Generalized Multiview Analysis: A Discriminative Latent Space

Abhishek Sharma[†] Abhishek Kumar Hal Daume III
David W. Jacobs
Institute for Advance Computer Studies, University of Maryland, USA
[bhokaal[†],abhishek,djacobs]@cs.umd.edu hal@umiacs.umd.edu

Abstract

This paper presents a general multi-view feature extraction approach that we call Generalized Multiview Analysis or GMA. GMA has all the desirable properties required for cross-view classification and retrieval: it is supervised, it allows generalization to unseen classes, it is multi-view and kernelizable, it affords an efficient eigenvalue based solution and is applicable to any domain. GMA exploits the fact that most popular supervised and unsupervised feature extraction techniques are the solution of a special form of a quadratic constrained quadratic program (QCQP), which can be solved efficiently as a generalized eigenvalue problem. GMA solves a joint, relaxed QCQP over different feature spaces to obtain a single (non)linear subspace. Intuitively, GMA is a supervised extension of Canonical Correlational Analysis (CCA), which is useful for cross-view classification and retrieval. The proposed approach is general and has the potential to replace CCA whenever classification or retrieval is the purpose and label information is available. We outperform previous approaches for text-image retrieval on Pascal and Wiki text-image data. We report state-of-the-art results for pose and lighting invariant face recognition on the MultiPIE face dataset, significantly outperforming other approaches.

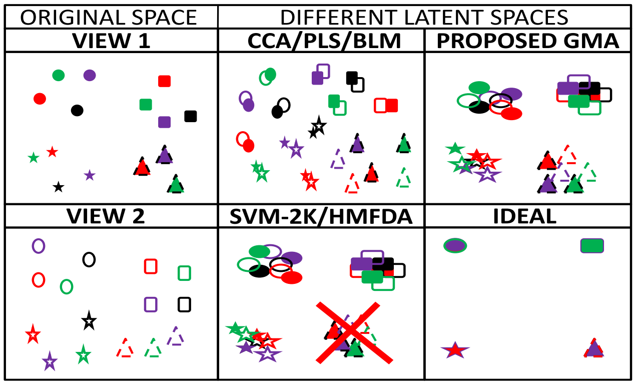

Figure 1. A simple pictorial demonstration of various multi-view approaches along with the proposed GMA and an ideal approach. Shapes represent classes, the same color and shape indicates paired samples in different views, dashed outline shapes (triangles) are the unseen classes (not used in training). Ideally, we would like different classes (seen and unseen) to be well separated while all the same-class samples collapse to a point. Unsupervised approaches like CCA, PLS and BLM try to unite paired samples only. Supervised approaches, like SVM-2K and HMFDA unite same-class samples and separate different classes but they cannot generalize to unseen classes. Our proposed GMA approach unites same class samples, separates different classes and generalizes to unseen classes. (Figure best viewed in color)

1. Introduction

Data often arrives in multiple *views or styles*. These different views may represent the same underlying *content*. For example, user tags or textual descriptions and image features (views) indicate the class of objects (content) contained in the image; face images of a person in different poses (views) and lighting conditions reveal the identity (content) and so on. In some applications (e.g., face recognition, multilingual or cross-media retrieval), we are interested in performing classification and retrieval where the gallery and query data belongs to different views. This is difficult because it is not a priori meaningful to directly compare the instances across different views since they span different feature spaces. Formally, given a trained model $\mathcal{M}$, database $\mathcal{D}_v$ in view v and query q_u in view u, *cross-view classification* refers to obtaining the label of q_u using a k-NN classification scheme from $\mathcal{D}_v$ (pose-invariant face recognition) and *cross-view retrieval* refers to retrieving samples from $\mathcal{D}_v$ that are closest to q_u (text-image retrieval). *Unseen class* refers to the class that is not used in obtaining $\mathcal{M}$. The use of a k-NN scheme makes it possible to classify q_u even if it belongs to an unseen class.

One popular solution is to learn view-specific projection directions using *paired* samples from different views to project samples from different views into a common latent space followed by classification/retrieval. Paired samples refer to samples in different views that are known to come from the same object, e.g. image features and associated tags for an image, face images of a person in two different poses. Successful cross-view classification/retrieval requires that samples from the same content are united and those from different content are separated in the common subspace, see Fig1.

