Abstract—Based on the consideration that multiset integrated canonical correlation analysis (MICCA) does not include the class information of the samples, this paper presents a discriminative learning version of MICCA, called discriminative-classification of multiset integrated canonical correlations (DMICC). The extracted features by DMICC not only contain the class information of training samples, but also possess more powerful discriminant ability than those by MICCA. The proposed DMICC method is evaluated on the AR, ORL face image databases, and the COIL-20 object database. The experimental results on face and object recognition demonstrate that DMICC is significantly superior to MICCA.

Keywords-pattern recognition; multiset canonical correlation analysis; discriminant analysis; feature extraction; feature fusion

I. INTRODUCTION

In pattern recognition, feature fusion has become an important research focus. In modern pattern classification systems, information fusion techniques can be broadly partitioned into three classes: pixel level fusion, feature level fusion and decision level fusion [1, 2]. Since feature fusion can reduce the redundant information existing in original features and enhance the recognition accuracy for classification, it is in conflict with other two kinds of fusion, that is, it has been of wide concern and some delightful research results [3-5] are obtained. In practice, feature fusion has been widely used in numerous applications. Multiset feature vectors, which are extracted from the same patterns by adopting different extraction methods, always reflect different characteristics of patterns. These features can be effectively fused by some given fusion strategies to form effective feature vectors for recognition. Therefore, feature fusion undoubtedly is very important for pattern classification in many applications.

Two classical techniques, principal component analysis (PCA) [6] and linear discriminant analysis (LDA) [7], have been widely used in feature extraction and fusion for face and digit image recognition. In earlier work [3-5], the way of feature fusion is to simply concatenate or integrate multi-group features together. Although the fusion way can improve the recognition accuracy to some extent in classification, it ignores the intrinsic relationship among multiple sets of features. To reveal the intrinsic correlation, canonical correlation analysis (CCA) [8] is used to fuse two-set different features for effective face and handwritten numeral recognition [1]. Subsequently, Sun et al. further improved the feature fusion method using CCA and proposed a generalized CCA (GCCA) [9] for two-set feature fusion. The experiment on handwritten numeral classification indicates that GCCA outperforms CCA. Additionally, a new fusion method [10] using partial least squares (PLS) has been proposed for two-set feature fusion. This method improves the existing feature fusion methods.

Recently, a novel multiset integrated canonical correlation analysis (MICCA) framework [11] has been presented for multiset feature fusion. This fusion method first extracts multiset features from the same patterns by different feature extraction methods, and then it extracts multiset integrated canonical correlation features (MICCFs) to form effective feature vectors by means of the given fusion strategies for classification. A series of experiments have demonstrated the effectiveness and robustness of the MICCA method. But despite this, the feature fusion method based on MICCA does not contain class information of samples, that is, the extracted MICCFs may not be optimal for pattern classification tasks.

Based on this consideration, we propose a discriminative learning version of MICCA, called discriminative-analysis of multiset integrated canonical correlations (DMICC). In DMICC, the class information of training samples has been added, which can significantly improve the recognition rates in classification tasks. The experimental results on the AR, ORL face image databases and the COIL-20 object database have demonstrated that the proposed DMICC method outperforms MICCA.

II. OUTLINE OF MCCA AND MICCA

A. Multiset Canonical Correlation Analysis

Multiset canonical correlation analysis (MCCA) [12] can be viewed as a natural extension of CCA. Mathematically, MCCA has many different optimization models under different metric functions and constraints [12, 13]. The following MCCA method which is known as SUMCOR has been examined in [13]. Specifically, given $m$ real random vectors $x_1, x_2, \ldots, x_m$ with dimensions $p_1, p_2, \ldots, p_m$, the objective of MCCA is to find projection directions $\alpha_1, \alpha_2, \ldots, \alpha_m$ that maximize the sum of the pair-wise correlation between the projected variables...
\( \alpha_i^T x_1, \alpha_2^T x_2, \ldots, \alpha_m^T x_m \) under some constraints. It is defined as the optimization problem:

\[
\max \rho(\alpha_1, \alpha_2, \ldots, \alpha_m) = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^T S_{ij} \alpha_j, \tag{1}
\]

s.t. \( \alpha_i^T S_{ij} \alpha_i = 1, i = 1, 2, \ldots, m. \)

where \( S_{ij} \) is a within-set covariance matrix of random vector \( x_i \), and \( S_{ij} (i \neq j) \) is a between-set covariance matrix between random vectors \( x_i \) and \( x_j \).

