IPSJ Transactions on Computer Vision and Applications Vol.7 138–150 (Nov. 2015)

[DOI: 10.2197/ipsjtcva.7.138]

Regular Paper

Auxiliary Training Information Assisted Visual Recognition

Qilin Zhang[1,a)] Gang Hua[1,4,b)] Wei Liu[2] Zicheng Liu[3] Zhengyou Zhang[3]

Received: January 26, 2015, Accepted: August 27, 2015, Released: November 26, 2015

Abstract: In the realm of multi-modal visual recognition, the reliability of the data acquisition system is often a concern due to the increased complexity of the sensors. One of the major issues is the accidental loss of one or more sensing channels, which poses a major challenge to current learning systems. In this paper, we examine one of these specific missing data problems, where we have a main modality/view along with an auxiliary modality/view present in the training data, but merely the main modality/view in the test data. To effectively leverage the auxiliary information to train a stronger classifier, we propose a collaborative auxiliary learning framework based on a new discriminative canonical correlation analysis. This framework reveals a common semantic space shared across both modalities/views through enforcing a series of nonlinear projections. Such projections automatically embed the discriminative cues hidden in both modalities/views into the common space, and better visual recognition is thus achieved on the test data. The efficacy of our proposed auxiliary learning approach is demonstrated through four challenging visual recognition tasks with different kinds of auxiliary information.

Keywords: collaborative auxiliary learning, canonical correlation, discriminative learning

1. Introduction

In the realm of multi-spectral/multi-modal visual recognition, we focus on an under-explored missing data problem with visual data of two views [*1], where a main view and an auxiliary view are present in the training data, but only the main view is available in the testing data. This particular problem attracts our attention, due to the recent popularity of multi-sensory and multiple spectrum imaging systems, such as the depth-sensing and infrared (IR) sensing cameras, and hence the accumulation of labeled multi-view visual data.

However, the increased complexity of multiple data acquisition sensors incurs reduced reliability, especially the relatively new IR/depth (i.e., the auxiliary) sensors. The higher probability of failure of the auxiliary view sensors prompts us to investigate proper measures to alleviate the impacts of the missing data.

Before a typical real-world deployment of these multi-view sensing systems, multi-view, labeled training data is often collected beforehand and readily available to the learning system. However, due to the failure-prone nature of the auxiliary sensors, the testing data could be completely deprived of the auxiliary view. This "missing-of-auxiliary-view" problem could occur in the test phase, possibly incurred by, for example, adversarial sensing environments or insufficient bandwidth allowing the transmission of only the main view data. In addition, the "missing view" problem could also arise while processing a backlog of historical data without the auxiliary sensing channel during the time of collection. Therefore, a question naturally emerges: How can the auxiliary view in the training data assist the visual recognition based only on the main view?

Unlike conventional learning assumptions where the training data and the testing data follow the identical probability distribution [26], this problem requires techniques that can incorporate the beneficial information from the auxiliary view into the training of a classification model that works only on the main view. We shall emphasize that this problem is different from the domain adaptation or the transfer learning problems [9], [16], [28]. The goal of most domain adaptation/transfer learning problems is to leverage existing abundant labeled data in one domain to facilitate learning a better model in the target domain with scarce labeled data, if at all. Essentially, the knowledge is transferred from one data domain to a related but statistically different one. In contrast, in our problem the data domain in the test phase (i.e., the main view) is a proper subset of the training data domain that contains the auxiliary view other than the main view.

In this sense, the problem of this paper is more closely related to the multi-view learning problems, but with the additional missing data obstacle. The previous works in multi-view learning [10], [25], [40] have demonstrated that synergistically modeling of all views leads to improved performance. As a matter of fact, the existing multi-view recognition methods [1], [10], [36] emphasize heavily on properly combining per-view information. In this paper, we adopt a verification-by-construction approach

[1] Stevens Institute of Technolog, NJ, USA
[2] IBM Thomas J. Watson Research Center, NY, USA
[3] Microsoft Research, Redmond, WA, USA
[4] Microsoft Research Asia, Beijing, China
[a)] qzhang5@stevens.edu
[b)] ganghua@gmail.com

[*1] Following the definition in the multi-view learning problems, multiple views in this paper are defined as multiple semantically related sensing modalities, such as images, depth maps and infrared images of the same object.

by showing that there exists at least one method that consistently outperforms in terms of recognition accuracy on four missing-data multi-view visual recognition datasets with various settings of view assignments, features and parameters.

In particular, we propose to seek a common latent space to capture the semantic relevance between the main view and the auxiliary view. This is achieved by a new discriminative canonical correlation analysis (DCCA) algorithm inspired by Ref. [34]. The new DCCA algorithm not only takes supervised label information into consideration, but also concurrently optimizes multiple nonlinear projections with a guaranteed convergence. Our DCCA algorithm is parallel to and even exceeds most previous CCA algorithms which do not explicitly exploit the label information or pursue multiple projections repeatedly in a one-by-one manner. After obtaining the desired common latent space, the auxiliary view information in the training data is carried into a classifier defined in this latent space. Subsequent tests are conducted by projecting the only available main view information of the test data onto the common space and then applying the classifier trained on the aforementioned latent space.

The primary contributions of this paper are:

(1) we focus on an under-explored multi-view visual recognition problem, i.e., the "missing-view-in-test-data" problem with real-world multisensory visual datasets;

(2) we propose a new discriminative canonical correlation algorithm together with a rigorous convergence proof, and its efficacy is validated on four benchmarks.

The rest of the paper is organized as follows. Section 2 briefly summarizes the related work. Section 3 formally defines the "missing-view-in-test-data" problem. Section 4 formulates the proposed DCCA algorithm. Section 5 gives a solution to the DCCA algorithm. Section 6 presents the experiments. Section 7 concludes the paper. The convergence proof is also included in the appendix of the paper.

2. Related Work

In this section, we summarize the related work in domain adaptation, transfer learning, multi-view learning, and also highlight their differences from the "missing-view-in-test-data" problem we investigate in the paper.

With the encouraging successes in applying metric learning techniques to various problems (e.g., machine translation [9], [10], multimedia information retrieval [13], and visual recognition [16], [28]), the metric learning methods have also found its way into the visual recognition tasks, especially in the form of domain adaptation and transfer learning [26], [28].

Among most of these domain adaptation and transfer learning visual recognition problems, it is often assumed that abundant labeled training samples are present in the source domain; while in the target domain, there are a limited number of labeled training samples. The two domains are related in some semantic sense, but in terms of the low-level features, the target and source domains have statistically different distributions. Therefore, it is infeasible to directly train an effective classifier using the scarce labeled data in the target domain or train the classifier using the labeled samples in the source domain of a different statistical distribution.

To overcome the aforementioned difficulties, these works (e.g., [16], [28]) attempt to transfer discriminative information from the source domain to the target domain, exploiting the corresponding relationship between the two domains to build a regularized cross-domain transform via techniques such as metric learning and kernel learning [17]. However, these domain transfer problems are different from the "missing-view-in-test-data" problem we are tackling in the paper. In our problem, instead of having the source/target domains, we have the auxiliary view and main view. In the training phase, both views are labeled but in the testing phase, only unlabeled main view is available.

Considering the existence of the main view and auxiliary view in our focused problem of this paper, it is related to the multi-view and multi-task learning problems. However, in our "missing-view-in-test-data" problem, there is systematic missing of testing data, leaving only the main view data for testing. The typical multi-view and multi-task learning problems, e.g., [1], [10], [36], require some principled fusion schemes which combine information from two or more related but statistically different views/tasks to achieve certain goals. Since the introduction of the early co-training framework [2] in dealing with multi-modal data, various techniques have been proposed in this specific application area. A typical example is the SVM-2K algorithm [10]. It was demonstrated in Ref. [10] that the recognition performance does benefit from explicitly leveraging the underlying semantic consistency among multiple views. In our experiments shown in Section 6, a simple modification of the SVM2K algorithm [10] is implemented and treated as a competing baseline. However, the additional missing data obstacle is generally not considered in these multi-view learning problems.

In Ref. [8], both the missing data and the domain transfer problems are considered, however, the objective function ignores the discriminative label information. A different RGBD-RGB data based object detection problem is addressed in Ref. [30] where explicit 3D geometry information is modeled. Tommasi et al. [35] focus on the dataset bias problem with a similar latent space model. In Ref. [11], a different missing feature problem is addressed, where the features are missing randomly in all dimensions, instead of the systematic absence of views in the multi-view settings considered in this paper. In the deep learning community, multi-modal deep learning systems such as Ref. [32] also show the robustness against missing views.

In another related field, Chen et al. [7] suggest boosting-based learning with side information. Unlike our proposed latent space model, this method is not straightforward in terms of handling multiple sets of side information. In the area of human perception and psychology, Shams et al. [29] show that humans recognize better in unisensory tests with previous multisensory experiences than with the unisensory counterparts. Through establishing our proposed auxiliary learning framework, we show that this benefit also exists in the context of machine recognition.

We also note that there exists a discriminative CCA [15] proposed for image set recognition where each sample consists of many observations/views. However, this CCA method exploits the correlations between sets of observations, which require abun-

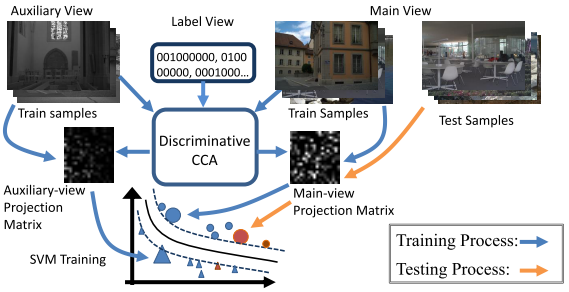
dant views for the robust estimation of sample covariance matrices. In contrast, our missing-view-in-test-data problem involves many observations but only a few views and even a single view in the test phase.

