A Dynamic Discriminative Canonical Correlation Analysis via Adaptive Weight Scheme

SHUZHI SU¹,², PENGLIAN GAO¹, YANMIN ZHU³, XINGZHU LIANG¹,⁴, AND JUN XIE¹

¹School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China
²Institute of Energy, Hefei Comprehensive National Science Center, Anhui, Hefei 230031, China
³School of Mechanical Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China
⁴Institute of Environment-Friendly Materials and Occupational Health, Anhui University of Science and Technology (Wuhu), Wuhu 241003, China

Corresponding author: Yanmin Zhu (zyanmin1988@163.com)

This work was supported in part by the Foundation of National Natural Science Foundation of China under Grant 61806006, in part by the Institute of Energy, Hefei Comprehensive National Science Center under Grant 19KZS203, in part by the Foundation of China Postdoctoral Science Foundation under Grant 2019M660149, and in part by the Science and Technology Research Project of Wuhu City under Grant 2020yF48.

ABSTRACT In multi-view learning, massive literature is devoted to exploring the intrinsic structure between cross-views. It is well known that canonical correlation analysis (CCA) is a conventional multi-view learning method, which considers the correlation between two views. However, it fails to utilize class information and is difficult to suit different issues to extract discriminative features. In this paper, we propose a novel cross-view discriminative feature learning method called dynamic discriminative canonical correlation analysis, which captures class information to yield discriminative features. More specifically, we develop an adaptive weight scheme of cross-view within-class and between-class scatters to make full use of distribution class information. In addition, an iterative algorithm with Cauchy inequalities and the Lagrange multiplier is proposed to handle the non-smooth objective function. Our method is applied to face recognition and multi-linguistic text classification tasks. Extensive experimental results reveal that the adaptive weight scheme plays a beneficial role and our method is an effective feature learning.

INDEX TERMS Feature learning, adaptive weight scheme, dynamic discrimination, canonical correlation analysis.

I. INTRODUCTION

One object can possess single or multiple data representations, and data can be classified as single-view data [1] and multi-view data [2] by the number of data representations corresponding to one same object. Single-view data can directly reflect one nature, while multi-view data can explore the same latent contents from different viewpoints. The fact is that most of the data existing in many real-world applications (e.g., hyperspectral anomaly detection [3], face recognition, and multi-linguistic text classification) are multi-view data. For instance, hyperspectral and multi-sensor data contain different discrimination with subtle spectral differences. In addition, we often translate an article into different languages to make readers from different regions understand the contents that authors want to deliver. The text representations from different languages belong to the multi-view data of the article. These multi-view data can be beneficial to improve a variety of practical applications including classification, retrieval, label prediction tasks [4]. Naturally, multi-view feature learning has been a popular and valuable field. Scholars show a strong interest in the study of the multi-view method and propose a lot of meaningful algorithms based on effective approaches and rigorous theory. With the success of convolutional neural networks (CNN [5], [6]), the computer vision community has witnessed a huge shift from conventional methods to deep learning in multi-view feature learning. Zhao et al. proposed a series of effective methods to promote the joint classification using hierarchical random walk network (HRWN) [7], fractional Gabor convolutional network [8], dual-tunnel CNN [9], Patch-to-Patch CNN [10]. On the other hand, in recent years traditional hashing also demonstrates its superiority of short binary representation for high-dimension data. To make full use of correlations of cross-view data, many effective cross-view hashing e.g semi-paired discrete hashing (SPDH) [11] and
dubbed multiview discrete hashing (MvDH) [12] has been proposed.

Nowadays multi-view feature learning is divided into three kinds: co-training, subspace learning, and multi-kernel learning. Canonical correlation analysis (CCA) [13] belonging to subspace learning has been widely regarded as a classic method with the capability of exploiting the linear correlation between two series of features. CCA based on multivariate correlation statistics maximizes the correlation between two canonical variates in common low-dimensional feature space [14]. Canonical projections of CCA are the conjugate orthogonal projections. Specifically, the \( i \)th projection pair is conjugate orthogonal with the \( j \)th (\( i \neq j \)) projection pair to preserve useful content in each feature space. Nevertheless, classic CCA fails to cope with the small sample size (SSS) problem [15]. The SSS problem and noise disturbance usually lead to poor performance in recognition and classification tasks. Orthogonality [16] has been proved as a useful criterion to solve the SSS problem. Shen et al. [17] proposed an extended version of orthogonal CCA (OCCA) that can compute the first canonical projection pair and constrained the other canonical projection direction pairs with orthogonality. OCCA gives effective solutions by using twin eigen decomposition and by replacing the conjugate orthogonality constraint. A large number of experiments have shown its validity. Another effective approach to improve the robustness is probabilistic CCA (PCCA) [18]. The novel strategy and solution of canonical correlation features extraction based on distribution assumption in the union subspace, PCCA, has been proved effective to improve the robustness of the model.

