Multiple Feature Fusion by Subspace Learning

Yun Fu, Liangliang Cao
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{yunfu2,cao4}@uiuc.edu

Guodong Guo
Computer Science
North Carolina Central University
Durham, NC 27707, USA
gdguo@nccu.edu

Thomas S. Huang
Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
huang@ifp.uiuc.edu

ABSTRACT

Since the emergence of extensive multimedia data, feature fusion has been more and more important for image and video retrieval, indexing and annotation. Existing feature fusion techniques simply concatenate a pair of different features or use canonical correlation analysis based methods for joint dimensionality reduction in the feature space. However, how to fuse multiple features in a generalized way is still an open problem. In this paper, we reformulate the multiple feature fusion as a general subspace learning problem. The objective of the framework is to find a general linear subspace in which the cumulative pairwise canonical correlation between every pair of feature sets is maximized after the dimension normalization and subspace projection. The learned subspace couples dimensionality reduction and feature fusion together, which can be applied to both unsupervised and supervised learning cases. In the supervised case, the pairwise canonical correlations of feature sets within the same classes are also counted in the objective function for maximization. To better model the high-order feature structure and overcome the computational difficulty, the features extracted from the same pattern source are represented by a single 2D tensor. The tensor-based dimensionality reduction methods are used to further extract low-dimensional discriminative features from the fused feature ensemble. Extensive experiments on visual data classification demonstrate the effectiveness and robustness of the proposed methods.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement—*feature representation, projections*; I.5 [**Pattern Recognition**]: Design Methodology

General Terms

Algorithms

CIVR'08, July 7–9, 2008, Niagara Falls, Ontario, Canada.

Keywords

Feature fusion, subspace learning, canonical correlation, face recognition, image set, tensor.

1. INTRODUCTION

In modern image/video retrieval or pattern classification systems, the multimodality fusion strategies are often categorized by feature fusion, model fusion, and decision fusion [17]. Because of the simplicity, multimodality sensory data and multiple feature fusion are prevalent in existing real-world systems. The data/feature fusion scheme typically achieves boosted system performance or robustness, which attracts much attention of researchers from multimedia, computer vision, audio-visual speech processing, biomedical imaging and pattern recognition [17, 23]. Although the importance of data/feature fusion is obvious, there are still not any techniques that can manipulate this idea in generalized ways. Most existing techniques are still case-by-case in solving specific real-world problems. So, multimodality data or multiple feature fusion is still an open problem technically.

Conventional feature fusion methods simply concatenate or integrate several kinds of features together. Despite of the simplicity of such methods, the systems that adopting such feature fusion may still not perform better (or even worse) or more robust than using single features. This is because the information conveyed by different features is not equally represented or measured. The element values of different features can be significantly unbalanced. So, the equally weighted concatenation or integration is a suboptimal solution. In some cases, one or two features may dominate the entire system performance. The simplest way for the feature weighting is to normalize different feature value ranges or scales so that they are well balanced. For example, the resulting normalized features could have zero mean and unit variance [34] or equal sum of eigenvalues on eigenvector-based features [28]. However, in most cases, those features extracted from the same datum might be highly correlated to each other. The simple normalization or weighting can not be sufficiently helpful to make the fused feature effective for the classification purpose. On the other hand, one set of feature may dominate the effectiveness of a feature ensemble. The simple normalization or weighting may not perform well in practice.

A possible way for valid weighting is to perform joint dimensionality reduction or subspace learning by preserving the correlation between different feature pairs. For example, in [17, 21], Canonical Correlation Analysis (CCA) [21,

15] is used to fuse audio-visual features with joint subspace learning, in which the representations of projected audio and visual features in their own subspaces will be able to preserve the correlation conveyed from the original audio and visual feature spaces. This correlation can be explained as that the lip region within the face region is more related to speech signal than other facial parts. By maximizing the canonical correlation between the projected audio and visual features, the fused audio-visual feature is demonstrated to be efficient and effective for person verification based on a probabilistic classification model. In [24], CCA based feature pair fusion is used for effective face and handwritten Arabic numerals recognition. Another related method in [23] presented to use Partial Least Squares (PLS) [3] regression for feature pair fusion. Given two sets of features, the basic idea of PLS is to find a pair of directions such that the covariance between the projections of two feature sets is maximized. As proved by [3], PLS can be actually thought as penalized CCA with basically the Principal Components Analysis (PCA) [4] in the original two feature spaces providing the penalties. This method was used to fuse different feature pairs for face recognition and handwritten Arabic numerals classification, which improves the existing feature fusion methods.

In addition, feature pair fusion can also be achieved by parallel strategy [32], in which a complex vector is defined to represent the parallel combined features. Conventional subspace learning methods can be used jointly with this parallel feature fusion for generalized feature extraction in the complex feature space. Probabilistic fusion methods [20], Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Machine (SVM) [9] have also been proposed for classification-driven feature fusion.