3. Missing View Resilient Latent Space Model

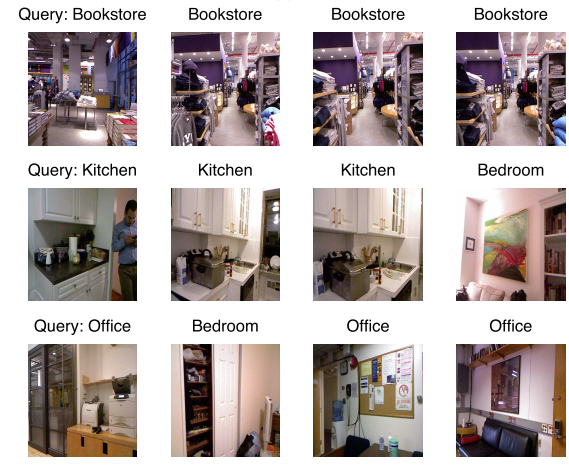
Suppose that m sensing channels are collecting data on a specific task, these observations are denoted as $\mathbf{z}(i)=(\mathbf{x}_1(i), \mathbf{x}_2(i))$ $(i = 1, 2, \cdots, m)$. Without loss of generality, Let $\mathbf{x}_1(i) \in \mathcal{X}_1$ and $\mathbf{x}_2(i) \in \mathcal{X}_2$ denote observations obtained from the main view and auxiliary view, respectively. We assume $\mathcal{X}_1 \subset \mathbb{R}^{n_1}$ and $\mathcal{X}_2 \subset \mathbb{R}^{n_2}$, and define $\mathcal{Z}=\mathcal{X}_1 \times \mathcal{X}_2$, so $\mathbf{z}(i) \in \mathcal{Z}$. Within the training dataset, there comes with a label $l(i) \in \mathcal{L}$ for each observation $\mathbf{z}(i)$.

Unlike the conventional supervised learning problem settings where the classifier $f_\theta : \mathcal{Z} \to \mathcal{L}$ is normally obtained by the inference of the parameters $\theta = \theta(\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(m))$ based on the training observations, the "missing-view-in-test-data" problem requires a classifier in the form of $\tilde{f}_{\tilde{\theta}} : \mathcal{X}_1 \to \mathcal{L}$ due to the missing auxiliary view in the test data. To effectively incorporate the available information in the entire training set $\{\mathbf{z}(i), l(i)\}_{i=1}^m$, this paper focuses on constructing the classifier $\tilde{f}_{\tilde{\theta}}$, where $\tilde{\theta} = \tilde{\theta}(\mathbf{z}(1), \mathbf{z}(2), \cdots, \mathbf{z}(m), l(1), l(2), \cdots, l(m))$.

To capture the information hidden in the auxiliary view $\mathbf{x}_2(j)$ and training labels $l(j)$, we propose to construct an intermediate space $\mathcal{S}$ to maximally retain the discriminative information from both views. During this process, we construct projections



(a)



(b)

Fig. 1 (a) A latent space model to address the "missing-view-in-test-data" problem. The training and test processes are displayed in blue and yellow arrows, respectively. (b) k-NN retrieval from the NYU depth Indoor scenes data base. Actual scene class labels are listed above each image.

$p_1(\cdot)$ and $p_2(\cdot)$ that map $\{\mathbf{x}_1(i)\}_{i=1}^m$ and $\{\mathbf{x}_2(i)\}_{i=1}^m$ to $p_1(\mathbf{x}_1(i)) \in \mathcal{S}$ and $p_2(\mathbf{x}_2(i)) \in \mathcal{S}$, respectively. In $\mathcal{S}$, the classification problem becomes significantly easier, as illustrated in **Fig. 1** (a). A conventional supervised classifier $f : \mathcal{S} \to \mathcal{L}$ (e.g., the SVM classifier [5]) can be obtained by training with $\{p_1(\mathbf{x}_1(i)), l(i)\}_{i=1}^m \cup \{p_2(\mathbf{x}_2(i)), l(i)\}_{i=1}^m$. In the test phase, the test samples $\{\hat{\mathbf{x}}_1(j)\}_{j=1}^k$ are first projected to $p_1(\hat{\mathbf{x}}(j)) \in \mathcal{S}$ and subsequently fed to the trained classifier $f : \mathcal{S} \to \mathcal{L}$. In Fig. 1 (a), the training processes and the test processes are shown in blue arrows and yellow arrows, respectively.

The success of the aforementioned latent space based approach depends on not only the maximal preservation of the congruent information among views, but also the discriminative label information acquired in constructing $\mathcal{S}$. To achieve this, we propose a discriminative canonical correlation analysis (DCCA) algorithm, which simultaneously extracts multiple CCA projections and also incorporates the label information. In the following section, we formally formulate the optimization problem of DCCA and compare our proposed optimization method against previous ones.

4. DCCA Formulation

In this section, we formulate the DCCA algorithm and compare it with previous related ones. The classical Canonical Correlation Analysis (CCA) (see Refs. [13], [14], [38]) and its variants have been popular among practitioners for decades. In its basic form, the CCA algorithm finds a pair of linear projection vectors that maximize the linear correlation:

$$\max_{\alpha_j, \alpha_k} \quad \alpha_j^T \Sigma_{jk} \alpha_k \tag{1}$$

$$\text{s.t.} \quad \alpha_j^T \Sigma_{jj} \alpha_j = 1 \quad (j, k = 1, 2, \cdots, J) \tag{2}$$

where $\Sigma_{jk} = \mathcal{E}(\mathbf{x}_j \mathbf{x}_k^T)$, $\mathcal{E}$ denotes the mathematical expectation, and $\mathbf{x}_j$ denotes the observations in the jth view ($j = 1, 2, \cdots, J$, J denotes the number of distinct views in the multi-view data). In this way, α_j and α_k project the jth and kth views to a one-dimensional space by computing the inner products $\alpha_j^T \hat{\mathbf{x}}_j$ and $\alpha_k^T \hat{\mathbf{x}}_k$, respectively. In practice, a latent space of more of dimension d ($d > 1$) is often desired, hence this procedure in Eq. (2) needs to be repeated d times. Alternatively, as shown in Ref. [13], the following formulation can be used equivalently:

$$\max_{A_j, A_k} \quad \text{tr}(A_j^T \Sigma_{jk} A_k) \tag{3}$$

$$\text{s.t.} \quad A_j^T \Sigma_{jj} A_j = \mathbf{I} \quad (j, k = 1, 2, \cdots, J) \tag{4}$$

where $A_j = [\alpha_j(1), \cdots, \alpha_j(d)]$ contains all d projection vectors to form a d-dimensional latent space. The projections are computed as $A_j^T \hat{\mathbf{x}}_j$ and $A_k^T \hat{\mathbf{x}}_k$. To preserve the semantic consistency between the jth and kth views, we strive to make sure that $A_j^T \hat{\mathbf{x}}_j$ and $A_k^T \hat{\mathbf{x}}_k$ are as similar as possible, therefore maximizing $\text{tr}(A_j^T \Sigma_{jk} A_k)$ as Eq. (4).

These CCA algorithms are extended to handle more than two views, which are generally termed as the generalized CCA (gCCA) [13], [27]. However, among the gCCA algorithms where per-view "correlations" are combined nonlinearly, most of them cannot be solved in a closed form and require a computing scheme to optimize the projections one by one. Also, the naive

version of gCCA ignores the discriminative information in the class labels provided in the training process.

To address these two potential issues, we propose a new algorithm, named as *Discriminative CCA* (DCCA), which is inspired by Ref. [34]. It not only inherits the guaranteed optimization convergence property, but also has two extended properties: 1) the simultaneous optimization of all CCA projections coefficients contributes to numerically stable and optimal results; 2) the pursuit of DCCA incorporates the label view information. Empirically, we find that the combination of 1) and 2) leads to performance gains across four challenging visual datasets in Section 6.

Starting from leveraging discriminative information like Loog et al. [22], we first encode the training labels as the Jth view based on k-nearest neighbors (k-NNs) of the training samples, where observations from all views are normalized and concatenated and then used in computing k-NNs. Suppose that there are C classes. For each labeled training sample $\mathbf{x}(0)=[\mathbf{x}_1(0)^T,\cdots,\mathbf{x}_J(0)^T]^T$, we consider its k-NNs $\mathbf{x}(i)=[\mathbf{x}_1(i)^T,\cdots,\mathbf{x}_J(i)^T]^T$ ($i=0,\cdots,k$, including itself $\mathbf{x}(0)$). We record their labels as $\mathbf{L}_{(i)}$ ($i=0,\cdots,k$) accordingly. The encoded label view for this training sample is a binary vector consisting of $k+1$ length-C blocks.

After obtaining the label view, we can simultaneously optimize all d projections in the form of A_j ($j=1,2,\cdots,J$) as follows:

$$\arg\max_{\mathbf{A}} \quad \sum_{j,k=1,j\neq k}^{J} c_{jk}\, g\left(\mathrm{tr}(A_j^T\Sigma_{jk}A_k)\right) \tag{5}$$

$$\text{s.t.} \quad A_j^T\Sigma_{jj}A_j=\mathbf{I} \quad (j=1,2,\cdots,J) \tag{6}$$

where $\mathbf{A}=[A_1,\cdots,A_J]$, $A_j=[\alpha_j(1),\cdots,\alpha_j(d)]$ projects $\mathbf{x}_j$ to the d-dimensional common semantic space, c_{jk} is the selecting weights (either 0 or 1), and function $g(\cdot)$ is the view combination function (e.g., direct combination $g(x)=x$, or squared combination $g(x)=x^2$). The space dimension d is chosen such that each diagonal element of $A_j^T\Sigma_{jk}A_k$ is positive, which implies that the observations from all views are positively correlated. The Lagrangian of Eq. (6) is

$$F(\mathbf{A},\mathbf{\Lambda}) = \sum_{k\neq j} c_{jk}\, g\left(\mathrm{tr}(A_j^T\Sigma_{jk}A_k)\right) \tag{7}$$

$$-\phi\sum_{j}\frac{1}{2}\mathrm{tr}\left(\mathbf{\Lambda}_j^T(A_j^T\Sigma_{jj}A_j-\mathbf{I})\right) \tag{8}$$

where $\mathbf{\Lambda}=[\mathbf{\Lambda}_1,\cdots,\mathbf{\Lambda}_J]$, $\mathbf{\Lambda}_j\in\mathcal{R}^{d\times d}$ is the multiplier, and ϕ is a scalar which is equal to 1 if $g(x)=x$ and 2 if $g(x)=x^2$. The derivative of $g(x)$ is denoted as $g'(x)$.