In addition, a number of existing multi-view methods are committed to exploiting the mutual information toward one same object and unique character of different views such as cross-view locality preserved diversity and consensus learning (CvLP-DCL) [19], learning a joint affinity graph for multiview subspace clustering [20], and cost-sensitive canonical correlation analysis (CSCCA) [21]. CCA and OCCA are unsupervised feature learning methods [22], which fail to utilize label information to enhance their discrimination power for better recognition performance. In theory, supervisory information [23] could show real distribution, and more real structure information can be explored to obtain a more discriminative training model for feature learning. Abhishek Sharma et al. [24] presented a general multi-view feature learning approach namely generalized multi-view analysis (GMA) for cross-view classification and retrieval, which turns out to be supervised discriminative and effective. GMA pointed out that some existing feature learning methods could be regarded as certain special forms of quadratic constrained quadratic programming (QCQP). GMA first simply adds a balance parameter to extend a single-view model to a multi-view model, and then certain constraints are added for the convenience of computation in the latent low-dimensional feature space. Single-view feature learning methods such as principal component analysis (PCA) [25], linear discriminant analysis (LDA) [26], neighborhood preserving embedding (NPE) [27], and locality preserving projection (LPP) [28] can be extended to multi-view versions that can preserve their inherent characteristics still through the GMA framework. For instance, numerous experiments of generalized multi-view locality preserving projection (GMLPP) [24] have shown its validity, and GMLPP makes the low-dimensional features of multi-view data still maintain the discriminative local structural relationships of the original high-dimensional feature space in the low-dimensional feature space. Although the GMA schema solves the multi-view problem very well, it does not take into account the correlation between views. Different from GMA, discriminative canonical correlation analysis (DCCA [29]) makes full use of the advantage of supervision and takes the global correlation into account. The canonical projection directions in DCCA have the property of maximizing within-class similarity [30] and minimizing between-class similarity to eliminate the redundant information in the same view while canonical projection directions obtain the maximum between-view correlation.

However, the GMA framework and DCCA suffer from a lack of consideration towards label scatter [30]. Under some specific application backgrounds, it is significant to highlight the class information that mainly contributes to better discrimination. For instance, to judge whether an article is suitable for publication in scientific and technological journals, the training model based on published articles in the journal should be more sensitive to the semantics of science and technology in text classification [31]. For a vast of labeled data, it is a cumbersome task that we note the practical category weights modifying model performance instead of simply calculate the probabilities of each category. Besides, it is nearly impossible to obtain the consistent practical category weight according to two different views. Since the problem above is a serious challenge, in this paper we propose a novel supervised feature learning method called dynamic discriminative canonical correlation analysis (DDCCA). DDCCA sets the adaptive weight [32] for each category and automatically learns the contribution of categories and finds out the primary classes to be identified in classification to improve classification performance. In addition, for simplifying the solving process DDCCA decomposes a multivariable problem into two univariate problems with mutual dynamic constraints [33]. The decomposition based on Cauchy-Schwarz inequality [34] and maximum correlation between cross-view makes DDCCA fully learn the dispersion of supervisory information to realize class sensitivity and complete the fast dynamic global optimal solution. Numerous experiments of image recognition and text classification can reveal the superiority of our method. We summarize the main contribution of this work as follows:

1) For better exploiting the inner structure of classes and practical contribution of categories hidden in the vast labeled data, a dynamic discriminative conical correlation analysis via adaptive scheme is designed to
learn the consistent contribution of different classes and preserve the class scatter in the union low-dimensional subspace from two views, which is helpful to improve the discrimination power of features.

2) In order to adjust the adaptive weights to obtain the discriminative feature quickly, we propose an effective iterative solution with alternating optimization by Cauchy-Schwarz inequality and Lagrange multiplier, which is of low computation complexity.

3) We carefully conduct extensive experiments on several faces and multi-language text datasets to demonstrate the superiority of our method to existing related methods.

The rest of this paper is organized as follows: In Section 2, we briefly review conventional CCA. In Section 3, a dynamic discriminative canonical correlation analysis via adaptive weight scheme is formulated and elaborated in detail. In Section 4, we perform several experiments on the public datasets with some other methods to demonstrate the effectiveness of our method. Finally, in Section 5, we conclude the paper.

II. REVIEW OF CCA

In univariate statistical analysis, the correlation coefficient [35] is used to measure the linear correlation of two random variables, and the complex correlation coefficient is used to study the linear correlation between one random variable and multiple random variables. On the other hand, CCA is a multivariate statistical analysis method that uses the correlation between canonical variable pairs to reflect the overall correlation between the two groups of indicators. More specifically, given N pairs of samples \( \{(x_i, y_i)\}_{i=1}^{N} \) coming from c classes from two views, CCA aims to find pairs of projection directions \( \alpha \) and \( \beta \) that maximize the correlation between the joint projections \( \alpha^T x \) and \( \beta^T y \). Denote \( \rho \) as canonical correlation:

\[
\rho = \max_{\alpha, \beta} \frac{\alpha^T C_{xy} \beta}{\sqrt{\alpha^T C_{xx} \alpha} \sqrt{\beta^T C_{yy} \beta}}
\]

where \( C_{xx} \) and \( C_{yy} \) are variance matrices of \( X \) and \( Y \), and \( C_{xy} \) represents the covariance matrix between \( X \) and \( Y \) sets respectively. From Eq. (1), we can find that the value of \( \rho \) has nothing to do with the length of \( \alpha \) and \( \beta \), and the value only matters with the angular relationship between \( \alpha \) and \( \beta \) [36]. It is easy to constrain the denominator \( \alpha^T C_{xx} \alpha = 1 \) and \( \beta^T C_{yy} \beta = 1 \) like the operation of LDA. The optimization problem of CCA can be solved by the Lagrange multiplier method.

III. A DYNAMIC DISCRIMINATIVE CANONICAL CORRELATION ANALYSIS VIA ADAPTIVE WEIGHT SCHEME

Although CCA can reduce the dimension of raw data to extract features, CCA fails to utilize supervised information such as label information, which has been proved a superior strategy. To overcome such a shortage of CCA, Bai et al. [37] proposed a supervised combined feature learning method known as DCCA. The features extracted by DCCA can be proved to be more discriminative, which benefits image recognition. However, DCCA fails to exploit the distribution of classes [38] and react to the difference of classes.