The foregoing methods provide us the preliminary idea to fuse multimodality feature pairs in generalized joint subspace learning ways. However, they also remain challenging. For example, in addition to fusing a feature pair, how can we go further to fuse *multiple* feature sets (more than two) by measuring the canonical correlation between feature pairs? In this paper, we solve this problem by providing a generalized subspace learning solution instead of individual subspaces for each feature sets. By projecting all the features into a linear subspace, the sum of pairwise feature sets canonical correlation is maximize. The learned subspace couples dimensionality reduction and feature fusion together. This feature fusion method is designed for both unsupervised and supervised learning. In the supervised case, the pairwise feature sets canonical correlation for the same classes are also counted in the objective function for maximization. To deal with the computational difficulty introduced by existing feature fusion methods when the number of fused features is large, the features coming from the same pattern source are represented by a single 2D tensor. The tensor structure, for high-order feature patterns, may also introduce more powerful properties to represent the fused feature to boost the discriminating power. On the other hand, it may also avoid the curse-of-dimensionality dilemma and the small sample size problem [31, 30, 10, 15]. The tensor-based dimensionality reduction methods are adopted to jointly reduce dimensionality and extract low-dimensional discriminative features of the fused feature ensemble. We perform face recognition experiments to demonstrate the effectiveness and robustness of the proposed methods.

The contributions of the paper are summarized as follows.

- The multiple feature fusion problem is formulated as generalized subspace learning with canonical correlation based feature set measurement.
- A subspace learning method is presented to couple dimensionality reduction and feature fusion together, which can be used for both unsupervised and supervised learning.
- Tensor based analysis is applied to the feature fusion framework to better achieve the learning purpose.

In the rest of the paper, we first formulate the multiple feature fusion problem in section 2. The generalized subspace learning algorithm for multiple feature fusion is presented in section 3. In section 4, a tensor-based discriminative subspace learning method is proposed to deal with the computational difficulty for multiple feature fusion. Section 5 presents the feature extraction techniques used in the paper. Extensive experimental results on face recognition is reported in section 6. We make the conclusion of the paper in the last section.

2. PROBLEM FORMULATION

Suppose a given high-dimensional data set (original feature set) is denoted by $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^D$. For the supervised case, each datum is labeled by $l_i \in \mathcal{L}$, where $\mathcal{L} = \{l_1, l_2, \ldots, l_c\}$ for the c classes. After applying the m different feature extraction operations on each original datum, we obtain n feature sets $\mathcal{F}_i^{(1)} = \{\mathbf{f}_i^{(k)}\}_{k=1}^m$ for $i = 1, 2, \ldots, n$. The first feature $\mathbf{f}_i^{(1)}$ usually represents the original (raw) feature $\mathbf{x}_i$. On the other hand, we can also consider the feature sets as $\mathcal{F}_k^{(2)} = \{\mathbf{f}_i^{(k)}\}_{i=1}^n$ for $k = 1, 2, \ldots, m$, each of which contains the same features for different data vectors. Note that different feature extraction operations may generate the $\mathbf{f}_i^{(k)}$ with different dimensions, which needs to be taken care of by dimension normalization. The next table illustrates the above definition more clearly by a feature matrix. From different directions, row or column, we can represent the feature sets as $\{\mathcal{F}_i^{(1)}\}_{i=1}^n$ and $\{\mathcal{F}_k^{(2)}\}_{k=1}^m$ respectively. For convenience, we call this feature matrix as F-Matrix in the rest of the paper.

	$\mathcal{F}_1^{(1)}$	$\mathcal{F}_2^{(1)}$	$\cdots$	$\mathcal{F}_n^{(1)}$
$\mathcal{F}_1^{(2)}$	$\mathbf{f}_1^{(1)}$	$\mathbf{f}_2^{(1)}$	$\cdots$	$\mathbf{f}_n^{(1)}$
$\mathcal{F}_2^{(2)}$	$\mathbf{f}_1^{(2)}$	$\mathbf{f}_2^{(2)}$	$\cdots$	$\mathbf{f}_n^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathcal{F}_m^{(2)}$	$\mathbf{f}_1^{(m)}$	$\mathbf{f}_2^{(m)}$	$\cdots$	$\mathbf{f}_n^{(m)}$

We can see that the feature sets $\mathcal{F}_i^{(1)}$ with same labels are highly correlated from each other since they all represent the same class. In addition, all the $\mathcal{F}_k^{(2)}$ are highly correlated since they all represent the same data set. Hence, in the unsupervised case, the basic objective for multiple feature fusion is to find a general subspace matrix $\mathbf{P} \in \mathbb{R}^{D_0 \times d}$, where $d \leqslant D_0$ and D_0 denotes the normalized feature dimension, so that the pairwise canonical correlation of the feature sets $\mathcal{F}_k^{(2)}$ is maximized, while in the supervised case, the basic objective is to maximize the pairwise canonical correlation of both $\mathcal{F}_k^{(2)}$ and $\mathcal{F}_i^{(1)}$ with the same labels.

3. SUBSPACE LEARNING FOR MULTIPLE FEATURE FUSION

3.1 Feature Dimension Normalization

As we mentioned, the dimension of the feature may be different in all the $\mathcal{F}_i^{(1)}$. This may cause computational problem. So, before fusing the features, we first need to normalize the feature dimension throughout the given feature space. Since the feature dimensions in each $\mathcal{F}_k^{(2)}$ are usually identical, an effective way is to use Principal Component Analysis (PCA) [26] for dimensionality reduction on each $\mathcal{F}_k^{(2)}$ separately. The normalized dimension of the entire feature space can be chosen as the smallest dimension of all the full PCA subspaces for $\mathcal{F}_k^{(2)}$. Without losing generality, in the paper, we assume the dimension of $\mathcal{F}_k^{(2)}$ is already normalized to D_0.