From Eq. (8), the following stationary eqnarrays hold:

$$\frac{1}{\phi}\sum_{k\neq j} c_{jk}g'\left(\mathrm{tr}(A_j^T\Sigma_{jk}A_k)\right)\Sigma_{jk}A_k=\Sigma_{jj}A_j\mathbf{\Lambda}_j \tag{9}$$

$$A_j^T\Sigma_{jj}A_j=\mathbf{I} \tag{10}$$

In practice, a kernelized version of DCCA is often favored, because linear correlations are not sufficient in modeling the nonlinear interplay among different views. Suppose that $\mathbf{K}_j$ is the Gram matrix of the centered data points $[\mathbf{x}_1,\cdots,\mathbf{x}_J]$. The empirical covariance is $\frac{1}{n}\alpha_j^T\mathbf{K}_j\mathbf{K}_k\alpha_k$, where α_j is the coefficient vector and n is the number of training samples. Let $A_j=[\alpha_j(1),\cdots,\alpha_j(d)]$ and the empirical covariance matrix be $\frac{1}{n}A_j^T\mathbf{K}_j\mathbf{K}_kA_k$. Suppose that $\mathbf{K}_j$ can be decomposed as $\mathbf{K}_j=\mathbf{R}_j^T\mathbf{R}_j$. We define the projection matrix $\mathbf{W}_j=\mathbf{R}_jA_j$, and similarly define $\mathcal{W}=[\mathbf{W}_1,\cdots,\mathbf{W}_J]$. The optimization objective function of the kernelized DCCA is

$$\arg\max_{\mathcal{W}} \quad \sum_{j,k=1,j\neq k}^{J} c_{jk}\, g\left(\mathrm{tr}(\tfrac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)\right) \tag{11}$$

$$\text{s.t.} \quad \mathbf{W}_j^T\left[(1-\tau_j)\tfrac{1}{n}\mathbf{R}_j\mathbf{R}_j^T+\tau_j\mathbf{I}_n\right]\mathbf{W}_j=\mathbf{I} \tag{12}$$

Let us define $N_j=(1-\tau_j)\frac{1}{n}\mathbf{R}_j\mathbf{R}_j^T+\tau_j\mathbf{I}_n$, where $0<\tau<1$ is a pre-specified regularization parameter. Similar to Eq. (10), the following stationary eqnarrays hold

$$\frac{1}{\phi}\sum_{k\neq j} c_{jk}g'\left(\mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)\right)\frac{1}{n}\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k=N_j\mathbf{W}_j\mathbf{\Lambda}_j \tag{13}$$

$$\mathbf{W}_j^TN_j\mathbf{W}_j=\mathbf{I} \tag{14}$$

whose solution [*2] is presented in Section 5.

5. DCCA Solution

In this section, a monotonically convergent iterative algorithm to solve the DCCA optimization problem is presented. For generic $g(\cdot)$ format and c_{jk} value assignments, there are no closed-form solutions to Eq. (10) or Eq. (13). However, following [34], a similar "PLS-type", monotonically convergent iterative algorithm can be formulated. For conciseness, we only present the details of this algorithm with the solution to the problem in Eq. (13). Define the outer component $\mathbf{Y}_j$ and inner component $\mathbf{Z}_j$, respectively, as

$$\mathbf{Y}_j=\mathbf{R}_j^T\mathbf{W}_j \tag{15}$$

$$\mathbf{Z}_j=\frac{1}{\phi}\sum_{k\neq j} c_{jk}g'\left(\mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)\right)\mathbf{Y}_k \tag{16}$$

Differentiating the Lagrangian with respect to $\mathbf{W}_j$ and setting the gradient to zero, we obtain

$$\mathbf{R}_j\mathbf{Z}_j=N_j\mathbf{W}_j\mathbf{\Lambda}_j \tag{17}$$

$$\mathbf{W}_j^TN_j\mathbf{W}_j=\mathbf{I} \tag{18}$$

From Eq. (18) and Eq. (16), we have

$$\mathbf{\Lambda}_j=\mathbf{W}_j^T\mathbf{R}_j\mathbf{Z}_j \tag{19}$$

$$=\frac{1}{\phi n}\sum_{k\neq j} c_{jk}g'\left(\mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)\right)\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k \tag{20}$$

where $\mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)$ is assumed to be positive, $c_{jk}=0$ or 1, and due to the definition of g', $\mathbf{\Lambda}_j$ is a positive semi-definite matrix. From Eq. (18) and Eq. (20), we have $\mathbf{\Lambda}_j^T\mathbf{\Lambda}_j=\mathbf{Z}_j^T\mathbf{R}_j^TN_j^{-1}\mathbf{R}_j\mathbf{Z}_j$. Since $\mathbf{\Lambda}_j$ has non-negative eigenvalues, $\mathbf{\Lambda}_j$ can be obtained via the matrix square root $[\mathbf{Z}_j^T\mathbf{R}_j^TN_j^{-1}\mathbf{R}_j\mathbf{Z}_j]^{1/2}$. Therefore,

$$\mathbf{W}_j=N_j^{-1}\mathbf{R}_j\mathbf{Z}_j\left(\left[\mathbf{Z}_j^T\mathbf{R}_j^TN_j^{-1}\mathbf{R}_j\mathbf{Z}_j\right]^{1/2}\right)^{\dagger} \tag{21}$$

where † denotes the pseudoinverse of a matrix.

A monotonically convergent iterative algorithm is described in Algorithm 1. Let $f(\mathcal{W})=\sum_{k\neq j}c_{jk}g\left(\mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)\right)$. Importantly, we have the following proposition:

Proposition 1

$f(\mathcal{W}(s=1))\leq f(\mathcal{W}(s=2))\leq f(\mathcal{W}(s=3))\leq\ldots\leq C_u<\infty$ holds for all $s\in\mathcal{N}$, where s denotes the iteration index and C_u is a constant bound. This guarantees Algorithm 1 to converge monotonically. The proof of Proposition 1 is included in the appendix.

[*2] Note that Eq. (10) is a linear version of Eq. (13) and has a very similar solution. For conciseness, the solution to Eq. (10) is omitted.

Algorithm 1: A monotonically convergent iterative algorithm for DCCA

Input: Observations from all J views: $\mathbf{x}_j, j = 1, \cdots, J$.
Output: J projection matrices $\mathbf{W}_j, j = 1, \cdots, J$.
Initialization:
Randomly initialize $\mathbf{W}_j(0)$, normalize them by $\mathbf{W}_j(0) \leftarrow N_j^{-1}\mathbf{W}_j(0)\left(\left[\mathbf{W}_j(0)^T N_j^{-1}\mathbf{W}_j(0)\right]^{1/2}\right)^{\dagger}$, and compute the initial outer components $\mathbf{Y}_j(0)$ by Eq. (15).
for $s = 0, 1, \cdots$ *until the convergence of* $\mathbf{W}_j$ **do**
  for $j = 1, 2, \cdots J$ **do**
    update the inner components by Eq. (16):
    $\mathbf{Z}_j(s) \leftarrow \frac{1}{\phi}\sum_{k=1}^{j-1} c_{jk} g'(\mathrm{tr}(\frac{1}{n}\mathbf{Y}_j^T(s)\mathbf{Y}_k(s+1)))\mathbf{Y}_k(s+1) + \frac{1}{\phi}\sum_{k=j+1}^{J} c_{jk} g'(\mathrm{tr}(\frac{1}{n}\mathbf{Y}_j^T(s)\mathbf{Y}_k(s)))\mathbf{Y}_k(s)$;
    update the outer weighs by Eq. (21):
    $\mathbf{W}_j(s+1) \leftarrow N_j^{-1}\mathbf{R}_j\mathbf{Z}_j(s)\left(\left[\mathbf{Z}_j(s)^T\mathbf{R}_j^T N_j^{-1}\mathbf{R}_j\mathbf{Z}_j(s)\right]^{1/2}\right)^{\dagger}$;
    update the outer components by Eq. (15):
    $\mathbf{Y}_j(s+1) \leftarrow \mathbf{R}_j^T\mathbf{W}_j(s+1)$;
  end
end
return

6. Experiments

In this section, the efficacy of the proposed DCCA algorithm is verified in four visual recognition tasks and we also isolate each component of the DCCA algorithm and analyze their contributions to the overall performance.

6.1 Compared Methods and Datasets

We compare the performances of the kernelized DCCA algorithm (paired with the RBF SVM classifier) against the following algorithms:

- SVM: The vanilla "SVM" is trained on the main view of the visual information only.
- SVM-2K*: This is a variant of "SVM-2K" [10], which we train with two views in the train phase but only use one of the classifier due to the missing auxiliary view in the test phase [*3].
- KCCA: This is the kernel CCA based on the main view and the auxiliary view.
- KCCA+L: a kernel CCA based on the main view and the encoded label view, it ignores the auxiliary view.
- RGCCA: Regularized Generalized Canonical Correlation Analysis [34], a kernel variant of the regularized generalized CCA based on the main view and the auxiliary view. As an extended version of gCCA, it iteratively optimizes each one-dimensional projection vectors (one column of $\mathbf{W}_j$), and all d columns of $\mathbf{W}_j$ are pursued one by one, similar to the algorithm presented in Ref. [38].
- RGCCA+L: similar to "RGCCA", except it is based on the main view and the encoded label view.
- RGCCA+AL: similar to the proposed DCCA, except it iteratively optimizes each one-dimensional projection vectors (one column of $\mathbf{W}_j$).

In selecting the competing algorithms, we choose the "SVM-2K*" variant to represent the application of a crudely modified multi-view learning algorithm. We select "KCCA" and "RGCCA" to represent classical and recent variants of the CCA-type algorithms, both of which are based on the correlation between the main view and the auxiliary view, i.e., without explicitly considering the label view. In addition, to isolate the effects of the auxiliary view and encoded train label view, we have also included a series of semi-finished DCCA algorithms, i.e., the "KCCA+L", "RGCCA+L" and "RGCCA+AL", which have different components of the proposed DCCA algorithms. A brief comparison table of these algorithms are given in **Table 1**.

In the following experiments, the Radial Basis Function (RBF) kernel is applied in both the KCCA algorithm and the SVM algorithms. Parameters such as c_{jk}, $g(\cdot)$, the bandwidth parameter in the RBF kernel, and those parameters in the SVM-2K* algorithm, are all selected by a 4-fold cross validation. The experiments are conducted on four different datasets, i.e., the "NYU Depth V1" Indoor Scenes dataset ([31]), the RGBD Object dataset [18], the multi-spectral scene dataset [4] and the Binghamton University 3D Facial Expression dataset [39].

The NYU Depth V1 dataset consists of RGBD images of indoor scenes collected by a modified Kinect sensor [31]. With this dataset, we demonstrate that the depth information in the train phase can benefit the scene classification based solely on the RGB images. The RGBD object dataset [18] consists of a large collection of paired RGB images and depth maps of common objects. We focus on the instance level object recognition, and demonstrate that the additional depth information during the train phase can facilitate better recognition based on the RGB information only. The multi-spectral scene dataset [4] consists of 477 registered and aligned RGB and near-infrared (IR) images from 9 different scene categories, i.e., country, field, forest, indoor, mountain, old-building, street, urban and water. In this experiment, we demonstrate that the auxiliary information hidden in the IR channel can help to train a better scene recognition model that operates only on the main view. The Binghamton University 3D Facial Expression dataset [39] consists of 3D face models and corresponding face images. With this dataset, we focus on three tasks of visual recognition (gender, ethnic/racial ancestries, expressions) with the aforementioned "missing-view-in-test-data" scenarios.