In this section, we explore the discriminative structures [39] of classes to access sensitivity towards the contribution of classes by means of natural logic hidden in data rather than artificial classification logic. Then the novel cross-view discriminative feature learning method i.e. DDCCA is proposed. DDCCA utilizes the mathematical operation of Cauchy–Schwarz inequality [34] to reach effective self-learning, and the theoretical analysis will be given to prove its validity.

A. PROBLEM FORMULATION

Let \( X = [x_1^{(1)}, \ldots, x_n^{(1)}, \ldots, x_1^{(c)}, \ldots, x_n^{(c)}] \), \( Y = [y_1^{(1)}, \ldots, y_n^{(1)}, \ldots, y_1^{(c)}, \ldots, y_n^{(c)}] \), are two sets across different views where \( N \) pairs of zero-means samples from \( c \) classes and \( n_i \) is the number of samples from \( i \)th class. The greater the within-class correlation of canonical variates is, the more profound and obvious the characteristics of classes are. The smaller the between-class correlation of canonical variates is, the greater difference and the higher degree of discrimination between different classes are, which is conducive to classification and recognition. Aim at extracting discriminant features, we can construct an objective function by maximizing cross-view within-class scatters and minimizing cross-view-between-class scatters:

\[
\max_{\alpha, \beta} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_j} \alpha^T x_j^{(i)} y_k^{(i)} T \beta - \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_j} \alpha^T x_j^{(i)} y_k^{(i)} T \beta
\]

We constrain the within-class structures and between-class relationships of canonical variables according to Eq. (2). The class weights of the formula default to 1 for each class, which means the equal effectiveness of each class in discriminative feature learning. However, different classes have different sensitivity for classification or recognition tasks. For better exploring the inner structures in each class and the relationships between different classes, we seek a feasible linear joint combination of the column vectors of \( X \) and \( Y \) for representing cross-view within-class and between-class scatters. In addition, we set a tunable parameter that reveals the relative sensitivity between the within-class scatters and the between-class scatters. Then we modify the correlation criterion of pairs of canonical variates:

\[
\max_{\alpha, \beta} \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_j} \alpha^T \tilde{\delta}_i x_j^{(i)} y_k^{(i)} T \beta - \sum_{i=1}^{c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_j} \alpha^T \tilde{\eta}_j x_j^{(i)} y_k^{(i)} T \beta
\]
where $\delta_i$ and $n_i$ denotes the weights affected by the distribution of classes. Although more details on different classes can be explored, it will result in higher overall complexity. It can be easily solved by a mathematical connection between cross-view within-class scatters and cross-view between-class scatters since the samples have been mean-normalized. Thus, the between-class scatter matrix can be rewritten as follows:

\[
C_b = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} x_k^{(i)} y_l^{(j)^T}
\]

\[
= \sum_{i=1}^{c} \sum_{k=1}^{n_i} \sum_{j=1}^{n_j} x_k^{(i)} y_l^{(j)^T} - \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} x_k^{(i)} y_l^{(j)^T}
\]

\[
= -\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} x_k^{(i)} y_l^{(j)^T}
\]

\[
= -C_w
\]

Obviously, the between-class scatter matrix can represent the within-class scatter matrix, which indicates the connection between local and global distribution of classes. Furthermore, in order to adjust the contribution between the classes with the smaller scatter matrix and the classes with a larger scatter matrix, an adaptive weight is assigned for each class $\omega = \{w_i, i = 1, \ldots, c\}$ globally instead of simply adding regularization for specific classes manually. Thus, $C_w$ is denoted as

\[
C_w = \sum_{i=1}^{c} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \frac{x_k^{(i)} y_l^{(j)^T}}{w_i}
\]

\[
= \sum_{i=1}^{c} \left( \frac{Xe_{n_i}^{(i)}}{n_i} \right) \left( Ye_{n_j}^{(j)^T} \right)^T
\]

\[
= XWY^T
\]

represents an adaptive weight matrix that is symmetric, positive semi-definite, and blocked diagonal \cite{40}. rank( $W$ ) = $c$ is always less than the number of samples. And then, we minimize the weight summation function and simplify the equation by only analyzing the cross-view within-class scatters as follows to obtain the higher correlation:

\[
f (w) = \min_w \frac{\alpha^T XWY^T \beta}{\sqrt{\alpha^T XX^T \alpha} \sqrt{\beta^T YY^T \beta}}
\]

\[
s.t. \omega^T 1_c = 1, \quad w_i \geq 0
\]

We notice that the correlation increases with the decrease of weight $1/w_i$, which can intentionally reduce the deviation of the classes with small correlations. The projection directions can be obtained by maximizing $f (w)$ and the objective function can be noted as follows:

\[
f (\alpha, \beta, w) = \max_{\alpha, \beta} \min_w \frac{\alpha^T XWY^T \beta}{\sqrt{\alpha^T XX^T \alpha} \sqrt{\beta^T YY^T \beta}}
\]

\[
s.t. \omega^T 1_c = 1, \quad w_i \geq 0
\]

$f (\alpha, \beta, w)$ has nothing to do with the length of $\alpha$ and $\beta$ \cite{36}. That is, $f (\alpha, \beta, w)$ possesses the property of the scale invariance. Thus, the criterion function is rewritten as

\[
\max_{\alpha, \beta} \min_w \frac{\alpha^T XWY^T \beta}{\sqrt{\alpha^T XX^T \alpha} \sqrt{\beta^T YY^T \beta}}
\]

\[
s.t. \omega^T 1_c = 1, \quad w_i \geq 0
\]

Specifically, Eq. (2) can be approximately transformed into a supervise canonical correlation analysis when each element of $w$ equal to 1. It can be easily found that its optimal solutions $\alpha$ and $\beta$ not only minimize the intra-class aggregation but also maximize the inter-class separability. Since the optimal projection directions are based on the dynamic weight $w$, the objective function is difficult to be directly solved. Thus, an iterative approach is constructed for dealing with the problem in the next subsection.