3.2 Similarity Measure of Feature Sets

Following the existing work in [14, 13, 29], we use canonical correlation to measure the similarity of two feature sets. Suppose $\mathcal{F}_1$ and $\mathcal{F}_2$ represent two arbitrary feature sets, which can also be represented as feature matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ with feature vectors in the columns. Define two orthonormal basis matrices $\mathbf{P}_1 \in \mathbb{R}^{D_0 \times d_1}$ and $\mathbf{P}_2 \in \mathbb{R}^{D_0 \times d_2}$. So we have $\mathbf{F}_1\mathbf{F}_1^T = \mathbf{P}_1\mathbf{\Lambda}_1\mathbf{P}_1^T$ and $\mathbf{F}_2\mathbf{F}_2^T = \mathbf{P}_2\mathbf{\Lambda}_2\mathbf{P}_2^T$, where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ denote the diagonal matrices of corresponding eigenvalues. Choose the same dimension, denoted as d_0, of the two subspaces. The SVD of $\mathbf{P}_1^T\mathbf{P}_2 \in \mathbb{R}^{d_0 \times d_0}$ is $\mathbf{Q}_{12}\mathbf{\Lambda}_0\mathbf{Q}_{21}^T$, where $\mathbf{Q}_{12}^T\mathbf{Q}_{12} = \mathbf{Q}_{21}^T\mathbf{Q}_{21} = \mathbf{Q}_{12}\mathbf{Q}_{12}^T = \mathbf{Q}_{21}\mathbf{Q}_{21}^T$ and $\mathbf{\Lambda}_0$ denotes the diagonal matrix of singular values. According to the definition in [14], the similarity measure $S(\mathcal{F}_1, \mathcal{F}_2)$ of feature sets $\mathcal{F}_1$ and $\mathcal{F}_1$ is determined by the sum of canonical correlations, which can be written as

$$S(\mathcal{F}_1, \mathcal{F}_2) = \max\ \mathrm{Tr}(\mathbf{Q}_{12}^T\mathbf{P}_1^T\mathbf{P}_2\mathbf{Q}_{21}), \quad (1)$$

where Tr denotes the trace operation.

3.3 Multiple Feature Fusion

As we mentioned in the previous section, we want to learn a general subspace matrix $\mathbf{P}$ for multiple feature fusion. Before introducing the objective formulations, we need to first normalize the foregoing $\mathbf{P}_1$ and $\mathbf{P}_2$ to respectively make the columns of $\mathbf{P}^T\mathbf{P}_1$ and $\mathbf{P}^T\mathbf{P}_2$ orthonormal as suggested by [14]. This is handled by QR-decomposition. So, the normalized $\mathbf{P}_1$ and $\mathbf{P}_2$ are $\mathbf{P}_1\mathbf{R}_1^{-1}$ and $\mathbf{P}_2\mathbf{R}_2^{-1}$ respectively, where $\mathbf{R}_1 \in \mathbb{R}^{d_0 \times d_0}$ and $\mathbf{R}_2 \in \mathbb{R}^{d_0 \times d_0}$ are the invertible upper-triangular matrices from the QR-decompositions on $\mathbf{P}^T\mathbf{P}_1 =$ and $\mathbf{P}^T\mathbf{P}_2$. Without losing generality, we assume in the rest of the paper all the $\mathbf{P}_1$ and $\mathbf{P}_2$ are normalized.

3.3.1 Unsupervised Multiple Feature Fusion

For unsupervised case with m feature sets $\{\mathcal{F}_k^{(2)}\}_{k=1}^m$ and unknown labels $\mathcal{L}$, we only consider the row direction of the F-Matrix. So, the subspace matrix $\mathbf{P}$ is found by solving the optimization problem in Eq. (2)

$$\mathbf{P} = \arg\max_{\mathbf{P}} \mathcal{J}_1 = \arg\max_{\mathbf{P}} \sum_{k_1=1}^{m}\sum_{k_2=1}^{m} S_{\mathbf{P}}(\mathcal{F}_{k_1}^{(2)}, \mathcal{F}_{k_2}^{(2)}), \quad (2)$$

where

$$S_{\mathbf{P}}(\mathcal{F}_{k_1}^{(2)}, \mathcal{F}_{k_2}^{(2)}) = \max\ \mathrm{Tr}(\mathbf{Q}_{k_1k_2}^T\mathbf{P}_{k_1}^T\mathbf{P}\mathbf{P}^T\mathbf{P}_{k_2}\mathbf{Q}_{k_2k_1}). \quad (3)$$

3.3.2 Supervised Multiple Feature Fusion

For supervised case with $\{\mathcal{F}_k^{(2)}\}_{k=1}^m$ and known labels $\mathcal{L}$ for n feature sets $\{\mathcal{F}_i^{(1)}\}_{i=1}^n$, we consider both row and column directions of the F-Matrix. So, the subspace matrix $\mathbf{P}$ is found by solving the optimization problem in Eq. (4)

$$\begin{aligned}\mathbf{P} &= \arg\max_{\mathbf{P}}(\mathcal{J}_1 + \mathcal{J}_2) \\ &= \arg\max_{\mathbf{P}}(\mathcal{J}_1 + \sum_{i_1=1}^{n}\sum_{l_{i_2}=l_{i_1}} S_{\mathbf{P}}(\mathcal{F}_{i_1}^{(1)}, \mathcal{F}_{i_2}^{(1)}, \mathcal{L})), \quad (4)\end{aligned}$$

where

$$S_{\mathbf{P}}(\mathcal{F}_{i_1}^{(1)}, \mathcal{F}_{i_2}^{(1)}, \mathcal{L}) = \max\ \mathrm{Tr}(\mathbf{Q}_{i_1i_2}^T\mathbf{P}_{i_1}^T\mathbf{P}\mathbf{P}^T\mathbf{P}_{i_2}\mathbf{Q}_{i_2i_1}). \quad (5)$$