Table 1 A comparison of the competing algorithms and the proposed algorithm in terms of training data and CCA-optimization method. √ and × denote the specific part of training data is explicitly used and ignored in the algorithm, respectively. There are no CCA-optimization involved in the SVM or SVM-2K*, so 'N/A' is used, meaning 'Not Applicable'.

Algorithms	Main View	Aux. View	Label View	CCA-Optimization
SVM	√	×	×	N/A
SVM-2K* [10]	√	√	×	N/A
KCCA	√	√	×	Eigen decomposition
KCCA+L	√	×	√	Eigen decomposition
RGCCA [34]	√	√	×	vector by vector
RGCCA+L	√	×	√	vector by vector
RGCCA+AL	√	√	√	vector by vector
DCCA	√	√	√	simultaneous

[*3] The original form of SVM-2K is not directly applicable to the missing view problem

Table 2 NYU Depth V1 Indoor Scenes Classification, the highest and second highest values are colored red and blue, respectively.

Features	GIST		Spatial Pyramid, K=200		Spatial Pyramid, K=800	
Views	L+D	RGB+D	L+D	RGB+D	L+D	RGB+D
SVM	59.57±3.31	60.79±3.12	64.11±3.11	64.71±3.95	64.73±2.79	65.34±3.18
SVM-2K* [10]	57.52±3.88	60.01±3.71	59.62±3.23	60.42±4.55	58.00±3.58	60.64±3.84
KCCA	58.16±6.55	62.58±3.55	64.94±4.58	64.00±4.92	64.77±4.69	65.01±4.86
KCCA+L	58.48±3.37	59.95±3.62	62.99±3.80	60.67±4.23	62.26±3.56	60.55±4.40
RGCCA [34]	58.66±5.93	59.75±4.11	60.49±5.21	60.31±5.75	61.70±4.00	60.42±3.68
RGCCA+L	59.12±4.11	59.82±4.50	63.34±4.18	62.48±3.49	63.81±4.51	61.04±4.99
RGCCA+AL	**59.82**±6.10	**62.85**±4.24	**65.61**±4.22	**65.31**±4.23	**65.38**±4.22	**65.66**±3.04
DCCA	**60.26**±3.86	**63.60**±3.43	**66.20**±3.69	**65.35**±4.72	**66.09**±4.18	**66.28**±4.16

6.2 NYU-Depth-V1-Indoor Scene Dataset

On the NYU Depth V1 indoor scenes dataset [31], we carry out the multi-spectral scene recognition task. Following Ref. [31], these observations are randomly split into 10 folds with approximately equal number of the training samples and the testing samples. Subsequently, we extract both the GIST [24] features and the spatial pyramid bag-of-feature representation [19] independently from each imaging channel (Red, Green, Blue and Depth channels). For the latter case concerning the spatial pyramid, we densely extract SIFT descriptors from 40×40 patches with a stride of 10 pixels, and use two k-means dictionary sizes of 200 and 800, which are denoted as "Spatial Pyramid, K=200" and "Spatial Pyramid, K=800" in **Table 2**, respectively. While grouping the imaging channels into views, we investigate the following two settings:

- **L+D**: Grayscale image features are assigned as the main view, while the depth features are assigned as the auxiliary view.
- **RGB+D**: RGB image features are concatenated and assigned as the main view, while the depth features are assigned as the auxiliary view.

We first demonstrate the *k*-NN retrieval results in Fig. 1 (b) to visualize some typical images from this dataset. The query images (from the test set) are displayed on the left and the corresponding 3 nearest neighbors (in the train set) are displayed on the right. As demonstrated in Fig. 1 (b), this dataset consists of highly cluttered indoor scenes, making this scene recognition task a challenging one.

In Table 2, the means and standard deviations of the recognition accuracy (in percentage) are reported. We observe that higher dimensional features generally offer better recognition accuracy and the color information also helps recognition slightly. Generally, experiments based on the "RGB+D" features achieve slightly higher accuracies than their "L+D" counterparts. In Table 2, the SVM-2K* variant produces lower accuracies than the SVM baseline, we speculate that the loose "prediction consistency" regularization of SVM-2K* variant is only helpful when two views satisfy certain distributions. In addition, neither the KCCA nor the RGCCA approach sees significant performance improvements. With the encoded label view alone, neither the "KCCA+L" nor the "RGCCA+L" algorithm achieves any advantage over the SVM baseline. Intuitively, the information embedded in the encoded label view is far less significant than that in the auxiliary view. However, with both the label view and the auxiliary view, the "RGCCA+AL" algorithm is capable of achieving a small advantage over the baseline, though not as significant as the proposed DCCA algorithm, whose projections are optimized and computed simultaneously and more accurately. The large standard deviation in Table 2 stem from the difficult nature of the datasets, which can be estimated from the baseline SVM performance in Table 2.

6.3 RGBD Object Dataset

With this RGBD Object dataset from Ref. [18], we focus on the instance level object recognition. There are multiple instances of the same class of object for all the 51 object categories; the recognition target is to correctly recover the object instance labels in the test set. We follow the "leave-one-sequence-out" scheme detailed in Ref. [18] and split recordings with camera mounting angle of 30° and 60° as the training set and the remaining as the testing set (the train/test sets are fixed, hence the standard deviation are not applicable). In this section, we present the results with both the EMK-based features [18] and the state-of-the-art HMP-based features [3] extracted from the RGB channels as the main view, and from the depth channel as the auxiliary view, respectively.

As is seen in **Table 3**, the recognition accuracy (in percentage) using the EMK-based features within each category fluctuates significantly. This is due to the characteristic of the object categories: e.g., different instances of limes resemble each other a lot more than different instances of cereal boxes with different logos.

With some of the easy categories (e.g., "pitcher"), the baseline SVM algorithm already achieves perfect recognition. However, with some of the challenging categories (e.g., "food bag" and "lime"), the proposed DCCA offers the most significant performance boost. Overall, the "KCCA+L" and the "RGCCA+L" algorithms achieve a small advantage over the SVM baseline, both of which are inferior to the RGCCA algorithm that only maximizes the main view and auxiliary view correlation. However, the "RGCCA+AL" algorithm performs much better, though not as good as the proposed DCCA algorithm. Among the 51 categories in Table 3, the "RGCCA+AL" algorithm achieves the best and second best accuracies in 8 and 24 categories, earning itself an overall average accuracy of 85.7%. The proposed DCCA achieves the best and second best recognition accuracies in 28 and 19 categories, acquiring an average accuracy of 86.6% across all categories, highest among all algorithms. Alternatively, with the new HMP-based features [3], the recognition results are summa-

Table 3 Accuracy Table for the Multi-View RGBD Object Instance recognition with EMK features, the highest and second highest values are colored red and blue, respectively.

Category	SVM	SVM-2K* [10]	KCCA	KCCA +L	RG-CCA [34]	RG -CCA +L	RG-CCA +AL	DCCA
apple	65.2	72.4	**77.6**	64.8	**79.0**	68.1	76.7	**77.6**
ball	95.9	97.3	97.8	**99.5**	94.2	95.3	97.8	**98.4**
banana	74.2	61.1	**80.3**	71.7	**77.3**	**77.3**	**80.3**	**80.3**
bell pepper	**71.3**	59.8	**69.3**	65.7	66.1	68.9	**69.3**	**69.3**
binder	67.3	36.7	74.1	52.4	**75.5**	68.7	74.1	**74.8**
bowl	85.0	85.8	87.7	81.2	**90.0**	86.2	**88.1**	**88.1**
calculator	**99.4**	88.3	**99.4**	**100**	97.8	**99.4**	**99.4**	**99.4**
camera	91.7	49.6	**97.5**	90.1	**96.7**	95.9	**97.5**	**97.5**
cap	91.8	88.9	**95.9**	91.2	**96.5**	93.0	**95.9**	**95.9**
cellphone	93.2	81.7	**96.3**	93.2	94.2	**95.3**	**96.3**	**96.3**
cereal box	77.4	75.7	**82.5**	**89.8**	**82.5**	80.8	81.9	**82.5**
coffee mug	82.7	81.7	**89.2**	62.5	81.7	82.7	**87.3**	**89.2**
comb	97.3	96.0	**99.3**	**100**	98.7	98.7	**99.3**	**100**
dry battery	**90.3**	79.3	86.3	87.7	**92.5**	86.3	88.1	87.7
flashlight	77.1	75.0	80.3	74.5	78.2	77.7	**81.4**	**82.4**
food bag	72.7	69.1	80.8	82.5	84.9	77.7	**86.6**	**89.2**
food box	75.1	72.6	78.0	83.8	84.1	76.1	**84.4**	**86.4**
food can	70.0	63.7	70.7	78.2	73.7	66.2	**81.0**	**83.7**
food cup	87.1	86.4	84.9	**94.5**	83.8	84.2	90.1	**91.5**
food jar	84.5	81.0	**88.6**	86.1	85.8	85.4	88.3	**88.9**
garlic	**95.5**	**93.3**	92.2	**95.5**	89.0	91.2	92.8	**93.3**
glue stick	**100**	89.3	99.4	95.6	**100**	93.7	**99.7**	99.4
greens	75.7	70.3	**82.2**	**84.9**	74.6	80.5	81.1	**82.2**
hand towel	80.1	74.5	80.9	82.4	79.0	78.7	**83.9**	**85.4**
instant noodles	83.1	78.7	**97.1**	88.8	89.5	88.5	**95.4**	**97.1**
keyboard	88.1	82.7	90.6	88.1	**93.6**	88.1	**95.0**	**95.0**
kleenex	**96.6**	90.9	95.1	89.8	94.7	93.6	**95.8**	**95.8**
lemon	45.8	44.2	43.4	45.0	44.2	43.0	**51.8**	**53.0**
light bulb	93.2	92.5	**95.9**	**98.6**	**95.9**	95.2	95.2	**95.9**
lime	37.8	38.3	42.2	37.2	45.0	40.0	**48.3**	**50.0**
marker	**48.4**	39.3	46.0	**48.4**	**49.9**	46.2	46.6	46.9
mushroom	**100**	**99.4**	**100**	**100**	**100**	**99.4**	**100**	**100**
notebook	80.5	73.3	**86.5**	82.0	**85.7**	82.7	**85.7**	**86.5**
onion	91.5	86.4	88.6	**94.0**	93.1	88.6	**93.4**	**93.4**
orange	41.1	42.0	**57.0**	49.8	19.3	48.3	**51.7**	**57.0**
peach	**100**	74.2	**100**	97.4	**99.3**	98.7	**100**	**100**
pear	68.7	61.6	79.0	70.1	79.0	74.0	**81.9**	**84.0**
pitcher	**100**	**100**	**100**	**100**	**100**	**98.3**	**100**	**100**
plate	89.8	74.2	96.3	85.8	96.9	91.9	**97.6**	**98.0**
pliers	78.2	62.4	**86.5**	79.5	72.1	80.8	84.7	**87.3**
potato	62.9	65.6	**70.3**	**71.8**	64.5	67.6	69.1	**70.3**
rubber eraser	**99.0**	75.5	95.1	96.6	93.1	96.1	97.5	**98.5**
scissors	87.5	84.9	**96.7**	**96.7**	**96.1**	92.8	**96.1**	**96.7**
shampoo	84.5	82.9	94.5	82.3	**95.5**	89.0	94.2	**94.8**
soda can	89.6	87.8	91.0	**97.7**	92.8	88.7	95.5	**96.8**
sponge	76.4	64.8	75.2	69.4	**78.6**	72.8	76.4	**77.4**
stapler	71.8	67.7	73.3	70.6	74.2	71.5	**77.4**	**78.0**
tomato	81.9	70.0	79.6	76.8	**82.2**	77.1	**85.0**	**85.0**
tooth brush	**78.4**	69.6	76.8	70.1	**79.4**	74.7	77.8	77.8
tooth paste	84.5	68.3	**90.0**	86.0	81.9	87.1	88.9	**90.4**
water bottle	88.2	88.2	**90.6**	82.6	80.2	87.4	**89.3**	**90.6**
average	81.3	74.4	84.5	81.6	83.0	81.8	**85.7**	**86.6**