**B. OPTIMIZATION**

In this subsection, a capable and fast iterative approach is proposed to reach the optimal stable solution of Eq. (8).

Firstly, adaptive weights $\omega = \{w_i : w_i = 1/c, i = 1, \ldots, c\}$ are initialized and then update $\omega$ and $(\alpha, \beta)$ iteratively by updating one item while fixing the other one.

When $\omega$ is fixed, the objective function of Eq. (8) can be simplified to

\[
\max_{\alpha, \beta} \alpha^T XWY^T \beta
\]

\[
s.t. \alpha^T XX^T \alpha = 1, \quad \beta^T YY^T \beta = 1
\]

It is not difficult to utilize the Lagrange multiplier method to solve it. And we can get the corresponding Lagrange
function:
\[
L(\alpha, \beta) = \alpha^T X W Y T \beta - \frac{\lambda_1}{2} \left( \alpha^T X X^T \alpha - 1 \right) - \frac{\lambda_2}{2} \left( \beta^T Y Y^T \beta - 1 \right) - \frac{\lambda_3}{2} (\omega^T 1 - 1) \tag{10}
\]

The above equation can further compute the partial derivations of \(\alpha\) and \(\beta\), and the partial derivations are set to zero:
\[
\frac{\partial L}{\partial \alpha} = X W Y T \beta - \lambda_1 X X^T \alpha = 0 \tag{11}
\]
\[
\frac{\partial L}{\partial \beta} = Y W^T X T \alpha - \lambda_2 Y Y^T \beta = 0 \tag{12}
\]

Both sides of Eq. (11) and Eq. (12) are multiplied by and respectively, and combined with the constraint of Eq. (9), we can get
\[
\alpha^T X W Y T \beta = \lambda_1 \alpha^T X X^T \alpha = \lambda_1 \tag{13}
\]
\[
\beta^T Y W^T X T \alpha = \lambda_2 \beta^T Y Y^T \beta = \lambda_2 \tag{14}
\]

Since adaptive weight matrix is symmetric block diagonal matrix, that is \(\lambda_1 = \lambda_2^T = (\alpha^T X W Y T \beta)^T = \beta^T Y W^T X T \alpha = \lambda_2 = \lambda_2\). Now the Eq. (11) and Eq. (12) can be converted to the two general decomposition problems:
\[
X W Y^T (Y Y^T)^{-1} Y W^T X T \alpha = \lambda_2 X X^T \alpha \tag{15}
\]
\[
Y W^T X^T (X X^T)^{-1} X W Y T \beta = \lambda_2 Y Y^T \beta \tag{16}
\]

Obviously, we can do an eigen decomposition of Eq. (15) and Eq. (16) to get projector matrices \(A\) and \(B\) consisting of optimal project direction \((\alpha_i, \beta_i)_{d=1}^d\).

When \(A\) and \(B\) is fixed, Eq. (8) can reformulate the following objective function with the characteristic of trace
\[
\min_{w} \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \frac{\text{tr}(A^T x_i^T y_l^T B)}{w_i} \tag{17}
\]

s.t. \(\sum_{i=1}^c w_i = 1, \quad w_i \geq 0\)

Due to \(\sum_{i=1}^c w_i = 1\), problem (13) can be equivalently translated as follow for further analysis:
\[
\min_{w} \sum_{i=1}^c \sum_{k=1}^{n_i} \frac{\text{tr}(A^T x_k^T y_l^T B)}{w_i} \left( \sum_{i=1}^c w_i \right) \tag{18}
\]

s.t. \(\sum_{i=1}^c w_i = 1, \quad w_i \geq 0\)

According to the property of matrix trace, we can know that \(tr(mA + nB) = mtr(A) + ntr(B)\). The property makes the solution of Eq. (13) feasible by utilizing the property of Cauchy-Schwarz inequality shown in Theorem 1.

**Theorem 1**: In Euclidean space \(\mathbb{R}^n\) with the standard inner product, Cauchy-Schwarz inequality is
\[
\left( \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \frac{\text{tr}(A^T x_k^T y_l^T B)}{w_i} \right)^2 \geq \left( \sum_{i=1}^c \left( \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \sqrt{\text{tr}(A^T x_k^T y_l^T B)} \right) \right)^2.
\]

And
\[
\left( \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \frac{\text{tr}(A^T x_k^T y_l^T B)}{w_i} \right)^2 = \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \sqrt{\text{tr}(A^T x_k^T y_l^T B)}.
\]

holds if and only if
\[
\left( \sum_{i=1}^c \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \frac{\text{tr}(A^T x_k^T y_l^T B)}{w_i} \right)^2 \geq \left( \sum_{i=1}^c \left( \sum_{k=1}^{n_i} \sum_{l=1}^{m_i} \sqrt{\text{tr}(A^T x_k^T y_l^T B)} \right) \right)^2.
\]