Popular unsupervised approaches to learn such directions are Canonical Correlational Analysis (CCA) [19, 8], Bilinear Model (BLM) [22] and Partial Least Squares (PLS) [16, 19, 18]. Specifically, CCA has been the workhorse for learning a common latent space which is evident from its wide-spread use in vision [19, 17, 18], cross-lingual retrieval[8], cross-media retrieval [12, 15], etc.... These citations are just the few we had space to include. Unfortunately, the above mentioned approaches only care about pair-wise closeness in the common subspace so they are not well suited for classification/retrieval. Especially, when within-class variance is large, these methods are bound to perform poorly for classification/retrieval because classification and retrieval both require that within-class samples are united. Moreover, the costly label information that might be available during training is unharnessed. Locality preserving CCA (LPCCA) was introduced to capture the non-linearity present in the data by forcing nearby points in the original feature space to be close in the latent space as well [21]. However, they did not use the label information and we will see that it is a special instance of our general model. Discriminative CCA (DCCA) uses multi-dimensional labels as the second view, which is just single view scenario with multidimensional labels [20]. CCA is used to match sets of images by maximizing within-set correlation and minimizing between set correlation, which is again a single view scenario with set membership information [13]. We are interested in scenarios in which the data has two different views, along with label information.

A number of supervised approaches to multi-view analysis have also been proposed. Multi-view Fisher Discriminant Analysis (MFDA) learns classifiers in different views by maximizing the agreement between the predicted labels of these classifiers[4]. But, MFDA can only be used for two-class problems. To cope with this, [3] extended MFDA to a multi-class scenario using a Hierarchical clustering approach. In [6], the authors obtained a multi-view version of SVM by constraining the one-dimensional outputs of individual SVM's to be equal. These approaches however, use multi-view data to learn classifiers in each view that are better than the classifiers learned using single-view data only. With some non-trivial adaptation they can be used for cross-view classification and retrieval, but originally the authors have used them as single-view classifiers trained with multi-view data. The prime objective of this paper is cross-view classification and retrieval. Most importantly, none of MFDA, SVM-2K or HMFDA can classify samples from unseen classes, which is required in many real-world applications such as face recognition, cross-view retrieval and domain adaptation. For example, practical face recognition often requires a classifier that can compare images of unseen subjects (not used in training) at testing time, while cross-view retrieval also requires retrieval of unseen categories.

Table 1. Properties of popular approaches for classification and feature extraction. Note that only the proposed GMA approach has all the required properties. **S**: **S**upervised, **G**: **G**eneralizable, **MV**: **M**ulti-**V**iew, **E**: Efficient, **K**: **K**ernelizable, **DI**: **D**omain-**I**ndependent ('✓' indicates presence of property).

Method	**Properties**					
	S	**G**	**MV**	**E**	**K**	**DI**
PCA [23]		✓		✓	✓	✓
LDA [1]	✓	✓		✓	✓	✓
MFA [25]	✓	✓		✓	✓	✓
LPP [10]	✓	✓		✓	✓	✓
BLM [22]		✓	✓	✓	✓	✓
CCA [19]		✓	✓	✓	✓	✓
PLS [19, 16]		✓	✓	✓	✓	✓
SVM-2K [6]	✓			✓	✓	✓
MFDA [4]	✓			✓	✓	✓
HMFDA [3]	✓			✓	✓	✓
LPCCA [21]		✓	✓	✓	✓	✓
DCCA [20]	✓	✓		✓	✓	✓
SetCCA [13]	✓			✓	✓	✓
GMA	✓	✓	✓	✓	✓	✓

Finally, some *domain-specific* approaches use domain information to learn discriminative cross-view classifiers. Lighting invariant features are used in [14]. Synthetic virtual images in new pose and lighting conditions are used to train LDA for pose and lighting invariant face recognition in [17]. Geometry assisted hashing is used to counter pose and lighting change in [26]. Use of logistic regression with topic modeling features to obtain semantically meaningful features is used in [15] to extract text and image features for cross-media retrieval. Unfortunately, it might not work for unseen classes or when topic modeling is not effective e.g. face recognition. These approaches are customized to a particular task and such domain information may not be available in general.