The Lagrangian of this optimization problem is

\[
L = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^T S_{ij} \alpha_j - \lambda_i (\alpha_i^T S_{ii} \alpha_i - 1), \tag{2}
\]

where \( \{\lambda_i\}_{i=1}^{m} \) are Lagrange multipliers. By setting \( \partial L / \partial \alpha_i = 0 \), we obtain

\[
\sum_{j=1}^{m} S_{ij} \alpha_j = \lambda_i S_{ii} \alpha_i, \quad i = 1, 2, \ldots, m. \tag{3}
\]

Since (3) is a very complex and abnormal generalized eigenvalue problem, it is difficult for MCCCA to gain the accurate solution vectors. At present, there have been some iterative algorithms [14-16] for solving (3).

### B. Multiset Integrated Canonical Correlation Analysis

MICCA is a technique that analyzes linear integrated relations among multiple sets of random variables. It computes a set of projection directions \( \alpha_1, \alpha_2, \ldots, \alpha_m \), called multiset integrated canonical vector (MICVec), to maximize the multiset integrated canonical correlation among the projected variables \( \alpha_1^T x_1, \alpha_2^T x_2, \ldots, \alpha_m^T x_m \), as follows:

\[
\rho = \max_{\alpha_1, \ldots, \alpha_m} \left[ -\det(G(\alpha_1^T x_1, \alpha_2^T x_2, \ldots, \alpha_m^T x_m)) \right]^{1/2}, \tag{4}
\]

where \( G(\cdot) \) represents a Gram matrix, \( \det(f) \) denotes the determinant of square matrix \( f \), and \( \| f \| \) is the 2-norm of \( f \).

According to [11], if we assume that \( (s-1) \) sets of MICVecs \( \{\alpha_1(1)\}_{i=1}^{m}, \{\alpha_2(1)\}_{i=1}^{m}, \ldots, \{\alpha_{s-1}(1)\}_{i=1}^{m} \) are obtained, then the \( s \)th set of MICVec \( \{\alpha_s(1)\}_{i=1}^{m} \) can be found by solving the following optimization problem:

\[
\min \{ \det(S_{\text{proj}}) \}
\]

s.t. \( \alpha_i^T S_{ij} \alpha_i = 1 \), \( \alpha_i^T S_{ij} \alpha_j = 0 (j = 1, 2, \ldots, s - 1) \), \( \alpha_i \in \mathbb{R}^p \) \( (i = 1, 2, \ldots, m) \)

where \( S_{\text{proj}} \) represents the covariance matrix of the projected variables \( \alpha_1^T x_1, \alpha_2^T x_2, \ldots, \alpha_m^T x_m \), i.e.,

\[
S_{\text{proj}} = \begin{bmatrix}
\alpha_1^T S_{11} \alpha_1 & \alpha_1^T S_{12} \alpha_2 & \cdots & \alpha_1^T S_{1m} \alpha_m \\
\alpha_2^T S_{21} \alpha_1 & \alpha_2^T S_{22} \alpha_2 & \cdots & \alpha_2^T S_{2m} \alpha_m \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_m^T S_{m1} \alpha_1 & \alpha_m^T S_{m2} \alpha_2 & \cdots & \alpha_m^T S_{mm} \alpha_m 
\end{bmatrix}. \tag{5}
\]

### III. Discriminative-Analysis of Multiset Integrated Canonical Correlations (DMICC)

MICCA is an unsupervised technique in subspace learning. In order to improve the classification accuracy of MICCA, we will introduce class information into the multiset integrated canonical correlation analysis to construct its discriminative learning algorithm for feature fusion.

Assume \( m \) high-dimensional feature representations from the same \( n \) patterns are given as \( \{X_i \in \mathbb{R}^{p \times n} \}_{i=1}^{m} \), where \( X_i = (x_{i1}^{(1)}, x_{i2}^{(1)}, \ldots, x_{in}^{(1)}) \) and \( x_{ij}^{(s)} \in \mathbb{R}^p \) \( (s = 1, 2, \ldots, n) \), and \( p_i \) denotes the dimension of the \( i \)th feature representation. Each training sample \( x_{ij}^{(s)} \) in feature representation \( X_i \) is labeled by \( l_i \), where \( k = 1, 2, \ldots, c \) for the \( c \) classes. Let \( \{\alpha_i^{(s)}\}_{i=1}^{m} \) be one set of projection directions which linearly project \( \{X_i\}_{i=1}^{m} \) to \( \{\alpha_i^{(s)} X_i\}_{i=1}^{m} \). In DMICC, our goal is to introduce class information of training samples into within-set and between-set covariance matrices for benefiting pattern classification.