**Proof of Theorem 1**: Let \(\eta_i = \frac{\sqrt{\text{tr}(A^T x_k^T y_l^T B)}}{w_i} - 1\), \(\theta_l = \sqrt{w_i}\), \(i = 1, \ldots, c\) and \(\theta_l = \sqrt{w_i}, i = 1, \ldots, c\) i.e. \(\eta = (\eta_1, \eta_2, \ldots, \eta_c), \quad \theta = (\theta_1, \theta_2, \ldots, \theta_c)\). It is well-known to us that \(\eta^T \theta = \eta_1 \theta_1 + \eta_2 \theta_2 + \cdots + \eta_c \theta_c\). Obviously
\[
\eta^T \theta \leq |\eta| \cdot |\theta| \tag{20}
\]

Square the left side of inequality (20) \((\eta^T \theta)^2 = (\eta_1 \theta_1 + \eta_2 \theta_2 + \cdots + \eta_c \theta_c)^2\) and do the same mathematical operation for the right formula \(|\eta|^2 \cdot |\theta|^2 = (\eta_1^2 + \eta_2^2 + \cdots + \eta_c^2)(\theta_1^2 + \theta_2^2 + \cdots + \theta_c^2)\) in the meantime, inequality still holds. Note the condition of the equal sign: If and only if \(\cos^2 \langle \eta, \theta \rangle = 1\), \(\eta\) and \(\theta\) two vectors are collinear i.e. \(\theta = \frac{\theta_1}{\theta_1} = \frac{\theta_2}{\theta_2} = \cdots = \frac{\theta_c}{\theta_c}\). Based on the above analysis, theorem 1 is proved.
By combining Eq. (19) with $w_1 + w_2 + \cdots + w_c = 1$, we can solve the analytic solutions of $w_i (i = 1, 2, \ldots, c)$:

$$w_i = \frac{\sqrt{tr \left( \sum_{k=1}^{n_k} \sum_{l=1}^{n_l} A^T x_k^{(i)} y_l^{(i)T} B \right)}}{\sum_{j=1}^{c} tr \left( \sum_{k=1}^{n_k} \sum_{l=1}^{n_l} A^T x_k^{(j)} y_l^{(j)T} B \right)}$$

(21)

In summary, the specific algorithm steps are listed in Algorithm 1.

**Algorithm 1 Dynamic Discriminative Canonical Correlation Analysis**

**Input:**
1. Training data: $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d_1 \times n}$, $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{d_2 \times n}$.
2. Label information: $Z = [z_1, \ldots, z_n]^T \in \mathbb{R}^c$.
3. The maximum number of iterations $T$.

**Initialization:** $t = 0$ and $w_i = 1/\sqrt{c}$, $(i = 1, 2, \ldots, c)$
1. Obtain the correlation projection matrices $A^{(t)}$ and $B^{(t)}$ by solving Eq. (15) and (16);
   while $t < T$
2. Update the adaptive weight matrix $W \in \mathbb{R}^{n \times n}$ by Eq. (21);
3. Obtain the correlation projection matrices $A^{(t+1)}$ and $B^{(t+1)}$ by solving Eq. (15) and (16);
4. if $\|A^{(t)} - A^{(t+1)}\|_2^2 \leq 0.0001$ and $\|B^{(t)} - B^{(t+1)}\|_2^2 \leq 0.0001$,
   then stop;
4. else $t = t+1$;
**end while;**

**Output:** The optimal correlation projection direction matrices $A^{(t+1)}$ and $B^{(t+1)}$.

C. **COMPLEXITY ANALYSIS**

The computation cost of our method mainly depends on the cost of the loop body. The computation complexity of Step 2 and Step 3 are both $O \left( \max \left( d_1^3, d_2^3 \right) \right)$, where $d_1$ and $d_2$ is the larger dimension between $X$ and $Y$ respectively. In general, the overall computation complexity of DDCCA is $O \left( T \max \left( d_1^3, d_2^3 \right) \right)$, where $T$ is the number of iteration.

Benefiting from the pre-process, the data of two views have suffered a proper dimension reduction in the practical procedure. That is to say, the total computation cost is not enough affected by the high dimensions of the raw data. Thus DDCCA has low computation complexity.

IV. **EXPERIMENTAL RESULTS AND ANALYSIS**

In this section, a variety of face recognition [41] and text classification [42] experiments are carried out on several real-world image and text datasets to evaluate the effectiveness of DDCCA. For real-world image datasets, we employ the view strategy [41] to obtain two view data of each image. Concretely, by using Coiflets and Daubechies wavelet transform technology [44], we can obtain two low-frequency sub-images of each image, and then K-L transformation [45] is utilized to reduce the dimensionality of two sub-images so that two view data of each image are obtained. Simultaneously, the performance of DDCCA is compared with that of CCA, DCCA, OCCA, GMALPP, and graph regulation multi canonical correlation analysis (GrMCCA) [51]. In the final recognition stage, we utilize serial feature fusion strategy [43], [41] and K nearest neighbor classifiers based on Euclidean distance [46].

A. **INTRODUCTION TO DATA**

We performed face recognition and text classification on these three image datasets and a text dataset respectively, and the detailed dataset form will be elaborated in this subsection.

Umist [47] dataset. The dataset contains 575 grayscale face images of 20 different people. These images belonging to the same individual vary according to the angle from the side face to the front face. Raw data had been manually cropped and resized to 112 x 92 pixels by Daniel Graham at Umist.