Based on the above discussion we conclude that an ideal cross-view classification approach *must* be

- **Supervised(S)**: Use label information for class based discrimination.
- **Generalizable (G)**: Able to analyze new classes that are not used during training.
- **Multi-view (MV)**: Applicable to cross-view classification and retrieval, rather than just using multi-view

data for learning.

- **Efficient (E)**: Have an efficiently computed optimal solution.
- **Kernelizable (K)**: Have a kernel extension to model non-linearities.
- **Domain-Independent(DI)**: Applicable to general problems.

There are numerous feature extraction and classification techniques proposed so far, but none of them satisfies all the above mentioned requirements, see Table 1.

We approach the problem of cross-view classification by learning a common discriminative subspace and propose Generalized Multiview Analysis or GMA, with all the properties we have mentioned. We show that CCA, BLM and PLS are specific instances of our generic framework. Additionally, GMA can be used to extend a broad class of feature extraction techniques (*supervised and unsupervised*), including PCA [23], Linear Discriminant Analysis (LDA) [1], Locality Preserving Projections (LPP) [10], Neighborhood Preserving Embedding (NPE)[9] and Marginal Fisher Analysis (MFA) [25] into their multiview counterparts. The formulation involves solving a generalized eigenvalue problem, which leads to the globally optimal solution. For example, an extension of Linear Discriminant Analysis (LDA+GMA = GMLDA) will find a set of projection directions in each view that try to separate different contents' class means and unite different views of the same class in the common subspace.

Our generic GMA approach produces state-of-the-art results, outperforming several generic and domain-specific approaches for simultaneous pose and lighting invariant face recognition on the MultiPIE face dataset. We also report similar to state-of-the-art results on text-image retrieval on Wiki text-image data [15]. The paper is organized as follows - Section 2 presents the proposed approach, Section 3 describes experiments, and Section 4 presents conclusion and discussion.

2. Proposed Approach

Our approach is motivated by the fact that popular supervised and unsupervised feature extraction techniques can be cast as a special form of a quadratically constrained quadratic program (QCQP). Specifically, the optimal projection direction $\hat{\mathbf{v}}$ can be obtained as

$$\begin{aligned} \hat{\mathbf{v}} &= \underset{\mathbf{v}\neq 0}{\operatorname{argmax}}\, \mathbf{v}^T A \mathbf{v} \\ &s.t.\ \mathbf{v}^T B \mathbf{v} = 1 \ or\ \mathbf{v}^T \mathbf{v} = 1 \end{aligned} \quad (1)$$

Here, A is some symmetric square matrix and B is a square symmetric Definite Matrix i.e. no eigenvalue of B is equal to 0. Methods that fit this equation include PCA [23, 25], LDA [1, 25], LPP [10, 25], CCA, and MFA [25]. So, we first extend Eqn1 to a multi-view scenario and then use it with different (A, B) combinations to obtain different common subspaces with desired properties. For ease of understanding, we derive the results for two views and later extend it to multiple views.

Throughout this paper, superscripts are used for indexing and subscripts denote views. Vectors are denoted as straight bold lowercase ($\mathbf{x}$), variables/constants as lowercase italic (a) and matrices as capital italic (A). Hence, a sample in view p belonging to class i is denoted as $\mathbf{x}_p^i$ and a matrix of samples in view p as X_p.

2.1. Generalized Multiview Analysis

We now present a generalization of this framework to a multi-view setting. We first extend Eqn1 to a multi-view setting in Eqn2, combining two optimization problems without yet coupling them. Then, in Eqn6 we constrain the samples from the same content to project to similar locations in the latent space.