**Definition 1.** The matrix \( S_{ii}^{(w)} \) is referred to as within-set within-class covariance matrix (WWCM) of the \( i \)th representation \( X_i \) if and only if

\[
S_{ii}^{(w)} = \sum_{k=1}^{c} \sum_{j=1}^{n} \frac{1}{n_k} (x_{ij}^{(s)} - m_{ik}^{(s)})(x_{ij}^{(s)} - m_{ik}^{(s)})^T, \tag{6}
\]

where \( x_{ij}^{(s)} \) represents the \( j \)th training sample of the class \( k \) in the \( i \)th feature representation, \( n_k \) is the number of training samples in class \( k \), and \( m_{ik}^{(s)} \) denotes the mean vector of training samples of class \( k \) in the \( i \)th feature representation.

Obviously, \( S_{ii}^{(w)} \) is a symmetric and nonnegative definite matrix. Here, we suppose that all WWCMs, i.e., \( \{S_{ii}^{(w)}\}_{i=1}^{m} \), are positive definite. If \( S_{ii}^{(w)} \) is not positive definite in some applications, then a small perturbation is added into \( S_{ii}^{(w)} \) to avoid its singularity according to the perturbation idea in [17].

**Definition 2.** The matrix \( S_{ij}^{(b)} \) is referred to as between-set within-class covariance matrix (BWCM) of two different feature representations \( X_i \) and \( X_j \) if and only if
where \( i \neq j \) and \( i, j = 1, 2, \ldots, m \), and the rest of parameters have the same meaning as those in (7).

Clearly, \( S_y^{(b)} \) satisfies \( S_y^{(b)T} = S_y^{(b)} \) in Definition 2. Using (7) and (8) in Definitions 1 and 2, we can obtain such within-set and between-set covariance matrices in which the supervised information is contained. Under the case of discriminative learning, the optimization problem in (5) can be rewritten as

\[
\min \{ \det(S_D) \} \\
\text{s.t.} \quad \langle \alpha_i^T S_y^{(w)} \rangle = 0 (j = 1, 2, \ldots, s - 1), \\
\alpha_i \in R^p (i = 1, 2, \ldots, m)
\]

where \( S_D = S_D^T \), and

\[
S_D = \begin{bmatrix}
\alpha_1^T S_y^{(w)} & \alpha_2^T S_y^{(w)} & \cdots & \alpha_m^T S_y^{(w)} \\
\alpha_1^T S_y^{(w)} & \alpha_2^T S_y^{(w)} & \cdots & \alpha_m^T S_y^{(w)} \\
\vdots & \vdots & \ddots & \vdots \\
\alpha_1^T S_y^{(w)} & \alpha_2^T S_y^{(w)} & \cdots & \alpha_m^T S_y^{(w)}
\end{bmatrix}.
\]

As each \( S_y^{(w)} \) is symmetric and positive definite, the Cholesky decomposition \( S_y^{(w)} = C_s C_s^T \) exists. Let \( \beta_i = C_s^T \alpha_i \), then the optimization problem in (9) can be formulated equivalently as

\[
\min \{ \det(S_D) \} \\
\text{s.t.} \quad \langle \beta_i^T \beta_i \rangle = 1 \\
\beta_i \in R^p (i = 1, 2, \ldots, m)
\]

where \( S_D = \hat{S}_D \)

\[
\hat{S}_D = \begin{bmatrix}
\beta_1^T \beta_1 & \beta_1^T H_{12} \beta_2 & \cdots & \beta_1^T H_{1m} \beta_m \\
\beta_2^T H_{21} \beta_1 & \beta_2^T \beta_2 & \cdots & \beta_2^T H_{2m} \beta_m \\
\vdots & \vdots & \ddots & \vdots \\
\beta_m^T H_{m1} \beta_1 & \beta_m^T H_{m2} \beta_2 & \cdots & \beta_m^T \beta_m
\end{bmatrix}
\]

where \( \beta_i = C_s^{-1} S_y^{(b)} C_s^T (i \neq j) \).

In the light of the idea in [11], to solve the optimization problem in (10), we approximately transform it into the following optimization model without constraints:

\[
\max L = \sum_{i=1}^{m} \sum_{j=1}^{m} (\beta_i^T H_{ij} \beta_j)^2 - \sum_{i=1}^{m} \sum_{i=1}^{m} \xi_i \beta_i^T \beta_i - \sum_{i=1}^{m} \xi_i \beta_i^T \beta_i - \sum_{i=1}^{m} \xi_i \beta_i^T \beta_i,
\]

where \( \lambda_i \) and \( \xi_i \) are Lagrange multipliers, and when \( t = 0 \), let \( \xi_i = 0 \) and \( \beta_i \) be a zero vector. By setting \( \partial L / \partial \beta_i = 0 \), we obtain

\[
\sum_{j=1}^{m} (\beta_i^T H_{ij} \beta_j) H_{ij} \beta_j - \sum_{i=1}^{m} \xi_i \beta_i = \lambda_i \beta_i,
\]

where \( i = 1, 2, \ldots, m \).