Yale [48] dataset. The dataset is made up of facial images of 15 people. Each person takes 11 facial images with different kinds (normal front, wearing glass, lighting conditions, expressions, eyes closed). The size of each image is 100x100 pixels.

YaleB [49] dataset. It includes 38 different individuals. And for each people, nearly 64 frontal images are under different lighting directions. Each image is 32 x 32 pixels grayscale.

Reuters [50] Multilingual dataset. It contains six subsets of 1200 documents over six labels or topics (E21, CCAT, M11, GCAT, C15, and ECAT) and is described in five languages of 2000 words each.

Without the loss of generality, we adopt average accuracy and standard deviation for better evaluating the performance of our method.

B. **EXPERIMENTS USING THE UMIT DATASET**

In this experiment, we randomly select $u$ ($u = 2, 3, 4$) images per individual as training samples, and the rest images are utilized as testing samples. Besides, ten independent experiments have been finished, and the average recognition rates of 10 experiments are displayed in Table 1.

From Table 1, we can see that the recognition results of DDCCA are the best among all the methods. In addition, Table 1 proves that the correlation of cross-view data CCA considered can do well to recognition. DCCA utilizes the label information to approximate the distribution of data in the real world so that DCCA outperforms traditional CCA. Although GMALPP considers the locality information of data, 2 train samples per class are too few to help GMALPP construct a reliable manifold benefitting training a discriminative model. Besides, the graph information constructed by GrMCCA with 2 train samples per class as well fails to extract effective discriminative features from Table 1. As the number...
of samples in each individual increases, the recognition rate also increases. It is consistent with theory.

The results of each random experiment are shown in Fig 1. The accuracy of 10 independent experiments denotes a relatively stable fluctuation and higher recognition rates of DDCCA compared with the other four methods. Besides, OCCA performed the worst, but OCCA performed steadily in 3 training samples per individual and 4 training samples per individual, which shows the stability of orthogonality.

C. EXPERIMENTS USING THE YALE DATASET

In this experiment, we randomly select \( u (u = 2, 3, 4) \) images per individual as training samples. The rest images are utilized as testing samples. Besides, ten independent
S. Su et al.: Dynamic Discriminative CCA via Adaptive Weight Scheme

D. EXPERIMENTS USING THE YALEB DATASET

In the implementation, we randomly select \( u(u = 2, 3, 4) \) images per individual as training samples, the rest images are utilized as testing samples. Besides, ten independent experiments have been finished, and the average recognition rates of 10 experiments are displayed in Table 3 and Fig 2 demonstrates the accuracy of methods versus variation of the dimension.

As can be seen from Table 3, DDCCA still has better recognition performance most time. The standard deviation of OCCA and GrMCCA in Tables 2 and 3 performs relatively

TABLE 5. Experimental results on the Reuters dataset.

| Method   | DDCCA     | CCA       | DCCA     | OCCA     | GMALPP    | GrMCCA    |
|----------|-----------|-----------|----------|----------|-----------|-----------|
| Subset1  | 67.49±1.15 | 54.47±1.31 | 67.54±1.12 | 59.6±1.20 | 30.38±3.85 | 43.58±1.22 |
| Subset2  | 67.82±1.75 | 52.71±1.90 | 67.63±1.23 | 57.11±2.18 | 31.92±3.16 | 43.14±2.47 |
| Subset3  | 68.97±1.54 | 58.18±1.06 | 68.92±2.06 | 61.37±0.98 | 33.49±4.79 | 45.63±1.89 |
| Subset4  | 69.71±1.15 | 57.96±1.54 | 69.46±1.41 | 61.90±1.39 | 30.17±3.20 | 56.46±2.74 |
| Subset5  | 65.93±0.77 | 56.4±1.73 | 65.79±0.59 | 60.67±2.15 | 30.54±3.15 | 48.30±1.79 |
| Subset6  | 66.26±1.37 | 52.13±1.61 | 66.02±1.17 | 56.89±1.24 | 27.64±5.75 | 42.18±1.51 |

1 A ± D: A is average classification accuracy (%), and D is the standard deviation.
TABLE 6. Experimental results on the Reuters dataset.

| Method  | DDCCA    | CCA       | DCCA     | OCCA     | GMALPP   | GrMCCA   |
|---------|----------|-----------|----------|----------|----------|----------|
| EN&FR   | 66.02±1.10 | 54.78±1.76 | 65.71±1.19 | 58.96±1.67 | 30.80±2.85 | 42.86±1.15 |
| EN&GR   | 66.93±1.16 | 56.76±1.38 | 66.50±1.00 | 61.19±1.97 | 29.92±2.49 | 43.99±2.10 |
| EN&IT   | 65.74±1.37 | 55.71±1.12 | 65.54±1.25 | 60.17±0.88 | 30.99±2.55 | 44.43±2.02 |
| EN&SP   | 66.88±0.87 | 54.66±1.63 | 66.69±1.40 | 59.63±1.17 | 29.90±2.65 | 43.24±1.71 |
| FR&GR   | 65.93±0.77 | 56.40±1.73 | 65.79±0.59 | 60.67±2.15 | 30.54±3.15 | 43.30±1.79 |
| FR&IT   | 64.77±1.15 | 55.51±1.29 | 64.90±1.23 | 60.22±1.68 | 31.47±3.60 | 44.00±1.64 |
| FR&SP   | 65.36±1.00 | 54.88±1.59 | 65.46±0.84 | 59.96±1.88 | 30.37±3.38 | 42.83±1.52 |
| GR&IT   | 65.36±0.94 | 56.09±1.63 | 65.33±0.97 | 61.08±1.32 | 30.76±2.53 | 43.94±1.97 |
| GR&SP   | 66.60±1.34 | 54.89±0.88 | 66.39±1.54 | 60.31±1.53 | 29.84±3.26 | 44.31±1.63 |
| IT&SP   | 76.22±1.38 | 54.38±1.26 | 65.49±1.03 | 60.68±1.24 | 31.22±2.65 | 43.37±1.90 |

1 A ± D: A is average classification accuracy (%), and D is the standard deviation.
2 EN, FR, GR, IT, and SP stand for English, French, German, Italian, and Spanish.

low in experiments, which also indicates that OCCA and GrMCCA have great robustness to the different datasets in face recognition.