A joint optimization of two objective functions over two different vector spaces can be written as

$$\begin{aligned} [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] &= \underset{\mathbf{v}_1,\mathbf{v}_2}{\operatorname{argmax}}\ \mathbf{v}_1^T A_1 \mathbf{v}_1 + \mu \mathbf{v}_2^T A_2 \mathbf{v}_2 \\ &s.t.\ \mathbf{v}_1^T B_1 \mathbf{v}_1 = \mathbf{v}_2^T B_2 \mathbf{v}_2 = 1 \end{aligned} \quad (2)$$

The positive term μ is to bring a balance between the two objectives, because if $\max \mathbf{v}_1^T A_1 \mathbf{v}_1 \gg \max \mathbf{v}_2^T A_2 \mathbf{v}_2$, the joint objective will be biased towards optimizing $\mathbf{v}_1$ and vice-versa. Unfortunately, both the constraints are non-linear and there is no closed form solution in the current form. So, we couple the constraints with $\gamma = \frac{tr(B_1)}{tr(B_2)}$ to obtain a relaxed version of the problem with a single constraint as

$$\begin{aligned} [\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] &= \underset{\mathbf{v}_1,\mathbf{v}_2}{\operatorname{argmax}}\ \mathbf{v}_1^T A_1 \mathbf{v}_1 + \mu \mathbf{v}_2^T A_2 \mathbf{v}_2 \\ &s.t.\ \mathbf{v}_1^T B_1 \mathbf{v}_1 + \gamma \mathbf{v}_2^T B_2 \mathbf{v}_2 = 1 \end{aligned} \quad (3)$$

When $\hat{\mathbf{v}}_1^T B_1 \hat{\mathbf{v}}_1 = \hat{\mathbf{v}}_2^T B_2 \hat{\mathbf{v}}_2$, the constraints in Eqn2 and Eqn3 are equivalent. When $\hat{\mathbf{v}}_1^T B_1 \hat{\mathbf{v}}_1 \neq \hat{\mathbf{v}}_2^T B_2 \hat{\mathbf{v}}_2$, the constraint in Eqn3 is an approximation of the constraints in Eqn2. We empirically observed that parameter γ did not have much effect on overall performance.

Intuitively, the resulting problem in Eqn. 3 is solving the relaxed version of the original optimization problem in two different vector spaces (views). To facilitate understanding, let's consider a multi-view extension of LDA. In this case, $A_i = S_{bi}$, $B_i = S_{wi}$ for $i = 1, 2$ where S_{bi} and S_{wi} are between and within class scatter matrices and $\mathbf{v_1}$ and $\mathbf{v_2}$ are the projection directions in view 1 and 2 respectively. Eqn. 3 is jointly solving for LDA projection directions $\hat{\mathbf{v}}_\mathbf{1}$ and $\hat{\mathbf{v}}_\mathbf{2}$ to maximize between class separation and minimize within class variation in each view.

Now we introduce a constraint to couple these projection directions. For cross-view classification we require that the projections (a_1^i and a_2^i) of the exemplars (z_1^i and z_2^i) of the i^{th} content in different views should be close to each other in the projected latent space. a_1^i and a_2^i are defined as

$$a_1^i = \mathbf{v}_1^T \mathbf{z}_1^i \qquad and \qquad a_2^i = \mathbf{v}_2^T \mathbf{z}_2^i \tag{4}$$

We chose to maximize covariance between the exemplars from different views to obtain directions to achieve closeness between multi-view samples of the same class. This leads to a closed form solution and better preserves the between class variation as argued in [18]

$$[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] = \underset{\mathbf{v}_1,\mathbf{v}_2}{\operatorname{argmax}} \, \mathbf{v}_1^T Z_1 Z_2^T \mathbf{v}_2 \tag{5}$$

Here, Z_i's are the matrices constructed such that i^{th} column in both Z_1 and Z_2 contains exemplars corresponding to the same content. The exemplars can be chosen to suit the problem and feature extraction techniques. For instance, LDA represents a class as the mean of class samples, so class mean can be used as the exemplar.