3.3.3 Iterative Learning

With simple rearrangement of the formulation, $\mathcal{J}_1$ and $\mathcal{J}_2$ can be rewritten as

$$\mathcal{J}_1 = \mathrm{Tr}(\mathbf{P}^T\mathbf{A}\mathbf{P}), \quad \mathcal{J}_2 = \mathrm{Tr}(\mathbf{P}^T\mathbf{B}\mathbf{P}), \quad (6)$$

where

$$\begin{aligned}\mathbf{A} &= \sum_{k_1=1}^{m}\sum_{k_2=1}^{m}(\mathbf{P}_{k_1}\mathbf{Q}_{k_1k_2} - \mathbf{P}_{k_2}\mathbf{Q}_{k_2k_1}) \\ &\quad (\mathbf{P}_{k_1}\mathbf{Q}_{k_1k_2} - \mathbf{P}_{k_2}\mathbf{Q}_{k_2k_1})^T \\ \mathbf{B} &= \sum_{i_1=1}^{n}\sum_{l_{i_2}=l_{i_1}}(\mathbf{P}_{i_1}\mathbf{Q}_{i_1i_2} - \mathbf{P}_{i_2}\mathbf{Q}_{i_2i_1}) \\ &\quad (\mathbf{P}_{i_1}\mathbf{Q}_{i_1i_2} - \mathbf{P}_{i_2}\mathbf{Q}_{i_2i_1})^T.\end{aligned}$$

The matrix $\mathbf{P} = [\mathbf{p}_1\mathbf{p}_2\ldots\mathbf{p}_d]$ is obtained by solving the eigen-decomposition problem

$$\mathbf{A}\mathbf{p} = \lambda\mathbf{p} \text{ or} \quad (7)$$

$$\mathbf{C}\mathbf{p} = \lambda\mathbf{p}, \text{ where } \mathbf{C} = \mathbf{A} + \mathbf{B}. \quad (8)$$

Here $\{\mathbf{p}_i\}_{r=1}^d$ are eigenvectors corresponding to the d largest eigenvalues.

The general algorithm for learning $\mathbf{P}$ is in an iterative manner [14, 15]. First, initialize the $\mathbf{P}$ by identity matrix $\mathbf{I} \in \mathbb{R}^{D_0 \times D_0}$. Second, for each iteration, normalize $\mathbf{P}_{k_1}$, $\mathbf{P}_{k_2}$ or $\mathbf{P}_{i_1}$, $\mathbf{P}_{i_2}$ using $\mathbf{R}_1$ or $\mathbf{R}_2$ obtained by QR-decomposition; find $\mathbf{Q}_{k_1k_2}$, $\mathbf{Q}_{k_2k_1}$ or $\mathbf{Q}_{i_1i_2}$, $\mathbf{Q}_{i_2i_1}$; calculate $\{\mathbf{p}_i\}_{r=1}^{D_0}$ through Eq. (2) or (4). Last, select top $\{\mathbf{p}_i\}_{r=1}^d$ to form $\mathbf{P}$. Empirically, it requires a small number of iterations to converge, such as 5 or 6.

3.4 Discussion

The existing work that most relates to our idea is called Discriminant-analysis of Canonical Correlations (DCC), proposed by Kim in [14]. Although both methods use canonical correlation to measure the similarity between data sets, the differences between our method and DCC are distinct. Basically, the two methods are designed for different application purposes. Our method is developed for multiple feature fusion, while DCC is developed for discriminant analysis cross different image sets. So, the data set structures and objective functions for subspace learning are all different. In addition, our method can be either unsupervised or supervised, while DCC can only be supervised. For the unsupervised pattern classification applications, our method could be flexibly combined with any dimensionality reduction methods

and classifiers. But, DCC has its own classification scheme which is fixed in an intrinsical way. More specifically by mathematical formulations, we can see that DCC follows the objective function in Eq. (9)

$$\mathbf{P}_{\mathrm{DCC}} = \arg\max_{\mathbf{P}} \frac{\mathcal{J}_2}{\mathcal{J}_3}, \tag{9}$$

where

$$\mathcal{J}_3 = \sum_{i_1=1}^{n} \sum_{l_{i_2} \neq l_{i_1}} S_{\mathbf{P}_{\mathrm{DCC}}}(\mathcal{F}_{i_1}^{(1)}, \mathcal{F}_{i_2}^{(1)}, \mathcal{L}). \tag{10}$$

The DCC objective only considers the column direction in the F-Matrix. It introduces the different class canonical correlation measure by $\mathcal{J}_3$ and has nothing to do with $\mathcal{J}_1$ introduced by our proposed method. It performs the subspace learning by Fisher criterion [31] for discriminating which is not for the feature fusion purpose. For our proposed methods, Fisher criterion doesnot necessarily to be considered.

4. TENSOR-BASED DISCRIMINATIVE SUBSPACE LEARNING ON MULTIPLE FEATURES

Several important problems of multiple feature fusion are caused by the computational difficulty introduced by the concatenation operation and the small sample size case (or curse of dimensionality). Existing feature fusion methods often concatenate different features of single datum to be a long vector. If the number of features is too large, the final fused feature vector may be too long to be handled by the limited machine computation. On the other hand, if the number of features is too small, the final fused feature vector may be statistically insufficient for capturing the intrinsic high-order feature structure. An effective solution is to consider the fused features of each single datum as a 2D tensor [15, 30, 6, 10].