From Figure 2, we can see that the recognition rate basically rises as the dimensionality rise. Besides, the recognition rate corresponding to the projection vector dimension of DDCCA rises rapidly in 1-20 dimensions, reaches a peak in 20-40 dimensions early, and maintains a certain level in 40-100 dimensions. It shows the dynamic strategy helps our method solved quickly and obtain a stable optimal solution.

In addition, the results of DDCCA are consistent with that of DCCA, which demonstrates DCCA has taken a great response to class distribution in this dataset.

In total, thanks to the adaptive weight scheme and dynamic strategy, DDCCA obtains excellent performance handling with face recognition and suits different class distribution in different face datasets.

E. EXPERIMENTS ON TEXT CLASSIFICATION

To evaluate the effectiveness of the methods again, we select multiple-linguistic text classification in this section. The Reuters multilingual dataset has six subsets, and according to the order of the website article, the six subsets are from Subset1 to Subset6.

From Table 4, we can see that DDCCA outperforms other feature learning methods in most cases. As the number of samples per topic increases, the classification accuracy also increases.

Second, we randomly select 50 articles written in English and French versions per topic under six subsets. Table 5 presents the classification performance of DDCCA is superior to other methods in all subsets. It proves that DDCCA can fit different datasets.

At last, we randomly select 50 articles per topic written in two languages (all 10 combinations of 5 languages here) on the final subset. As shown in Table 6, DDCCA still performs greater than other methods under experiments in most cases. It further shows that DDCCA can fit different languages.

The results in Tables 4, 5 and 6 from the point of view of training samples, different subsets, and different languages fully showed the superiority of self-adaptation. Besides, it is observed that the supervisory can be conducive to text classification through the performance of DDCCA and DCCA. And its dynamic solution promises its robustness and great performance of classification according to the standard deviation and average accuracy.

V. CONCLUSION

Supervisory information and multiple correlations are beneficial to the discrimination power of cross-view feature learnings. However, the methods above fail to be sensitive to class distribution. In this paper, we propose the novel cross-view dynamic discriminative feature learning method, i.e. DDCCA. To make full use of supervisory information, each class pair is assigned an automatic learning adaptive weight in the low-dimensional feature space. And we adopt Cauchy-Schwarz inequality and Lagrange multiplier obtaining the iterative local optimal solutions to gain a global stable optimal solution maximizing the correlation between two views in this iterative model. A serial of experiments on both face image datasets and multi-linguistic text datasets demonstrates that DDCCA has positive accuracy improvement compared with existing competing methods. The future research will focus on the view extension with self-learning correlation rules.

ACKNOWLEDGMENT

(Shuzhi Su and Penglian Gao contributed equally to this work.)

REFERENCES

[1] M. Aghaei, M. Bustreo, Y. Wang, G. Bailo, P. Moreiro, and A. Del Bue, “Single image human proxemics estimation for visual social distancing,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2021, pp. 2785–2795.

[2] A. Kadkhodamohammadi and N. Padoy, “A generalizable approach for multi-view 3D human pose regression,” Mach. Vis. Appl., vol. 32, no. 1, pp. 1–14, Jan. 2021.
[3] R. Tao, X. Zhao, W. Li, H. Li, and Q. Du, “Hyperspectral anomaly detection by fractional Fourier entropy,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 12, pp. 4920–4929, Dec. 2019.

[4] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, “Multilabel prediction via cross-view search,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 4, pp. 4324–4338, Sep. 2017.

[5] M. Zhang, W. Li, and Q. Du, “Diverse region-based CNN for hyperspectral image classification,” IEEE Trans. Image Process., vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[6] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, “Information fusion for classification of hyperspectral and LiDAR data using IP-CNN,” IEEE Trans. Geosci. Remote Sens., early access, Jul. 13, 2021, doi: 10.1109/TGRS.2021.3093334.

[7] X. Zhao, R. Tao, W. Li, H.-C. Li, Q. Du, W. Liao, and W. Philips, “Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 10, pp. 7355–7370, Oct. 2020.

[8] X. Zhao, R. Tao, W. Li, W. Philips, and W. Liao, “Fractional Gabor convolutional network for multsource remote sensing data classification,” IEEE Trans. Geosci. Remote Sens., early access, Mar. 23, 2021, doi: 10.1109/TGRS.2021.3085507.

[9] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, “Multisource remote sensing data classification based on convolutional neural network,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 937–949, Feb. 2018.

[10] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, “Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN,” IEEE Trans. Cybern., vol. 50, no. 1, pp. 100–111, Jan. 2020.

[11] X. Shen, F. Shen, Q.-S. Sun, Y. Yang, Y.-H. Yuan, and H. T. Shen, “Semi-paired discrete hashing: Learning latent hash codes for semi-paired cross-view retrieval,” IEEE Trans. Cybern., vol. 47, no. 12, pp. 4275–4288, Dec. 2017.