Without any constraints on $\mathbf{v}_1$ and $\mathbf{v}_2$ the objective in Eqn5 can be increased indefinitely. But we couple this objective with the constrained objective of Eqn3 to get the final constrained objective

$$\begin{aligned}[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2] = \underset{\mathbf{v}_1,\mathbf{v}_2}{\operatorname{argmax}} \; & \mathbf{v}_1^T A_1 \mathbf{v}_1 + \mu \mathbf{v}_2^T A_2 \mathbf{v}_2 + 2\alpha \mathbf{v}_1^T Z_1 Z_2^T \mathbf{v}_2 \\ & s.t. \; \mathbf{v}_1^T B_1 \mathbf{v}_1 + \gamma \mathbf{v}_2^T B_2 \mathbf{v}_2 = 1\end{aligned} \tag{6}$$

Projection directions $\mathbf{v}_1$ and $\mathbf{v}_2$ will tend to balance the original feature extraction optimization with latent space covariance between exemplars that represent the same content. The vector form of Eqn6 is

$$\begin{aligned}\begin{bmatrix}\hat{\mathbf{v}}_1\\ \hat{\mathbf{v}}_2\end{bmatrix} &= \underset{\mathbf{v}_1,\mathbf{v}_2}{\operatorname{argmax}} \begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix}^T \begin{bmatrix} A_1 & \alpha Z_1 Z_2^T \\ \alpha Z_2 Z_1^T & \mu A_2 \end{bmatrix} \begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix} \\ s.t. & \begin{bmatrix}\mathbf{v}_1^T & \mathbf{v}_2^T\end{bmatrix} \begin{bmatrix} B_1 & 0 \\ 0 & \gamma B_2\end{bmatrix} \begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\end{bmatrix} = 1\end{aligned} \tag{7}$$

Equivalently,

$$\begin{aligned}\hat{\mathbf{v}} &= \underset{\mathbf{v}}{\operatorname{argmax}} \; \mathbf{v}^T \tilde{A} \mathbf{v} \\ & s.t. \; \mathbf{v}^T \tilde{B} \mathbf{v} = 1 \\ \Rightarrow \; & \tilde{A}\hat{\mathbf{v}} = \lambda \tilde{B} \hat{\mathbf{v}}\end{aligned} \tag{8}$$

Here, $\hat{\mathbf{v}}^T = [\hat{\mathbf{v}}_1^T \; \hat{\mathbf{v}}_2^T]$ and matrices $\tilde{A}$ and $\tilde{B}$ are the square symmetric matrices in Eqn7.

The final objective function is a standard generalized eigenvalue problem that can be solved using any eigensolver. It will produce real eigenvectors and eigenvalues because both $\tilde{A}$ and $\tilde{B}$ are square symmetric matrices. When data dimensions are greater than the number of classes, $\tilde{B}$ could be positive semi-definite and the problem becomes ill-posed. We can add a regularizer to the $\tilde{B}$ or project the original feature vectors to a lower dimensional subspace to handle this.

2.2. Multiview Extensions

There are several unsupervised and supervised feature extraction techniques with different properties in a single view scenario such as PCA [23], LDA [1], LPP [10], NPE [9], MFA [25] and their kernel versions. Three popular unsupervised multi-view feature extraction techniques are CCA [19, 8], BLM [22] and PLS [16, 19, 18]. We showed in the last subsection that a feature extraction technique in the form of a QCQP (Eqn1) can be extended to a multiview scenario using our framework. Plugging in different (A, B) pairs for different feature extraction techniques in our framework we can obtain multi-view extensions of PCA [23], LDA [1], LPP [10], NPE [9] and MFA [25]. We also show the relation between CCA, BLM and PLS and Generalized Multiview PCA or GMPCA as specific instances of our general framework. For further discussion, we use X_i to denote the *data matrix* with columns that are data samples in view i with the mean subtracted.

2.2.1 CCA, BLM, PLS and GMPCA

PCA in the i^{th} view is the following eigen-value problem

$$X_i W_i X_i^T \mathbf{v}_i = \lambda \mathbf{v}_i \tag{9}$$

$W_i = I_i/N_i$ with N_i equal to number of samples and I_i is the identity matrix in the i^{th} view. With different A_i, B_i and Z_i's in Eqn7 we get

- **GMPCA** $A_i = X_i W_i X_i^T$, $B_i = I$, $Z_i = X_i$
- **CCA** $A_i = 0$, $B_i = X_i W_i X_i^T$ and $Z_i = X_i$.
- **BLM** $A_i = X_i W_i X_i^T$, $B_i = I$ and $Z_i = X_i$ i.e. *same as GMPCA*.
- **PLS** $A_i = 0$, $B_i = I$ and $Z_i = X_i$. The difference from our approach is that in PLS eigen-vectors are found using asymmetric deflation of X_i's [19].