Suppose the fused multiple features of a single datum $\mathbf{x}_i$ are stacked in a matrix $\mathbf{X}_{\mathcal{F}}^{(i)} = \mathbf{P}^T[\mathbf{f}_i^{(1)}\ \mathbf{f}_i^{(2)}\ \ldots\ \mathbf{f}_i^{(m)}] \in \mathbb{R}^{d\times m}$. All the $\{\mathbf{X}_{\mathcal{F}}^{(i)}\}_{i=1}^n$ form a tensor $\mathbf{X}_{\mathcal{F}} \in \mathbb{R}^{d\times m\times n}$. For the classification purpose, we have the label information $\mathcal{L}$ available in the training stage. Considering the Pearson correlation metric and $\mathbf{Y}_{\mathcal{F}}$ as the low-dimensional tensor representation, we have the objective function in Eq. (11) following our previous work in [10].

$$\begin{aligned}
&\varepsilon(\mathbf{U}_1, \mathbf{U}_2) \\
=\ & \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \langle \mathbf{Y}_{\mathcal{F}}^{(i_1)}, \mathbf{Y}_{\mathcal{F}}^{(i_1)} - \mathbf{Y}_{\mathcal{F}}^{(i_2)} \rangle \cdot (w_{i_1 i_2}^{(\mathrm{d})} - w_{i_1 i_2}^{(\mathrm{s})}) \\
=\ & \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \langle \mathbf{X}_{\mathcal{F}}^{(i_1)} \times_q \mathbf{U}_q|_{q=1}^2, \mathbf{X}_{\mathcal{F}}^{(i_1)} \times_q \mathbf{U}_q|_{q=1}^2 \\
& -\mathbf{X}_{\mathcal{F}}^{(i_2)} \times_q \mathbf{U}_q|_{q=1}^2 \rangle \cdot (w_{i_1 i_2}^{(\mathrm{d})} - w_{i_1 i_2}^{(\mathrm{s})}),
\end{aligned} \tag{11}$$

where $w_{i_1 i_2}^{(\mathrm{d})}$ and $w_{i_1 i_2}^{(\mathrm{s})}$ are different-class and same-class weights defined in [11] and $\mathbf{U}_1$, $\mathbf{U}_2$ are two subspaces for row and column directions respectively. Note here $w_{i_1 i_2}^{(\mathrm{d})}$ and $w_{i_1 i_2}^{(\mathrm{s})}$ are determined by correlation metric based neighborhood relation in the sample space. The two subspaces are found by solving Eq. (12)

$$\arg\max_{\mathbf{U}_1, \mathbf{U}_2} \varepsilon(\mathbf{U}_1, \mathbf{U}_2). \tag{12}$$

The above optimization problem is not in closed-form solution. An iterative calculation is derived. We consider the idea to first arbitrarily initialize $\mathbf{U}_1$ and $\mathbf{U}_2$; then assume $\mathbf{U}_1$ is known so that $\mathbf{U}_2$ is solved by fixing the other. One way to compute $\mathbf{U}_q$ is to solve the generalized eigenvalue decomposition problem in Eq. (13), which is equivalent to the optimization problem in Eq. (11).

$$\mathbf{Z}^{(q)}(\mathbf{D}_{\mathrm{d}} - \mathbf{W}_{\mathrm{d}})\mathbf{Z}^{(q)T}\mathbf{U}_q = \lambda \mathbf{Z}^{(q)}(\mathbf{D}_{\mathrm{s}} - \mathbf{W}_{\mathrm{s}})\mathbf{Z}^{(q)T}\mathbf{U}_q, \tag{13}$$

where $\mathbf{Z}^{(q)}$ denotes the q-mode unfolding [30] of tensor $\mathbf{X}_{\mathcal{F}}$ into a matrix, the weight matrices $\mathbf{W}_{\mathrm{s}}$ and $\mathbf{W}_{\mathrm{d}}$ are formed by filling $w_{i_1 i_2}^{(\mathrm{d})}$ or $w_{i_1 i_2}^{(\mathrm{s})}$ respectively, and $\mathbf{D}_{\mathrm{d}}$ and $\mathbf{D}_{\mathrm{s}}$ are diagonal with $\mathbf{D}_{\mathrm{d}}[i_1, i_1] = \sum_{i_2} w_{i_1 i_2}^{(\mathrm{d})}$ and $\mathbf{D}_{\mathrm{s}}[i_1, i_1] = \sum_{i_2} w_{i_1 i_2}^{(\mathrm{s})}$. More details can be retrieved from [30] and [10].

Figure 1: Patch-based features for robust appearance modeling. Top row: Face images with appearance changes. Second row: Partition images into patches for more robust modeling.

5. FEATURE EXTRACTION

In this paper, we are particularly interested in image-based feature fusion, especially for face images. Raw image pixels are intuitive features, however, are not robust subject to the lighting conditions, occlusions, and other small changes. Take Figure 1 as an example. Although the two images are of the same person, the raw gray values differ significantly due to the occlusion effect of glasses. However, if we treat the images as sets of patch features, we can still observe satisfying similarities over most of the patches, even with influences of the glasses or small pose changes.

For multiple feature fusion, we combine patch-based features with the raw feature. We use grid sampling to partition an image into patches. The descriptors of these patches are then concatenated to form a global description of the image, which retains important spatial information and still keeps capturing small pose changes. Figure 1 shows two examples of dividing facial images into rectangular regions. To describe each patch's appearance, we use two different kinds of local features, Histogram of Oriented Gradient descriptor (HOG) [7] and Local Binary Pattern (LBP) [18, 2], as discussed in the following subsections.