Table 4 Accuracy Table for the Multi-View RGBD Object Instance recognition with HMP features, the highest and second highest values are colored red and blue, respectively.

Category	SVM	SVM-2K* [10]	KCCA	KCCA +L	RG-CCA [34]	RG -CCA +L	RG-CCA +AL	DCCA
apple	87.62	**93.33**	**92.86**	77.14	92.38	87.62	**93.33**	**93.33**
ball	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
banana	73.74	77.78	**81.31**	**81.31**	**80.30**	75.25	**80.30**	**80.30**
bell pepper	82.07	**84.06**	78.88	77.29	73.31	80.88	73.31	**83.27**
binder	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
bowl	83.08	86.54	86.54	86.15	**87.31**	77.69	**87.69**	**87.69**
calculator	**100**	**100**	**100**	**99.44**	**100**	**100**	**100**	**100**
camera	**99.17**	**100**	**100**	**100**	**99.17**	**99.17**	**99.17**	**99.17**
cap	**100**	**100**	**100**	**100**	**100**	**97.08**	**100**	**100**
cellphone	**98.43**	95.29	**95.81**	94.76	**95.81**	**98.43**	**95.81**	**95.81**
cereal box	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
coffee mug	92.88	85.45	96.90	94.43	**99.69**	91.64	**99.69**	**100**
comb	96.67	**97.33**	**97.33**	96.67	**98.00**	95.33	**98.00**	**98.00**
dry battery	86.34	84.14	85.46	76.65	**87.67**	85.90	**88.55**	**88.55**
flashlight	97.34	97.34	**97.87**	95.21	**100**	97.34	**100**	**100**
food bag	**100**	**100**	**100**	**100**	**100**	**96.88**	**100**	**100**
food box	**100**	**99.84**	**100**	**99.84**	**100**	**100**	**100**	**100**
food can	98.08	93.02	94.39	81.67	**99.18**	98.08	**99.32**	**99.32**
food cup	96.32	**98.90**	97.06	96.32	97.43	97.06	97.43	**98.53**
food jar	**98.73**	**100**	**100**	91.46	**100**	97.15	**100**	**100**
garlic	90.91	**98.40**	**98.40**	**96.26**	95.99	91.98	95.99	95.99
glue stick	**100**	**100**	**100**	**100**	**100**	**99.68**	**100**	**100**
greens	76.76	89.73	89.19	85.41	85.41	81.08	**90.27**	**91.35**
hand towel	**99.63**	**100**	**100**	**100**	**100**	**99.63**	**100**	**100**
instant noodles	**100**	99.51	**99.76**	**99.76**	**100**	**100**	**100**	**100**
keyboard	94.55	95.05	95.05	94.55	**96.04**	94.55	**96.53**	**96.53**
kleenex	**98.87**	**97.36**	**97.36**	96.98	**98.87**	**98.87**	**98.87**	**98.87**
lemon	45.02	**47.41**	**47.41**	45.82	**47.81**	44.62	**47.81**	**47.81**
light bulb	87.67	**97.26**	**96.58**	95.21	91.78	89.04	92.47	92.47
lime	**67.78**	62.22	61.11	58.33	62.78	**68.33**	62.78	62.22
marker	**100**	93.93	98.26	96.75	99.57	**100**	**99.78**	99.57
mushroom	**99.42**	**100**	**100**	**100**	**100**	**99.42**	**100**	**100**
notebook	**100**	**100**	**100**	**100**	**99.62**	**100**	**100**	**100**
onion	**98.74**	94.64	95.90	95.90	96.53	**99.37**	**98.74**	**98.74**
orange	72.46	97.58	**99.03**	**98.07**	96.62	69.57	**99.03**	**99.03**
peach	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
pear	90.04	**93.24**	90.75	90.04	91.81	88.61	91.81	**93.95**
pitcher	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
plate	**99.66**	**100**	**100**	99.32	95.59	**100**	99.32	99.32
pliers	**92.58**	89.96	90.39	90.39	**93.45**	**92.58**	**93.45**	**93.45**
potato	63.71	**71.81**	**74.13**	65.64	65.25	67.57	71.04	71.43
rubber eraser	69.61	**75.49**	71.57	69.61	70.59	**74.02**	70.59	71.08
scissors	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
shampoo	**99.35**	**99.68**	**99.68**	98.39	**99.68**	94.19	**99.68**	**99.68**
soda can	**100**	**100**	**99.55**	99.10	**100**	**100**	**100**	**100**
sponge	99.49	**100**	**100**	**99.83**	**99.83**	**99.83**	**99.83**	**99.83**
stapler	80.12	73.59	76.56	76.56	81.90	81.90	**82.20**	**82.49**
tomato	69.69	**81.59**	**83.85**	**83.85**	80.45	68.56	80.45	80.45
tooth brush	**99.48**	**100**	**99.48**	96.91	98.97	**99.48**	98.97	98.97
tooth paste	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
water bottle	**97.32**	95.98	95.71	94.64	96.25	**96.78**	96.25	96.25
average	92.62	93.35	93.87	91.85	93.93	92.40	**94.35**	**94.62**

rized in **Table 4**. As is seen in Table 4, the overall recognition accuracy improves significantly across all methods, as compared to the EMK-based ones in Table 3. In a large portion of the categories, perfect recognition is achieved even with the naive baseline SVM algorithm. However, the advantage of the proposed method is revealed in some of the challenging categories.

Overall, with the HMP-based features, "KCCA+L" and the "RGCCA+L" algorithms cannot match the SVM baseline, and the "SVM-2K*" and "KCCA" algorithms are only marginally better than the baseline. The "RGCCA" and "RGCCA+AL" algorithms offer some improvements, while the proposed DCCA algorithm achieves the highest overall recognition accuracy. Considering there are more than 13,000 testing samples, the two percent performance improvement means correctly classifying an additional amount of more than 200 samples.

6.4 Multi-Spectral Scene Dataset

Following Ref. [4], we construct 10 random training and testing splits to evaluate our methods. For each random split, 99 RGB images (11 per category) are used as the testing set while the remaining 378 pairs of RGB-IR images as the training set. Before the feature extraction, each RGB image is converted to the **LAB** color space (similarly to Ref. [4]). Then the GIST [24] features are computed independently on each of the L, A, B and

Table 5 Multi-Spectral Scene recognition, the highest and second highest values are colored red and blue, respectively.

Views	SVM	SVM-2K* [10]	KCCA	KCCA+L
LAB+I	67.78±5.25	67.17±5.58	66.87±4.76	66.46±2.83
L+I	61.82±3.77	61.82±4.59	62.32±4.81	61.92±3.64
Views	RGCCA [34]	RGCCA+L	RGCCA+AL	DCCA
LAB+I	68.59±3.94	67.27±4.88	**69.90**±3.16	**70.51**±2.37
L+I	62.22±3.96	62.42±4.45	**64.55**±3.44	**65.66**±4.57

IR channels. To demonstrate the ubiquitous advantages of our proposed methods, we have tested both the following view assignment schemes concerning these four sensing channels,

- **L+I**: Grayscale GIST features and the IR channel GIST features are assigned as the main view and the auxiliary view, respectively.
- **LAB+I**: GIST features extracted independently from the L, A and B channels are concatenated as the main view, while those extracted from the IR channel are considered as the auxiliary view.

In **Table 5**, the mean and the standard deviation (both in percentage) of the recognition accuracies are reported. We observe that neither the KCCA nor the SVM-2K* algorithm achieves a significant advantage over the SVM baseline, in both the "LAB+I" and the "L+I" view assignments.

With the label view alone, the "KCCA+L" algorithm and the "RGCCA+L" algorithm achieve recognition accuracies on a par with that of the baseline SVM. We speculate that the auxiliary view is more informative: with the auxiliary view, the RGCCA algorithm is capable of outperforming the baseline by a small margin. Furthermore, with the additional label view information in "RGCCA+AL", this margin is enlarged. Overall, the proposed DCCA still outperforms all other competing algorithms in both the "LAB+I" and the "L+I" scenarios.

The large standard deviations in Table 5 could stem from the nature of this dataset: the outdoor scenes lack homogeneity. Indeed, in Ref. [4], Brown and Susstrunk also report large standard deviations on a par with ours.

6.5 Binghamton University 3D Facial Expression Dataset

There are both the 3D models and conventional photos of the 100 subjects (with both genders and various 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) in this dataset [39]. With this dataset, three types of recognition tasks with respect to three different labels (expression, gender, and race) are independently carried out.