Obviously, the \( m \) equations in (12) are very complex. It is difficult to obtain their analytical solutions. Thus, we give their iterative solutions based on the power-successive over-relaxation (PSOR) algorithm [11, 18].

From (12), it can be clearly seen that if \( \beta_{1i}, \beta_{2i}, \ldots, \beta_{mi} \) are the solution vectors of the equations in (12), then they must be the \( s \)th set of solutions of the optimization problem in (11). According to \( \beta_i = C_s^T \alpha_i \), we use \( \alpha_i = C_s^{-1} \beta_i \) to get the \( s \)th set of projection directions of DMICC.

After obtaining \( d \) sets of projection directions \( \{\alpha_i\}_{i=1}^{m} \), \( \{\alpha_{1i}\}_{i=1}^{m}, \ldots, \{\alpha_{di}\}_{i=1}^{m} \) from \( m \) training sample spaces (i.e., \( m \) different representations), let \( W_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{id}) \in R^{p \times d} \), then a projection matrix \( W = \text{diag}(W_1, W_2, \ldots, W_m) \) can be formed for any testing sample \( y^T = (y_{11}, y_{12}, \ldots, y_{1m}) \) with \( y_{ij} \in R^p \). Thus, the testing sample \( y \) can be extracted \( m \) sets of feature components, i.e., \( W'y \). In multiset feature fusion, two effective fusion strategies have been proposed, called serial feature fusion and parallel feature fusion [11], respectively. In this paper, we adopt the parallel feature fusion strategy to fuse multiset feature components, i.e.,

\[
y_{\text{fusion}} = \sum_{i=1}^{m} W_i^T y_i.
\]

The fused features obtained by (13) are called multiset integrated canonical-correlational discriminative features (MIDCFs), which are used for pattern classification.

IV. EXPERIMENT AND ANALYSIS

In this section, the performance of the proposed DMICC method is evaluated on three well-known databases (i.e., the AR, ORL face image databases, and the COIL-20 object database) and compared with the performance of the MICCA method.

A. Data Preparation

The AR database [19] contains over 4000 color face images of 126 individuals (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of 120 individuals are taken in two sessions and each section contains 13 color images. In our experiments, the images of 120 people are selected from both sessions and one person has 14 images, which are full facial images (i.e. not occluded). Each face image is manually cropped and normalized to 50×40 pixels.

The ORL database [20] contains 400 face images from 40 individuals. For each individual, there are 10 different grayscale images with a resolution of 92×112. Some of these
images are taken at different times. The facial expressions are also varied. The images are taken with a tolerance for some tilting and rotation of the face of up to 20 degrees, and have some variation in the scale of up to about 10 percent.

The COIL-20 database [21] is widely used in object recognition. This database consists of grayscale images of 20 different objects; each one is rotated with 5 degree angle interval. There are 72 images for each object and 1440 images for the whole database.

In our experiments, the Coiflets, Daubechies and Symlets wavelet transforms are used to extract three sets of low-frequency feature vectors (i.e., three feature representations) from original images. An important reason that we use low-frequency sub-images from wavelet transforms is that they contain more shape information than high-frequency sub-images. After obtaining three sets of feature vectors, the K-L transform is used to reduce their dimensions to 150, 150, and 150 dimensions, respectively. The final formed three-set feature vectors are used in the following experiments.

B. Experiment Using the AR Database

In this experiment, seven images per individual are randomly chosen for training, while the remaining seven images are used for testing. Therefore, the total number of training samples and testing samples is, respectively, 840 and 840. In this way, ten independent tests are run, and then the average recognition rates are computed for the performance evaluation. MICDFs and MICCFs are, respectively, extracted by DMICC and MICCA from above-mentioned three sets of features. After multiset feature fusion, the nearest neighbor (NN) classifiers with Euclidean and cosine distance metrics are used for recognition. The minimum average recognition error rates of MICCA and DMICC and the corresponding dimension are given in Table I. The average recognition rates of both methods versus the variation of the dimension are shown in Figs. 1 and 2.