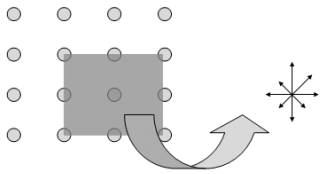

Figure 2: HoG features sampled on overlapping grids.

5.1 HOG Feature

Our first patch feature is inspired by the recent progress in image/video based human detection [7, 35] and general object recognition (such as cars, buses, bicycles and animals) [8, 5]. As reported in [7], local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. This so-called HoG feature obtains the state-of-the-art performance for human detection and object recognition. We believe that HoG feature is also useful for face recognition task, since it can model the local appearance well and tolerates small pose and appearance changes.

In our approach, we combine the idea of spatial sampling with orientation sampling, by concatenating the local features for each patch. To model each patch, we compute the histogram of oriented gradients, which represent edges with a magnitude-weighted histogram, grouped according to edge directions. Each patch is represented by an HoG feature, as a vector of length 8 describing the gradient in 8 orientations. Unlike the work in [7], we do not normalize each patch, which makes it possible to employ the technique of "integral image" [27] for fast computation. Note that HoG is designed for pedestrian detection [7, 35], but we will first adopt it for face recognition in this paper. Figure 2 shows our approach, where each circle denotes one sampling point, and the grid rectangle denotes one patch. Note that our sampled patches are 50% overlapping, which is supposed to overcome the boundary effects.

5.2 LBP Feature

The LBP operator is derived from a general definition of texture in a local neighborhood, which provides an appearance measure invariant against monotonic gray level changes. LBP is one of the best performing texture descriptors [18]. According to the definition, LBP operator assigns a label to every pixel of an image by comparing the neighborhood of each pixel with the center pixel value and considering the result as a binary number. Then the histogram of the labels are accumulated as a local descriptor. Such procedure is computationally efficient and simple to perform, which makes it attractive for many real-world applications on image/video processing.

By dividing the face image into patches, we can view the face as a composition of micro-patterns, which are described by uniform LBP for each patches. We can obtain another kind of local descriptors by applying the techniques of texture analysis.

Note that LBP has been applied successfully to face recognition [1, 2, 33, 25]. In this paper, we propose to fuse the LBP and HoG together with the raw feature to obtain a more effective description of the face images for the face recognition task.

Table 1: Description of some acronyms and abbreviations.

Method	Description
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
Tensor	Tensor-based subspace learning
UnSL	Unsupervised Subspace Learning
SuSL	Supervised Subspace Leraning
Raw	Raw image feature
HoG	Histogram of Oriented Gradient feature
LBP	Local Binary Pattern feature
Fusion	Fusion of the three features
EuNN	Euclidean-distance Nearest Neighbor
CorrNN	Correlation-distance Nearest Neighbor

6. EXPERIMENTS

Annotating faces from video or images is an important real-world application for image/video retrieval. Face recognition is a basic module of such kind of systems that can conduct this application. So, in the experiments, we demonstrate our proposed multiple feature fusion methods by face recognition evaluations on three benchmark databases. We compare several different kinds of methods in the evaluations. Table 1 shows the description of some acronyms and abbreviations representing those methods.

6.1 Data Sets

The Face Recognition Grand Challenge (FRGC) Ver1.0 [19] database contains 5000+ frontal face images of 275 subjects. The primary goal of the FRGC is to promote and advance face recognition technology designed to support existing face recognition efforts in the U.S. Government. Ver1.0 is designed to introduce participants to the FRGC challenge problem format and its supporting infrastructure. Face images are manually aligned, cropped out from the selected images and resized to be 32×32, with 256 gray levels per pixel. We randomly choose 10 images of each subject for training and 10 different images per subject for test (or the remaining images, when the subject class has less than 20 images).

The CMU PIE [22] database contains in total 41368 images of 68 subjects with 500+ images for each. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination, and expression. For each subject, we manually select 168 images from five near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. Face images are manually aligned, cropped out from the selected images and resized to be 32×32, with 256 gray levels per pixel. We randomly choose 40 images per individual to form a sub-database. There are 2720 images in total. A random database partition is performed with 20 images per individual for training, and the rest of the database for test.

The Yale Face Database B [12] contains 5760 single light source images of 10 subjects, each under 576 viewing conditions (9 poses $\times$ 64 illumination conditions). The extended

Table 2: Single feature vs. multiple feature fusion.

	FRGC Ver1.0		PIE		Yale-B	
Feature	Accuracy	Dim.	Accuracy	Dim.	Accuracy	Dim.
Raw+EuNN	73.9%	240	58.8%	400	70.3%	400
HoG+EuNN	76.4%	380	63.7%	380	74.6%	500
LBP+EuNN	76.1%	500	72.1%	500	53.2%	400
Fusion+EuNN	79.4%	440	69.6%	470	71.1%	700
Raw+CorrNN	73.9%	280	58.8%	400	70.4%	480
HoG+CorrNN	76.4%	380	63.7%	380	75.0%	480
LBP+CorrNN	76.1%	500	72.0%	380	53.2%	300
Fusion+CorrNN	79.3%	380	69.6%	500	70.9%	540

Table 3: Multiple feature fusion by subspace learning.