First the grayscale face images and the depth face maps are extracted and fed to the Eigen-PEP model [6], [20], [21] to generate features on both views independently. Unlike common 3D face features such as Refs. [23], [37] that require manual labelling of facial landmarks, our feature extraction is fully automatic without the need of manual intervention, making it more suitable for large scale real world applications.

We assign the features extracted from the grayscale images as the main view and the features extracted from the depth maps as the auxiliary view, respectively. For statistical stability, 50 random training and 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.

Table 6 Expression, gender and ethnic/racial ancestries recognition based on the Binghamton 3D Facial Expression dataset, the highest and second highest values are colored red and blue, respectively.

Algorithms	SVM	KCCA	DCCA
expression	72.2±4.1	**73.1**±3.7	**74.5**±4.1
gender	**92.1**±3.2	89.4±4.1	**92.6**±5.3
ethnic/racial	72.0±3.9	**74.1**±4.2	**75.2**±5.2

The recognition accuracies (in percentage) with respect to expression, gender and ethnic/racial ancestries recognition tasks are summarized in **Table 6**. In both the expression and the ethnic/racial ancestry recognition tasks, the proposed DCCA algorithm outperforms the competing ones by obvious margins. The gender recognition task is comparably easier in the sense that there are only two categories (male or female) and this could also contribute to larger intra-class variations which limit the effectiveness of the discriminative learning algorithms. In summary, the proposed DCCA algorithm is capable of achieving higher accuracies in various recognition tasks than the competing ones in Table 6.

6.6 Discussion

Overall, based on the aforementioned empirical results, we have the following observations.

- The latent space based model is capable of leveraging the information from the auxiliary view in training, therefore, the missing-view-in-test-data problem can be effectively addressed.
- Without the auxiliary view, the encoded label view alone is not significant enough to evidently boost the recognition performance.
- Incorporating the encoded label view with the auxiliary view yields some additional boost in recognition performance.
- DCCA consists of three components: the incorporation of the auxiliary view, the encoded label view, and the simultaneous optimization. They jointly contribute to the performance gains.

7. Conclusions

In this paper, we explored a practical multi-view visual recognition problem, where we have multi-view data in the training phase but only the single view data in the test phase. We have verified that information from the auxiliary view in the training data can indeed lead to better recognition in the test phase even when the auxiliary view is entirely missing. As a part of our verification-by-construction proof, we have proposed a robust collaborative multi-view learning framework with missing data, in which a new discriminative canonical correlation analysis method is developed to integrate the semantic information from all views to a common latent space where all subsequent recognition is conducted. We have also investigated and isolated the effects of the encoded label view and the auxiliary view. The exper-

imental results demonstrate that the proposed approach achieves performance advantages on all four benchmarks.

References

[1] Argyriou, A., Evgeniou, T. and Pontil, M.: Convex multi-task feature learning, *Machine Learning*, Vol.73, No.3, pp.243–272 (2008).

[2] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, *Proc. eleventh annual conference on Computational learning theory*, pp.92–100, ACM (1998).

[3] Bo, L., Ren, X. and Fox, D.: Unsupervised feature learning for RGB-D based object recognition, *Experimental Robotics*, pp.387–402, Springer (2013).

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

[5] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology* (*TIST*), Vol.2, No.3, p.27 (2011).

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

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

[8] Chen, L., Li, W. and Xu, D.: Recognizing RGB Images by Learning from RGB-D Data, *2014 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), IEEE (2014).

[9] Davis, J.V., Kulis, B., Jain, P., Sra, S. and Dhillon, I.S.: Information-theoretic metric learning, *Proc. 24th international conference on Machine learning*, pp.209–216, ACM (2007).

[10] Farquhar, J., Hardoon, D., Meng, H., Shawe-taylor, J.S. and Szedmak, S.: Two view learning: SVM-2K, theory and practice, *Advances in neural information processing systems*, pp.355–362 (2005).

[11] Globerson, A. and Roweis, S.: Nightmare at test time: Robust learning by feature deletion, *Proc. 23rd international conference on Machine learning*, pp.353–360, ACM (2006).

[12] Hanafi, M.: PLS path modelling: Computation of latent variables with the estimation mode B, *Computational Statistics*, Vol.22, No.2, pp.275–292 (2007).

[13] Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, Vol.16, No.12, pp.2639–2664 (2004).

[14] Hotelling, H.: Relations between two sets of variates, *Biometrika*, Vol.28, No.3/4, pp.321–377 (1936).

[15] Kim, T.-K., Kittler, J. and Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.29, No.6, pp.1005–1018 (2007).

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

[17] Kulis, B., Sustik, M. and Dhillon, I.: Learning low-rank kernel matrices, *Proc. 23rd international conference on Machine learning*, pp.505–512, ACM (2006).

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

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

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

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

[22] Loog, M., van Ginneken, B. and Duin, R.P.: Dimensionality reduction of image features using the canonical contextual correlation projection, *Pattern Recognition*, Vol.38, No.12, pp.2409–2418 (2005).

[23] Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A. and Berretti, S.: Shape analysis of local facial patches for 3D facial expression recognition, *Pattern Recognition*, Vol.44, No.8, pp.1581–1589 (2011).

[24] Oliva, A. and Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope, *International journal of computer vision*, Vol.42, No.3, pp.145–175 (2001).

[25] Qi, Z., Yang, M., Zhang, Z.M. and Zhang, Z.: Mining noisy tagging from multi-label space, *Proc. 21st ACM international conference on Information and knowledge management*, pp.1925–1929, ACM (2012).

[26] Quanz, B. and Huan, J.: Large margin transductive transfer learning, *Proc. 18th ACM conference on Information and knowledge management*, pp.1327–1336, ACM (2009).

[27] Rupnik, J. and Shawe-Taylor, J.: Multi-view canonical correlation analysis, *Conference on Data Mining and Data Warehouses* (*SiKDD 2010*), pp.1–4 (2010).

[28] Saenko, K., Kulis, B., Fritz, M. and Darrell, T.: Adapting visual category models to new domains, *Computer Vision–ECCV 2010*, pp.213–226, Springer (2010).

[29] Shams, L., Wozny, D.R., Kim, R. and Seitz, A.: Influences of multisensory experience on subsequent unisensory processing, *Frontiers in psychology*, Vol.2:264 (2011).

[30] Shrivastava, A. and Gupta, A.: Building Part-based Object Detectors via 3D Geometry, *2013 IEEE International Conference on Computer Vision* (*ICCV*), pp.1745–1752, IEEE (2013).

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

[32] Srivastava, N. and Salakhutdinov, R.: Multimodal learning with deep boltzmann machines, *Advances in neural information processing systems*, pp.2222–2230 (2012).

[33] Tenenhaus, A.: Kernel Generalized Canonical Correlation Analysis, *JdS'10, May 2010, Marseille, France*, hal-00553602, pp. CD–ROM Proceedings (6 p.) (2010).

[34] Tenenhaus, A. and Tenenhaus, M.: Regularized generalized canonical correlation analysis, *Psychometrika*, Vol.76, No.2, pp.257–284 (2011).

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

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

[37] Wang, J., Yin, L., Wei, X. and Sun, Y.: 3D facial expression recognition based on primitive surface feature distribution, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.1399–1406, IEEE (2006).

[38] Witten, D.M., Tibshirani, R. et al.: Extensions of sparse canonical correlation analysis with applications to genomic data, *Statistical applications in genetics and molecular biology*, Vol.8, No.1, pp.1–27 (2009).

[39] Yin, L., Wei, X., Sun, Y., Wang, J. and Rosato, M.J.: A 3D facial expression database for facial behavior research, *7th international conference on Automatic face and gesture recognition, 2006, FGR 2006*, pp.211–216, IEEE (2006).

[40] Zhang, D., He, J., Liu, Y., Si, L. and Lawrence, R.D.: Multi-view transfer learning with a large margin approach, *Proc. 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.1208–1216 (2011).

Appendix

In this appendix, we present the proof of the Proposition 1 in Section 5. First, the existence of the upper bound is proven in Section A.1, then the proof that the sequence $f(\mathcal{W}(s)), s = 1, 2, \cdots$ is monotonic is presented in Section A.2. With the Bolzano-Weierstrass theorem and the conclusions of Section A.1 and Section A.2, the Proposition 1 is proven.

A.1 Proof: the existence of the bound C_u

From the constraint in Eq. (7) of the paper:

$$\mathbf{W}_j^T \left[(1-\tau_j)\frac{1}{n}\mathbf{R}_j\mathbf{R}_j^T + \tau_j \mathbf{I}_n \right] \mathbf{W}_j = \mathbf{I}_n, \tag{A.1}$$

where τ_j denotes the pre-specified regularization parameter, $0 < \tau_j < 1$ ($j = 1, 2, \cdots, J$), we have

$$(1-\tau_j)\mathrm{tr}\left(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_j^T\mathbf{W}_j\right) + \tau_j \mathrm{tr}\left(\mathbf{W}_j^T\mathbf{W}_j\right) = n, \tag{A.2}$$

and therefore $\forall j = 1, 2, \cdots, J$,

$$\mathrm{tr}\left(\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_j^T\mathbf{W}_j\right) \le \frac{1}{1-\tau_j},\ 0<\tau_j<1. \tag{A.3}$$

In addition, we have

$$\begin{aligned}\mathrm{tr}\left(\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k\right) &\le \frac{1}{2}\mathrm{tr}\left(\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_j^T\mathbf{W}_j\right) &\text{(A.4)}\\ &+ \frac{1}{2}\mathrm{tr}\left(\mathbf{W}_k^T\mathbf{R}_k\mathbf{R}_k^T\mathbf{W}_k\right) &\text{(A.5)}\\ &\le \frac{1}{2}\left[\frac{1}{1-\tau_j}+\frac{1}{1-\tau_k}\right]<\infty, &\text{(A.6)}\end{aligned}$$

where the inequality Eq. (A.5) follows the property

$$\mathrm{tr}(\mathbf{B}^T\mathbf{A}) \le \frac{1}{2}\left[\mathrm{tr}(\mathbf{A}^T\mathbf{A})+\mathrm{tr}(\mathbf{B}^T\mathbf{B})\right], \tag{A.7}$$

which comes from the fact that $\mathrm{tr}\left((\mathbf{A}-\mathbf{B})^T(\mathbf{A}-\mathbf{B})\right) \ge 0$, because the matrix $(\mathbf{A}-\mathbf{B})^T(\mathbf{A}-\mathbf{B}) \ge 0$. Since $\mathrm{tr}\left(\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k\right)$ is bounded, $g(x)=x$ or x^2, we have

$$g\left(\mathrm{tr}\left(\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k\right)\right) \le \frac{1}{4}\left[\frac{1}{1-\tau_j}+\frac{1}{1-\tau_k}\right]^2<\infty. \tag{A.8}$$

Considering $c_{jk}=0$ or 1, we have

$$\begin{aligned}&\sum_{j\neq k}^{J} c_{jk}\ g\left(\mathrm{tr}\left(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k\right)\right) &\text{(A.9)}\\ &\le \frac{1}{4n}\sum_{j\neq k}^{J}\left[\frac{1}{1-\tau_j}+\frac{1}{1-\tau_k}\right]^2 &\text{(A.10)}\\ &< \infty, &\text{(A.11)}\end{aligned}$$

which shows that the sequence $f(\mathcal{W}(s)), s=1,2,\cdots$ is upper bounded by

$$C_u = \frac{1}{4n}\sum_{j\neq k}^{J}\left[\frac{1}{1-\tau_j}+\frac{1}{1-\tau_k}\right]^2<\infty. \tag{A.12}$$

A.2 Proof that the sequence $f(\mathcal{W}(s))$ is monotonically increasing

In this section, the monotonic property of the sequence $f(\mathcal{W}(s)), s=1,2,\cdots$ is presented. Following [12], [33], [34], we first present a Lemma, and then prove that

$$f(\mathcal{W}(s)) \le f(\mathcal{W}(s+1)),\ s=1,2,\cdots, \tag{A.13}$$

where s is the iteration index, $s=1,2,\cdots$.