From Table I, it can be seen that the proposed DMICC method obviously outperforms MICCA, no matter what distance metric is used. Its maximal recognition rate is up to 98.1 percent under cosine distance metric, and beyond that of MICCA 0.8 percent. This indicates that DMICC is indeed very discriminatory and effective for classification.

Fig. 1 shows that the average recognition rates of DMICC are consistently superior to those of MICCA along with the increase of the dimension under Euclidean distance metric. Fig. 2 indicates that although the difference of the recognition rates between both methods is not obvious when the dimension is less than 95, DMICC performs better than MICCA under cosine distance metric when the dimension is over 95.

<table>
<thead>
<tr>
<th>Method</th>
<th>MICCA</th>
<th>DMICC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euclidean</td>
<td>Cosine</td>
</tr>
<tr>
<td>Error</td>
<td>0.041</td>
<td>0.027</td>
</tr>
<tr>
<td>Dim.</td>
<td>140</td>
<td>142</td>
</tr>
</tbody>
</table>

C. Experiment Using the ORL Database

In this experiment, for each individual, we randomly select five images as training samples, and the rest are considered as testing samples. Therefore, the total number of training samples is 200, and the total number of testing samples is 200. In this way, ten independent tests are run, and then the average recognition rates are computed for the performance evaluation. DMICC and MICCA are, respectively, employed for feature extraction. After multiset feature fusion, the cosine nearest neighbor classifier is employed for classification. The average recognition rates of MICCA and DMICC versus the variation of the dimension are shown in Fig. 3.
Fig. 3 shows that DMICC is consistently superior to MICCA as the number of the dimension increases continually. Meanwhile, it also indicates that when DMICC obtains the maximum average recognition rate, the corresponding dimension is less than 35.

D. Experiment Using the COIL-20 Database

In this experiment, \( l \) images (\( l = 24, 30, \) and \( 36 \)) per object are randomly chosen for training, while the remaining \( 72 - l \) images are used for testing. For each \( l \), ten independent runs are performed to evaluate the performance of DMICC and MICCA. After feature fusion based on DMICC and MICCA, the Euclidean and cosine NN classifiers are used for recognition. The minimal average recognition error rates of MICCA and DMICC and the corresponding dimension are shown in Table II. Taking the recognition results under 36 training samples per object as a representation, the average recognition rates versus the variation of the dimension are given in Figs. 4 and 5.

From Table II, we can see that DMICC significantly outperforms MICCA again under the different number of training samples. Especially under 36 training samples, the maximal average recognition rates of DMICC are, respectively, up to 99.2 percent and 99.1 percent, and exceed the corresponding those of MICCA 2.1 percent and 1.8 percent. This further demonstrates that the extracted features by DMICC possess more powerful discriminant ability in contrast to those by MICCA.

**TABLE II.** The minimal average recognition error rates of MICCA and DMICC across ten runs on the COIL-20 database and the corresponding dimensions (in parentheses)

<table>
<thead>
<tr>
<th>Method</th>
<th>MICCA</th>
<th>DMICC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Euclidean</td>
<td>Cosine</td>
</tr>
<tr>
<td>24 Train.</td>
<td>0.040(19)</td>
<td>0.037(19)</td>
</tr>
<tr>
<td>30 Train.</td>
<td>0.033(25)</td>
<td>0.032(25)</td>
</tr>
</tbody>
</table>

From Figs. 4 and 5, we can find that the recognition rates of DMICC are obviously better than those of MICCA whether the distance metric is Euclidean or cosine.

V. CONCLUSION

This paper has presented a discriminative version of MICCA, called discriminative-analysis of multiset integrated canonical correlations (DMICC). Since the proposed method includes within-class discriminant information, it can fuse multiset feature vectors under a discriminative learning framework. DMICC has been evaluated on face and object recognition. Its effectiveness has been demonstrated by a series
of experiments. DMICC significantly outperforms the MICCA method in pattern classification. For the future work, it will be interesting to introduce other discriminant information (e.g., class label information) into MICCA. Additionally, the nonlinear extension of DMICC should also be investigated for nonlinear feature fusion.

ACKNOWLEDGMENT

This work is partially supported by the National Science Foundation of China, and Doctoral Fund of Ministry of Education of China (RFDP) under Grant Nos. 60773172 and 200802880017. Meanwhile, it is also partially supported by the Scientific Research and Innovation Project Fund for Graduate Students of Jiangsu Provincial Higher Education Institutions under Grant No. CXZZ110260, and the Excellent Doctoral Graduate Education Fund of Nanjing University of Science & Technology (NUST).

REFERENCES