	PIE		Yale-B	
Method	Accuracy	Dim.	Accuracy	Dim.
PCA+EuNN	69.6%	470	71.1%	700
PCA+CorrNN	69.6%	500	70.9%	540
UnSL+EuNN	71.2%	340	71.8%	310
UnSL+CorrNN	71.3%	350	71.6%	310
SuSL+EuNN	71.5%	400	85.4%	300
SuSL+CorrNN	93.6%	390	94.3%	390

Yale Face Database B [16] contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. The data format of the two databases are the same. We combine the two databases and form a new one, called Yale-B database in the paper, including 38 subjects' face images. For this database, the images are cropped and resized to 32×32, with 256 gray levels per pixel. We randomly choose 40 images per individual to form a sub-database. There are 1520 images in total. A random database partition is performed with 20 images per individual for training, and the rest of the database for test.

6.2 Face Recognition

Before performing face recognition experiments, we normalize each feature dimension to be 500. So, all the feature vectors have the same length after the dimension normalization. All the results reported as follows are from best tuning the parameters of all the corresponding methods.

6.2.1 Single Feature vs. Multiple Feature Fusion

In this experiment, we want to demonstrate the advantage of multiple feature fusion over any single feature. The three features, raw, HoG, and LBP, are all used separately to combine with Euclidean distance nearest neighbor classifier and correlation distance nearest neighbor classifier for the face recognition tests on the three databases. The simple fusion of the three features using concatenation is also used to combine with the two different classifiers respectively on the tests. Table 2 summarizes the results of the experiments. We can see that the feature fusion performs better than any single feature on the FRGC Ver1.0 database. On the PIE database, since LBP dominates the performance, the feature fusion performs much better than the other two single features. But, it performs a little worse than LBP feature because the raw feature performs much worse than the other two. On the Yale-B database, since HoG dominates the performance, the feature fusion performs better than the other two single features. But, it performs a little worse than HoG feature because the LBP feature performs much worse than the other two. This result is reasonably acceptable since we do not have any theoretical conclusions (prior knowledge) that can tell us for which database to use what kinds of features for the best results. In other words, the fused feature has much robustness in real-world applications if we have no prior knowledge to select the best feature beforehand. Note here the "Dim." means the subspace dimension corresponding to the best result. We calculate the face recognition accuracy by sampling each 10-dimension interval.

6.2.2 Multiple Feature Fusion by Subspace Learning

To evaluate the performance of our proposed multiple feature fusion methods, we compare both unsupervised and supervised subspace learning methods with the simple feature fusion method by concatenation. For the simple feature fusion case, we perform PCA to reduce the redundant dimensionality. Euclidean distance nearest neighbor classifier and correlation distance nearest neighbor classifier are still used to combine with those methods for the face recognition tests on the PIE and Yale-B databases. Table 3 summarizes the results of the experiments. We can see that our proposed multiple feature fusion methods significantly outperform the feature fusion method by simple concatenation in both databases. Especially the supervised subspace learning plus correlation distance nearest neighbor classifier performs the best. This result is consistent with our previous work [11, 10], in which correlation metric significantly boosts the discriminating power of the classifier for face recognition. Again, we calculate the face recognition accuracy by sampling each 10-dimension interval.

6.2.3 Tensor-based Subspace Learning

In this experiment, tensor-based subspace learning algorithm is used to combine with our proposed multiple feature fusion methods. Since the goal of tensor-based discriminative analysis here is to capture high-order feature structure and deal with the computational difficulty when too many features are fused, the multiple feature fusion scheme followed by tensor-based subspace learning doesnot necessarily outperform the multiple feature fusion scheme itself. To be fair in the comparison, we also combine Linear Discriminant Analysis (LDA) [4] with simple feature fusion by concatena-

Table 4: Combine multiple feature fusion with tensor-based discriminant analysis.

	PIE		Yale-B	
Method	Accuracy	Dim.	Accuracy	Dim.
PCA+LDA	92.8%	65	94.1%	35
UnSL+LDA	93.0%	65	96.6%	30
SuSL+LDA	92.8%	65	97.2%	35
UnSL+Tensor	94.7%	70×3	97.4%	34×3
SuSL+Tensor	94.3%	70×3	97.5%	37×3

tion and our proposed unsupervised and supervised subspace learning methods for discriminative feature extraction. For the simple feature fusion case, we perform PCA to reduce the redundant dimensionality. Euclidean distance nearest neighbor classifier is used to combine with those methods for the face recognition tests on the PIE and Yale-B database. Table 4 summarizes the results of the experiments. We can see that the proposed subspace learning methods improve the face recognition performance by combining with the tensor-based subspace learning. We calculate the face recognition accuracy by sampling each 5-dimension interval for the non-tensor-based methods. For the tensor-based cases, we fix the second mode (3-D) of the feature tensor (since the dimension is already small) and calculate the face recognition accuracy by brute-force search of each dimension in the first mode.

6.3 Discussion

From the experimental results, we can see that simply concatenating different features may improve the robustness of face recognition performance. But, the feature fusion may also degenerate the performance due to the unbalance among the individual features. The proposed method learns a generalized subspace in which the low-dimensional representations of those individual features have a better balance to contribute to the improved performance by fusion. Here, the low-dimensional representations are learned in a linear way. A non-linear learning strategy is also feasible to extend if we assume the correlations among different features tend to be more complicated.

The tensor-based subspace learning algorithm that concatenates with the feature fusion is use to reduce the computational cost when the number of different features is large. In the training stage, training the tensor model requires extra computational cost than the single linear model, but the tensor structure, for high-order feature patterns, may introduce more powerful properties to represent the fused feature to boost the discriminating power. On the other hand, it may also alleviate the curse-of-dimensionality dilemma and the small sample size problem.