[12] X. Shen, F. Shen, L. Liu, Y.-H. Yuan, W. Li, and Q.-S. Sun, “Multiview discrete hashing for scalable multiview search,” ACM Trans. Intel. Syst. Technol., vol. 9, no. 5, pp. 1–21, Jul. 2018.

[13] M. Sorensen, C. I. Kanatsoulis, and N. D. Sidiropoulos, “Generalized canonical correlation analysis: A subspace intersection approach,” IEEE Trans. Signal Process., vol. 69, pp. 2452–2467, 2021.

[14] W. Dong, J. Wu, Z. Bai, Y. Hu, W. Li, W. Qiao, and M. Wozniak, “MobileGCN applied to low-dimensional node feature learning,” Pattern Recognit., vol. 112, Apr. 2021, Art. no. 107788.

[15] R. Tao, X. Zhao, W. Li, H. Li, and Q. Du, “Hyperspectral anomaly detection based on fuzzy granules,” Knowl.-Based Syst., vol. 29, no. 9, pp. 3048–3057, Sep. 2017.

[16] T. Lindeberg, “Scale invariant feature transform,” Int. J. Comput. Vis., vol. 50, no. 2, pp. 7–25, Nov. 2002.

[17] R. D’ouza, H. Moum, and F. X. Yeh, “Structural analysis and optimization of convolutional neural networks with a small sample size,” Sci. Rep., vol. 10, no. 1, pp. 1–13, Dec. 2020.

[18] M. L. W. Smeeck, J. Agramunt, and K. M. Bonger, “Recent developments in bioorthogonal chemistry and the orthogonality within,” Current Opinion Chem. Biol., vol. 60, pp. 79–88, Feb. 2021.

[19] X.-B. Shen, Q.-S. Sun, and Y.-H. Yuan, “Orthogonal canonical correlation analysis and its application in feature fusion,” in Proc. 16th Int. Conf. Inf. Fusion, Jul. 2013, pp. 151–157.

[20] D. S. Wilks, “Probabilistic canonical correlation analysis forecasts, with application to tropical Pacific sea-surface temperatures,” Int. J. Climatol., vol. 34, no. 5, pp. 1405–1413, Apr. 2014.

[21] C. Tang, X. Zheng, X. Liu, W. Zhang, J. Zhang, J. Xiong, and L. Wang, “Cross-view locality preserved diversity and consensus learning for multi-view unsupervised feature selection,” IEEE Trans. Knowl. Data Eng., early access, Jan. 1, 2021, doi: 10.1109/TKDE.2020.3048878.

[22] C. Tang, X. Zhu, X. Liu, M. Li, P. Wang, C. Zhang, and L. Wang, “Learning a joint affinity graph for multiview subspace clustering,” IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1724–1736, Jul. 2019.

[23] J. Wan and F. Zhu, “Cost-sensitive canonical correlation analysis for semi-supervised multi-view learning,” IEEE Signal Process. Lett., vol. 27, pp. 1330–1334, Dec. 2020.

[24] S. Solerio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “A review of unsupervised feature selection methods,” Artif. Intell. Rev., vol. 53, no. 2, pp. 907–948, Feb. 2020.

[25] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 161–168.

[26] A. Sharma, A. Kumar, H. Daumé, and D. W. Jacobs, “Generalized multi-view analysis: A discriminative latent space,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2012, pp. 2160–2167.
[48] M. Sharif, S. Mohsin, M. J. Jamal, and M. Raza, “Illumination normalization preprocessing for face recognition,” in Proc. 2nd Conf. Environ. Sci. Inf. Appl. Technol., Jul. 2010, pp. 44–47.

[49] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, “Is face recognition really a compressive sensing problem?” in Proc. CVPR, Jun. 2011, pp. 553–560.

[50] K. Yu, H. Li, and B. Oguz, “Multilingual Seq2seq training with similarity loss for cross-lingual document classification,” in Proc. 3rd Workshop Represent. Learn. (NLP), 2018, pp. 175–179.

[51] Y.-H. Yuan and Q.-S. Sun, “Graph regularized multiset canonical correlations with applications to joint feature extraction,” Pattern Recognit., vol. 47, no. 12, pp. 3907–3919, 2014.

SHUZHI SU received the Ph.D. degree from the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. He is currently an Associate Professor with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. His research interests include multimodal pattern recognition, information fusion, feature learning, and image processing.

PENGLIAN GAO received the B.S. degree from the Department of Internet of Things Engineering, Anhui University of Science and Technology, Huainan, China, where he is currently pursuing the M.S. degree in computer technology. His current research interests include dimensionality reduction, image recognition, and text classification.

YANMIN ZHU received the M.S. degree from the School of Mechanical and Electronic Engineering, Shandong Agricultural University, Tainan, China. She is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Anhui University of Science and Technology, Huainan, China. Her research interests include multi-modal information fusion, correlation projection theory, and image processing.

XINGZHU LIANG was born in Hefei, Anhui, China, in 1979. He received the M.S. degree in computer science from Daqing Petroleum Institute, China, in 2006. He is currently an Associate Professor with the School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, China. His research interests include computer vision, computer network security, and pattern recognition.

JUN XIE received the B.S. degree from the Department of Computer Science and Technology, Medical University, Jining, China. He is currently pursuing the M.S. degree in computer technology with Anhui University of Science and Technology, Huainan, China. His current research interests include dimensionality reduction and image recognition.