G(jj)\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \\
& \operatorname{tr}(\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_G(ji)) = \delta_{ji}, i, j = 1, \cdots, d \qquad (42) \\
& \mathbf{M}_G = \mathbf{M}_G^T, \\
& \begin{bmatrix} \mathbf{M}_G & \operatorname{vec}(\mathbf{G}) \\ \operatorname{vec}(\mathbf{G})^T & 1 \end{bmatrix} \succeq 0, j = 1, \cdots, d \qquad (43)
\end{aligned}$$

where δ_{ji} denotes the Kronecker delta and $\operatorname{vec}(\mathbf{G})$ denotes the vectorization (column by column) of matrix $\mathbf{G}$.

With this relaxation, the d^2 scalar equations in Eq. (8) are all kept in Eq. (42). Compared with Eq. (26), the number of constraints d^2 are so large that it is practically intractable. Worse still, as defined in this relaxation, the $\mathbf{M}_G$ is a huge rank-one matrix of size $2nd \times 2nd$, which is well beyond the capacity of current solvers.

B. ALTERNATING OPTIMIZATION

In this section, the derivation details of the proposed alternating optimization method in Section 4 are presented.

B.1 Updating $\mathbf{M}_G$

After obtaining the large scale SDP problem, the following modified **EXT** relaxation [23] is obtained by dropping the last positive semidefinite constraint. We modify the **EXT** relaxation to include an extra $\operatorname{tr}\left(\mathbf{G}^T\mathbf{M}_G\mathbf{G}\right)$ term, which links the constraints between the augmented variable $\mathbf{M}_G$ and the original variable $\mathbf{G}$. Even with the extra term, the modified **EXT** retains the property of a linear program and can be efficiently solved by an off-the-shelf solvers such as [9]. $\mathbf{M}_G$ is obtained by solving the following linear program,

$$\begin{aligned}
\max_{\mathbf{M}_G} \; & \operatorname{tr}\left(\mathbf{K}^{(1)}\mathbf{K}^{(2)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_1^T\right) \\
& + \mu\operatorname{tr}\left(\mathbf{G}^T\mathbf{M}_G\mathbf{G}\right) \qquad (44) \\
\text{s.t.} \; & \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \qquad (45) \\
& \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_1\mathbf{M}_G\mathbf{L}_1^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \qquad (46) \\
& \boldsymbol{\kappa}_i^{(1)T}\mathbf{L}_2\mathbf{M}_G\mathbf{L}_2^T\boldsymbol{\kappa}_j^{(2)} \begin{cases} \geq c_{ij} & \text{if } y_i = y_j \\ \leq c_{ij} & \text{otherwise} \end{cases} \qquad (47) \\
& \operatorname{tr}(\boldsymbol{\Gamma}(\mathbf{K})\mathbf{M}_G) = d \qquad (48) \\
& \mathbf{M}_G = \mathbf{M}_G^T \qquad (49) \\
& \mathbf{G}^T\mathbf{M}_G\mathbf{G} = \mathbf{I}. \qquad (50)
\end{aligned}$$

The optimization problem presented in Eq. (44)–(50) is a linear program [23], hence a global optimal solution is guaranteed. In our experiments, we use the OPTI toolbox [9] to solve this LP problem.

B.2 Updating $\mathbf{G}$

Discarding irrelevant terms, the $\mathbf{G}$ is updated by the following optimization,

$$\max_{\mathbf{G}} \operatorname{tr}\left(\mathbf{G}^T\mathbf{M}_G\mathbf{G}\right) \;\; \text{s.t. } \mathbf{G}^T\mathbf{G} = \mathbf{I}. \tag{51}$$

Note that this is exactly the canonical correlation analysis problem [13]. Note that $\mathbf{G} = \left[\mathbf{A}^{(1)T}, \mathbf{A}^{(2)T}\right]^T$ and $n > d$, $\mathbf{G}$ can be obtained by Incomplete Cholesky factorization of $\mathbf{M}_G$. Denote the SVD of $\mathbf{M}_G$ be

$$\mathbf{M}_G = \mathbf{U}\Sigma\mathbf{U}^T = \mathbf{U}\Sigma_1\mathbf{U}^T + \mathbf{U}\Sigma_2\mathbf{U}^T, \tag{52}$$

where $\Sigma, \Sigma_1, \Sigma_2$ are diagonal matrices with corresponding non-negative singular values on their diagonals, and Σ_1 contains the largest d singular values while Σ_2 containing the remaining ones. $\mathbf{G} = \mathbf{U}\Sigma_1^{1/2}$. In practise, an incomplete Cholesky factorization is favored due to its computational efficiency.