7. CONCLUSIONS

We have presented in the paper on how to fuse multiple features in a generalized subspace learning framework. The basic idea is to find a linear subspace in which the cumulative canonical correlation between any pair of feature sets is maximized. The proposed algorithm can be used for both supervised or unsupervised feature fusion task. Any dimensionality reduction and classifiers can also be flexibly concatenated with the feature fusion scheme for the pattern classification purpose. To further deal with the computational difficulty and avoid the curse-of-dimensionality dilemma or the small sample size problem, tensor-based discriminative subspace learning method is introduced to reduce dimensionality and extract discriminative features with high-order structures from the feature fusion. We performed extensive experiments on face recognition with multiple visual feature fusion, which have demonstrated the effectiveness and robustness of the proposed methods. It will be interesting to use more different kinds of features in the fusion for an extended work. We also plan to apply the proposed multiple feature fusion methods to image/video retrieval applications in the future. Especially we want to demonstrate the power of our proposed method in dealing with semantic video retrieval and concept annotation, which need more general features embodying more complicated inter-correlations. The extension of the presented subspace learning method will also be explored.

8. ACKNOWLEDGMENTS

This research was funded in part by the Beckman Graduate Fellowship, in part by the U.S. Government VACE program, and in part by the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

9. REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. on PAMI*, 28(12):2037–2041, 2006.

[3] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

[4] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. on PAMI*, 19(7):711–720, 1997.

[5] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proc. of ACM CIVR*, pages 401–408, 2007.

[6] H.-T. Chen, H.-W. Chang, and T.-L. Liu. Local discriminant embedding and its variants. In *IEEE Conf. on CVPR*, volume 2, pages 846–853, 2005.

[7] N. Dalal and B. Triggs. Histograms of oriented gradi-ents for human detection. In *IEEE Conf. on CVPR*, pages 886–893, 2005.

[8] N. Dalal and B. Triggs. Object detection using histograms of oriented gradients. In *European Conference on Computer Vision, Workshop on Pascal VOC'06*, 2006.

[9] Y. Fang, T. Tan, and Y. Wang. Fusion of global and local features for face verification. In *IEEE Conf. on ICPR*, pages 382–385, 2002.

[10] Y. Fu and T. S. Huang. Image classification using correlation tensor analysis. *IEEE Trans. on Image Processing*, 17(2):226–234, 2008.

[11] Y. Fu, M. Liu, and T. Huang. Conformal embedding analysis with local graph modeling on the unit hypersphere. In *IEEE Conf. on CVPR, workshop on Component Analysis*, 2007.

[12] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. on PAMI*, 23(6):643–660, 2001.

[13] T.-K. Kim, O. Arandjelovic, and R. Cipolla. Learning over sets using boosted manifold principal angles (bompa). In *British Machine Vision Conference*, pages 779–788, 2005.

[14] T.-K. Kim, J. Kittler, and R. Cippola. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. on PAMI*, 29(56):1005–1018, 2007.

[15] T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *IEEE Conf. on CVPR*, 2007.

[16] K.-C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. on PAMI*, 27(5):684–698, 2005.

[17] M. Liu, Y. Fu, and T. S. Huang. An audio-visual fusion framework with joint dimensionality reduction. In *IEEE Conf. on ICASSP*, 2008.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on PAMI*, 24(7):971–987, 2002.

[19] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conf. on CVPR*, pages 947–954, 2005.

[20] K. S. Rao and A. N. Rajagopalan. A probabilistic fusion methodology for face recognition. *EURASIP Journal on Applied Signal Processing*, 2005(17):2772–2787, 2005.

[21] M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. on Multimedia*, 9(7):1396–1403, 2007.

[22] T. Sim, S. Baker, and M. Bsat. The cmu pose, illuminlation, and expression database. *IEEE Trans. on PAMI*, 25(12):1615–1618, 2003.

[23] Q.-S. Sun, Z. Jin, P.-A. Heng, and D.-S. Xia. A novel feature fusion method based on partial least squares regression. *Lecture Notes in Computer Science 3686*, 3686/2005:268–277, 2005.

[24] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia. A new method of feature fusion and its application in image recognition. *Pattern Recognition*, 38(12):2437–2448, 2005.

[25] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *IEEE Conf. on AMFG*, pages 168–182, 2007.

[26] M. Turk and A. Pentland. Face recognition using eigenfaces. In *IEEE Conf. on CVPR*, pages 586–591, 1991.

[27] P. Viola and M. Jones. Robust real-time face detection. *Int'l Journal of Computer Vision*, 57(2):137–154, 2004.

[28] X. Wang and X. Tang. Using random subspace to combine multiple features for face recognition. In *IEEE Conf. on FGR*, pages 284–289, 2004.

[29] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4:913–931, 2003.

[30] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang. Discriminant analysis with tensor representation. In *IEEE Conf. on CVPR*, pages 526–532, 2005.

[31] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. on PAMI*, 29(1):40–51, 2007.

[32] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu. Feature fusion: parallel strategy vs. serial strategy. *Pattern Recognition*, 36(6):1369–1381, 2003.

[33] J. Zhao, H. Wang, H. Ren, and S.-C. Kee. Lbp discriminant analysis for face verification. In *IEEE Conf. on CVPR*, pages 167–167, 2005.

[34] X. Zhou and B. Bhanu. Feature fusion of face and gait for human recognition at a distance in video. In *IEEE Conf. on ICPR*, pages 529–532, Washington, DC, USA, 2006.

[35] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Conf. on CVPR*, pages 886–893, 2005.