Define the function $r(\mathbf{Y}_j,\mathbf{Y}_k)$ as $r(\mathbf{Y}_j,\mathbf{Y}_k) \stackrel{\text{def}}{=} \mathrm{tr}(\frac{1}{n}\mathbf{Y}_j^T\mathbf{Y}_k) = \mathrm{tr}(\frac{1}{n}\mathbf{W}_j^T\mathbf{R}_j\mathbf{R}_k^T\mathbf{W}_k)$, therefore,

$$f(\mathbf{W}_1(s),\cdots,\mathbf{W}_J(s)) = \sum_{j,k=1,k\neq j}^{J} c_{jk}g[r(\mathbf{Y}_j(s),\mathbf{Y}_k(s))]. \tag{A.14}$$

Using this notation, we have the following Lemma:

Lemma 1

Define

$$\begin{aligned}f_j(\mathbf{W}_j) &\stackrel{\text{def}}{=} \sum_{k=1}^{j-1} c_{jk}g\left[r(\mathbf{R}_j^T\mathbf{W}_j,\mathbf{Y}_k(s+1))\right] &\text{(A.15)}\\ &+ \sum_{k=j+1}^{J} c_{jk}g\left[r(\mathbf{R}_j^T\mathbf{W}_j,\mathbf{Y}_k(s))\right] &\text{(A.16)}\\ &\text{s.t.}\quad \mathbf{W}_j^TN_j\mathbf{W}_j=\mathbf{I}, &\text{(A.17)}\end{aligned}$$

then

$$f_j(\mathbf{W}_j(s)) \le f_j(\mathbf{W}_j(s+1)),\ j=1,\cdots,J. \tag{A.18}$$

Proof.

We prove Lemma 1 in two cases, i.e., $g(x)=x^2$ and $g(x)=x$.

Case 1: when $g(x)=x^2$, we have $f_j(\mathbf{W}_j)$ in the following form,

$$\begin{aligned}f_j(\mathbf{W}_j) &= \sum_{k=1}^{j-1} c_{jk}\left(r(\mathbf{R}_j^T\mathbf{W}_j,\mathbf{Y}_k(s+1))\right)^2 &\text{(A.19)}\\ &+ \sum_{k=j+1}^{J} c_{jk}\left(r(\mathbf{R}_j^T\mathbf{W}_j,\mathbf{Y}_k(s))\right)^2, &\text{(A.20)}\end{aligned}$$

which can be written as

$$\begin{aligned}f_j(\mathbf{W}_j(s)) &= \frac{1}{n}\sum_{k=1}^{j-1} c_{jk}\theta_{jk}^{(s)}\mathrm{tr}(\mathbf{W}_j(s)^T\mathbf{R}_j\mathbf{Y}_k(s+1)) &\text{(A.21)}\\ &+ \frac{1}{n}\sum_{k=j+1}^{J} c_{jk}\theta_{jk}^{(s)}\mathrm{tr}(\mathbf{W}_j(s)^T\mathbf{R}_j\mathbf{Y}_k(s)) &\text{(A.22)}\\ &= \frac{1}{n}\mathrm{tr}\Big(\mathbf{W}_j(s)^T\mathbf{R}_j\Big(\sum_{k=1}^{j-1} c_{jk}\theta_{jk}^{(s)}\mathbf{Y}_k(s+1) &\text{(A.23)}\\ &+ \sum_{k=j+1}^{J} c_{jk}\theta_{jk}^{(s)}\mathbf{Y}_k(s)\Big)\Big), &\text{(A.24)}\end{aligned}$$

where $\theta_{jk}^{(s)}$ are defined as

$$\begin{aligned}\theta_{jk}^{(s)} &= r\left(\mathbf{Y}_j(s),\mathbf{Y}_k(s+1)\right)\text{if}\ \ k=1,\cdots,j-1 &\text{(A.25)}\\ \theta_{jk}^{(s)} &= r\left(\mathbf{Y}_j(s),\mathbf{Y}_k(s)\right)\text{if}\ \ k=j+1,\cdots,J &\text{(A.26)}\end{aligned}$$

Note that in Eq. (A.24), the inner term $\left(\sum_{k=1}^{j-1} c_{jk}\theta_{jk}^{(s)}\mathbf{Y}_k(s+1) + \sum_{k=j+1}^{J} c_{jk}\theta_{jk}^{(s)}\mathbf{Y}_k(s)\right)$ is equivalent to the definition of $\mathbf{Z}_j(s)$, hence $f_j(\mathbf{W}_j)$ can be simplified as $\frac{1}{n}\mathrm{tr}(\mathbf{W}_j^T\mathbf{R}_j\mathbf{Z}_j(s))$. Considering the following optimization problem:

$$\max_{\mathbf{W}_j}\frac{1}{n}\mathrm{tr}(\mathbf{W}_j^T\mathbf{R}_j\mathbf{Z}_j(s)),\ \text{s.t.}\ \mathbf{W}_j^TN_j\mathbf{W}_j=\mathbf{I}, \tag{A.27}$$

whose solution is exactly

$$\mathbf{W}_j(s+1)=N_j^{-1}\mathbf{R}_j\mathbf{Z}_j(s)\left([\mathbf{Z}_j(s)^T\mathbf{R}_j^TN_j^{-1}\mathbf{R}_j\mathbf{Z}_j(s)]^{1/2}\right)^\dagger \tag{A.28}$$

we have

$$\mathrm{tr}(\mathbf{W}_j(s)^T\mathbf{R}_j\mathbf{Z}_j(s)) \le \mathrm{tr}(\mathbf{W}_j(s+1)^T\mathbf{R}_j\mathbf{Z}_j(s)). \tag{A.29}$$

Similarly, the following equations can be obtained:

$$\begin{aligned}f_j(\mathbf{W}_j(s)) &= \sum_{k=1}^{j-1} c_{jk}\theta_{jk}^{(s)}r(\mathbf{Y}_j(s),\mathbf{Y}_k(s+1)) &\text{(A.30)}\\ &+ \sum_{k=j+1}^{J} c_{jk}\theta_{jk}^{(s)}r(\mathbf{Y}_j(s),\mathbf{Y}_k(s)) &\text{(A.31)}\\ &\le \sum_{k=1}^{j-1} c_{jk}\theta_{jk}^{(s)}r(\mathbf{Y}_j(s+1),\mathbf{Y}_k(s+1)) &\text{(A.32)}\\ &+ \sum_{k=j+1}^{J} c_{jk}\theta_{jk}^{(s)}r(\mathbf{Y}_j(s+1),\mathbf{Y}_k(s)). &\text{(A.33)}\end{aligned}$$

Considering that c_{jk} is either 0 or 1, we hence have $c_{jk}=c_{jk}^2$. Applying the Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
& f_j\ (\mathbf{W}_j(s)) \\
& \le \sum_{k=1}^{j-1} c_{jk}^2 \theta_{jk}^{(s)} r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s+1)) \\
& + \sum_{k=j+1}^{J} c_{jk}^2 \theta_{jk}^{(s)} r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s)) \\
& \le \Big[\sum_{k=1}^{j-1} c_{jk} \left(\theta_{jk}^{(s)}\right)^2 + \sum_{k=j+1}^{J} c_{jk} \left(\theta_{jk}^{(s)}\right)^2 \Big]^{1/2} \\
& \cdot \Big[\sum_{k=1}^{j-1} c_{jk} \left(r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s+1))\right)^2 \\
& + \sum_{k=j+1}^{J} c_{jk} \left(r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s))\right)^2 \Big]^{1/2} \\
& = \Big[\sum_{k=1}^{j-1} c_{jk} \left(r(\mathbf{Y}_j(s), \mathbf{Y}_k(s+1))\right)^2 \\
& + \sum_{k=j+1}^{J} c_{jk} \left(r(\mathbf{Y}_j(s), \mathbf{Y}_k(s))\right)^2 \Big]^{1/2} \\
& \cdot \Big[\sum_{k=1}^{j-1} c_{jk} \left(r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s+1))\right)^2 \\
& + \sum_{k=j+1}^{J} c_{jk} \left(r(\mathbf{Y}_j(s+1), \mathbf{Y}_k(s))\right)^2 \Big]^{1/2} \\
& = \Big[f_j(\mathbf{W}_j(s)) \Big]^{1/2} \cdot \Big[f_j(\mathbf{W}_j(s+1)) \Big]^{1/2}.
\end{aligned}
$$

We immediately have $f_j(\mathbf{W}_j(s)) \le f_j(\mathbf{W}_j(s+1))$. This concludes the case 1 scenario.