So, we see that all four approaches are related to each other under the proposed GMA framework.

2.2.2 Generalized Multiview LDA or GMLDA

LDA in the i^{th} view is the following eigenvalue problem

$$X_i W_i X_i^T \mathbf{v}_i = \lambda X_i D_i X_i^T \mathbf{v}_i \tag{10}$$

W_i and D_i are $N_i \times N_i$ matrices with $W_i^{kl} = 1/N_i^c$ if X_i^k and X_i^l belong to class c, 0 otherwise, N_i^c is the number of samples for class c in view i and $D_i = I - W_i$ [25, 10]. So, $A_i = X_i W_i X_i^T$, $B_i = X_i D_i X_i^T$ in Eqn7. For Z_i we have different choices; we can align corresponding samples giving $Z_i = X_i$, or class means, giving $Z_i = M_i$, with M_i defined as the matrix with columns that are class means. We choose class mean as exemplars because LDA tries to collapse all the class samples to the class mean. So if we align class means in different views we expect the samples to be aligned. Under some situations the within-class variation may not be a unimodal Gaussians. In such cases, samples from the same class can be clustered, and the class can be represented by the cluster centers as exemplars.

2.2.3 Generalized Multiview Marginal Fisher Analysis

LDA assumes a Gaussian class distribution, a condition that is often violated in real-world problems. Marginal Fisher Analysis, or MFA, is a technique that does not make this assumption, and instead tries to separate different- and compress same-class samples in the feature space [25]. It leads to following eigenvalue problem

$$X_i(S_{bi} - W_{bi})X_i^T\mathbf{v} = \lambda X_i(S_{wi} - W_{wi})X_i^T\mathbf{v} \tag{11}$$

here, $S^{kk}_{(b/w)i} = \sum_{kl,k\neq l} W^{kl}_{(b/w)i}$. The *within class compression* or *intrinsic* graph for the i^{th} view is defined as

$$W^{kl}_{wi} = \begin{cases} 1 & : k \in R_i^{k1}(l) \;\; or \;\; l \in R_i^{k1}(k) \\ 0 & : \text{otherwise} \end{cases} \tag{12}$$

Here, $R_i^{k1}(l)$ indicates the index set of the $k1$ nearest neighbors of the sample x_i^l in the same class. The *between class separation* or *penalty* graph for i^{th} view is defined as

$$W^{kl}_{bi} = \begin{cases} 1 & : (k,l) \in P_i^{k2}(c_l) \;\; or \;\; (k,l) \in P_i^{k2}(c_k) \\ 0 & : \text{otherwise} \end{cases} \tag{13}$$

Here, $P_i^{k2}(l)$ is a set of data pairs that are the $k2$ nearest pairs among the set $\{(k,l) :$ k and l are not in the same class$\}$. Hence, $A_i = X_i(S_{bi} - W_{bi})X_i^T$, $B_i = X_i(S_{wi} - W_{wi})X_i^T$ and $Z_i = X_i$. Similarly, multi-view extensions of LPP [10](the same as LPCCA [21]) and NPE [9] can be derived.

2.3. Kernel GMA

Kernel GMA involves mapping to a non-linear space and then carrying out GMA in that mapped space to obtain projection directions ν_i for the i^{th} view. So, we replace X_i with $\Phi_i = [\phi(x_i^1), \phi(x_i^2) \ldots \phi(x_i^{N_i})]$ and observe that $\nu_i = \Phi_i \tau_i$. The exemplars in kernel space are the columns of $N_i \times z$ matrix $\mathcal{Z}_i = \Phi_i G_i$, $N_i = \#$ samples in view i), z (same for each view) is the number of exemplars in each view, and G_i is an appropriately chosen $N_i \times z$ matrix. For example - G_i is the $N_i \times N_i$ identity matrix if all the samples are chosen to be exemplars and $N_i \times C$ matrix with $G_i^{r,c} = 1/N_i^c$ if the r^{th} sample belongs to class c, $C = \#$ of classes and $N_i^c = \#$ of samples in class c. The resulting eigenvalue problem $\tilde{\mathcal{A}}\tau = \lambda\tilde{\mathcal{B}}\tau$ will give $N = \sum_{i=1}^{V} N_i$ dimensional eigenvectors τ, which can be broken down into V parts to obtain the dual form of the eigenvectors for V views. These dual vectors will be used to project test sample $\mathbf{t}_i^j$ into the common non-linear latent space as