Case 2: when $g(x) = x$, we have

$$
\begin{aligned}
f_j(\mathbf{W}_j) &= \sum_{k=1}^{j-1} c_{jk} r(\mathbf{R}_j^T \mathbf{W}_j, \mathbf{Y}_k(s+1)) && \text{(A.34)} \\
&+ \sum_{k=j+1}^{J} c_{jk} r(\mathbf{R}_j^T \mathbf{W}_j, \mathbf{Y}_k(s)). && \text{(A.35)}
\end{aligned}
$$

Therefore, we can have exactly the same equation as Eq. (A.24), except that $\theta_{jk}^{(s)} \equiv 1$ for all the cases. The same equation as in Eq. (A.29) can be obtained, which directly implies that $f_j(\mathbf{W}_j(s)) \le f_j(\mathbf{W}_j(s+1))$. This concludes both the case 2 scenario and the entire proof of Lemma. With the conclusion in Lemma 1, we proceed with the proof that the sequence $f(\mathcal{W}(s)), s = 1, 2, \cdots$ is monotonically increasing. Consider the following subtraction

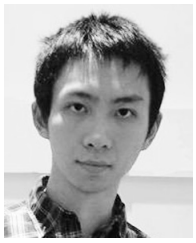
$$
\begin{aligned}
& \sum_{j=1}^{J} [f_j(\mathbf{W}_j(s+1)) - f_j(\mathbf{W}_j(s))] && \text{(A.36)} \\
& = \sum_{j=1}^{J} \sum_{k=1}^{j-1} c_{jk} g\left[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s+1)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s+1) \right) \right] && \text{(A.37)} \\
& + \sum_{j=1}^{J} \sum_{k=j+1}^{J} c_{jk} g\left[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s+1)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s) \right) \right] && \text{(A.38)} \\
& - \sum_{j=1}^{J} \sum_{k=1}^{j-1} c_{jk} g\left[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s+1) \right) \right] && \text{(A.39)} \\
& - \sum_{j=1}^{J} \sum_{k=1}^{j-1} c_{jk} g\left[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s+1)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s+1) \right) \right] && \text{(A.40)} \\
& = \frac{1}{2} \Big[\sum_{j,k=1, k\ne j}^{J} c_{jk} g\Big[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s+1)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s+1) \right) \Big] && \text{(A.41)} \\
& - \sum_{j,k=1, k\ne j}^{J} c_{jk} g\Big[\mathrm{tr}\left(\frac{1}{n} \mathbf{W}_j(s)^T \mathbf{R}_j \mathbf{R}_k^T \mathbf{W}_k(s) \right) \Big] \Big] \ge 0. && \text{(A.42)}
\end{aligned}
$$

The last equation in Eq. (A.42) follows the Lemma 1. This implies that

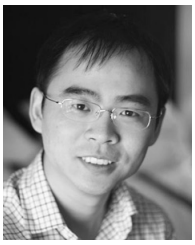
$$
\begin{aligned}
& f(\mathbf{W}_1(s), \cdots, \mathbf{W}_J(s)) \le f(\mathbf{W}_1(s+1), \cdots, \mathbf{W}_J(s+1)) && \text{(A.43)} \\
& \text{i.e., } f(\mathcal{W}(s)) \le f(\mathcal{W}(s+1)),\ s = 1, 2, \cdots. && \text{(A.44)}
\end{aligned}
$$

Using Eq. (A.11), Eq. (A.44), the bounded sequence $f(\mathcal{W}(s)), s = 1, 2, \cdots$ is monotonically increasing.

According to the Bolzano-Weierstrass theorem, the sequence will converge, i.e., Proposition 1 is proven.



Qilin Zhang received his B.Eng degree in electrical information engineering from the University of Science and Technology of China, Hefei, China, in 2009, and the M.S. degree in electrical and computer engineering from University of Florida, Gainesville, Florida, USA in 2011. He is currently a Ph.D. student at the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA.



Gang Hua (M'03-SM'11) was enrolled in the Special Class for the Gifted Young of Xian Jiaotong University (XJTU) in 1994 and received the B.S. degree in automatic control engineering from XJTU in 1999. He received the M.S. degree in Control Science and Engineering in 2002 from XJTU, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science at Northwestern University in 2006. He is a Senior Research Manager at Microsoft Research Asia. He also holds a Visiting Professor position at Stevens Institute of Technology. His research focuses on computer vision, pattern recognition, machine learning, and robotics, with primary applications in the cloud and mobile intelligence domain. Before he come back to Microsoft Research, he was an Associate Professor of Computer Science at Stevens Institute of Technology between 2011 and 2015. During the academic year 2014-2015, he took an on leave from Stevens and worked on a confidential corporate project at Amazon. He was an academic visiting researcher position at IBM Research T. J. Watson Center between 2011 and 2014. Before that, he was a Research Staff Member in IBM Research T. J. Watson Center from 2010 to 2011, a Senior Researcher in Nokia Research Center Hollywood from 2009 to 2010, and a Scientist in Microsoft Live labs Research from 2006 to 2009. He is the recipient of the 2015 IAPR Young Biometrics Investigator Award, and a recipient of the 2013 Google Research Faculty Award. He has been awarded Best Reviewers Award for many top international conferences including CVPR/ICCV/ACCV/BTAS, etc. He served as Area Chairs for CVPR'2014, ICCV'2011, ACM MM'2011&2012&2015, ICIP2012&2013, ICASSP2012&2013. He served as an Associate Editor for IEEE T-IP (2010-2014), and is currently serving as Associate Editors for IEEE Multimedia, CVIU, MVA, and VCJ. He has published more than 100 peer reviewed papers in top conferences such as CVPR/ICCV/ECCV, and top journals such as T-PAMI and IJCV. He holds 15 issued U.S Patents and also has more than 10 U.S. Patents Pending. He is a Senior Member of IEEE and a Life Member of ACM.

Wei Liu received his M.Phil. and Ph.D. degrees in electrical engineering from Columbia University, New York, NY, USA in 2012. Currently, he is a research staff member of IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, and holds adjunct faculty positions at Rensselaer Polytechnic Institute and Stevens Institute of Technology. His research interests include machine learning, data mining, computer vision, pattern recognition, image processing, multimedia, and information retrieval. Dr. Liu is the recipient of the 2011-2012 Facebook Fellowship and the 2013 Jury Award for best thesis of Department of Electrical Engineering, Columbia University. Dr. Liu has published over 70 papers in peer-reviewed journals and conferences including Proceedings of IEEE, NIPS, ICML, KDD, CVPR, ICCV, ECCV, MICCAI, DCC, ACM Multimedia, IJCAI, AAAI, SIGIR, SIGCHI, etc. His recent papers win CVPR Young Researcher Support Award and Best Paper Travel Award for ISBI 2014.

Zicheng Liu is a principal researcher at Microsoft Research, Redmond. His current research interests include human activity understanding, face modeling and animation, and human computer interaction. He received a Ph.D. in Computer Science from Princeton University, a M.S. in Operational Research from the Institute of Applied Mathematics, Chinese Academy of Science, and a B.S. in Mathematics from Huazhong Normal University, China. Before joining Microsoft Research, he worked at Silicon Graphics as a member of technical staff. He co-authored three books: Face Geometry and Appearance Modeling: concept and applications, Cambridge University Press, Human Action Recognition with Depth Cameras, Springer Briefs, and Human Action Analysis with Randomized Trees, Springer Briefs. He was a technical co-chair of 2010 and 2014 IEEE International Conference on Multimedia and Expo, and a general co-chair of 2012 IEEE Visual Communication and Image Processing. He is an associate editor of Machine Vision and Applications journal, and an associate editor of the Journal of Visual Communication and Image Representation. He is a fellow of IEEE.

Zhengyou Zhang received his B.S. degree in electronic engineering from Zhejiang University, Hangzhou, China, in 1985, the M.S. degree in computer science from the University of Nancy, Nancy, France, in 1987, and the Ph.D. degree in computer science and the Doctorate of Science (Habilitation diriger des recherches) from the University of Paris XI, Paris, France, in 1990 and 1994, respectively. He is a Principal Researcher with Microsoft Research, Redmond, WA, USA, and the Research

Manager of the Multimedia, Interaction, and Communication group. Before joining Microsoft Research in March 1998, he was with INRIA (French National Institute for Research in Computer Science and Control), France, for 11 years and was a Senior Research Scientist from 1991. In 1996-1997, he spent a one-year sabbatical as an Invited Researcher with the Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan. He served as an Adjunct Chair Professor with Zhejiang University, Hangzhou, China. He is also an Affiliate Professor with the University of Washington, Seattle, WA, USA. He has published over 200 papers in refereed international journals and conferences, and has coauthored the following books: 3-D Dynamic Scene Analysis: A Stereo Based Approach (Springer-Verlag, 1992); Epipolar Geometry in Stereo, Motion and Object Recognition (Kluwer, 1996); Computer Vision (Chinese Academy of Sciences, 1998, 2003, in Chinese); Face Detection and Adaptation (Morgan and Claypool, 2010), and Face Geometry and Appearance Modeling (Cambridge University Press, 2011). He has given a number of keynotes in international conferences and invited talks in universities. Dr. Zhang is a Fellow of the Institute of Electrical and Electronic Engineers (IEEE), a Fellow of Association for Computing Machinery (ACM), the Founding Editor-in-Chief of the IEEE Transactions on Autonomous Mental Development, an Associate Editor of the International Journal of Computer Vision, an Associate Editor of Machine Vision and Applications, and an Area Editor of the Journal of Computer Science and Technology. He served as Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2000 to 2004, an Associate Editor of the IEEE Transactions on Multimedia from 2004 to 2009, an Associate Editor of the International Journal of Pattern Recognition and Artificial Intelligence from 1997 to 2009, among others. He has been on the program committees for numerous international conferences in the areas of computer vision, audio and speech signal processing, multimedia, human-computer interaction, and autonomous mental development. He was a member of the Pre- and Interim Steering Committee, in 2009, in charge of revamping the International Conference of Multimedia and Expo (ICME), the flagship multimedia conference sponsored by four IEEE societies. He served as Area Chair, Program Chair, or General Chair of a number of international conferences, including recently a Program Co-Chair of the International Conference on Multimedia and Expo (ICME), July 2010, a Program Co-Chair of the ACM International Conference on Multimedia (ACM MM), October 2010, a Program Co-Chair of the ACM International Conference on Multimodal Interfaces (ICMI), November 2010, and a General Co-Chair of the IEEE International Workshop on Multimedia Signal Processing (MMSP), October 2011 and September 2015. He served as a Chair of a new track Technical Briefs of the ACM SIGGRAPH Asia Conference, Nov. 28 C Dec. 1st, 2012. He is serving as a General Chair of International Conference on Multimodal Interaction (ICMI) 2015, and a General Chair of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. He received the IEEE Helmholtz Test of Time Award at ICCV 2013 for his paper published in 1999 on camera calibration, now known as Zhang's method.

(Communicated by *Ming-Hsuan Yang*)