$$\mathbf{t}^j_{common} = \sum_{n=1}^{N_i} \varphi(\mathbf{t}_i^j.\mathbf{x}_i^n).\tau_i^n = \tau_i^T \phi(\mathbf{t}_i^j) \tag{14}$$

Here, $\phi(\mathbf{t}_i^j)$ is an $N_i \times 1$ vector of kernel evaluations of $\mathbf{t}_i^j$ with all the data samples in the i^{th} view.

2.4. More than two views

For more than two views simple algebra tells that we need to set $\tilde{A}$ and $\tilde{B}$ as

$$\tilde{A} = \begin{bmatrix} A_1 & \lambda_{12} Z_1 Z_2^T & \cdots & \lambda_{1n} Z_1 Z_n^T \\ \lambda_{12} Z_1^T Z_2 & \mu_2 A_2 & \cdots & \lambda_{2n} Z_2 Z_n^T \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1n} Z_n^T Z_1 & \lambda_{2n} Z_n^T Z_2 & \cdots & \mu_n A_n \end{bmatrix} \tag{15}$$

$$\tilde{B} = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_n B_n \end{bmatrix} \tag{16}$$

3. Experiments

In this section we test the proposed GMA approach on problems for cross-view classification with available class labels, showing improvement over other approaches.

3.1. Pose and Lighting Invariant Face Recognition

This is a problem with simultaneous cross view (pose) and within-class (lighting) variation. We use the MultiPIE [7] face dataset, which has 337 subjects' face images taken across 15 different poses, 20 illuminations, 6 expressions and 4 different sessions. We have done experiments using 5 poses ranging from frontal to profile (75°) at an interval of 15°. We have considered 18 lighting conditions for our experiments (illuminations 1 to 18). All the images are cropped (40 by 40 pixels) and aligned using 4 hand annotated fiducial points (eyes, nose tip and mouth) and affine transformations.

In the training phase, multiple images of a person (under different lighting conditions) in two different poses $p1$ and $p2$ are used to learn pairs of pose-specific projection directions $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$, respectively. At testing time, gallery and probe images are projected using learned pairs of pose-specific projection directions i.e. a face image in pose p is projected on $\hat{\mathbf{v}}_p$. 1-NN matching is done in the feature space using the normalized correlation score as a metric. We use two different modes for our recognition experiments.

Mode1 matches the conditions in a number of prior experiments and **Mode2** highlights our ability to generalize to unseen classes that were not used to obtain the latent space projection directions. In all our experiments, the gallery consists of a single image per individual, taken in the frontal pose with a frontal light (illum 7); probe images come from all poses and illuminations.

- **Mode1** We use training images of 129 subjects from session 01 (these 129 subjects were selected because they appear in all 4 sessions which allows future evaluation across sessions 03 and 04) under 5 lightings (1, 4, 7, 12 and 17) and testing images of the same subjects from session 02 under all 18 lightings.
- **Mode2** Training images of 120 subjects from session 01 (different than the one chosen in Mode1) under 5 lightings (1, 4, 7, 12 and 17); testing images are the same as Mode1 testing images.

We have used LDA and MFA with the proposed GMA approach and called the resulting approach GMMFA and GMLDA respectively. A naive way to obtain discriminant directions in two views is to learn a common subspace using CCA followed by LDA in the latent space (CCA+LDA) or LDA in individual spaces followed by CCA to get a common space (LDA+CCA). Surprisingly, neither of these approaches has been used before and we found that even these naive approaches outperform some competitive approaches. LDA, PCA, CCA, BLM, CCA+LDA and LDA+CCA are implemented by us. PLS, BLM and CCA have been used before for pose invariant face recognition to achieve state-of-the-art results on the CMU PIE dataset using PLS (code[1]) [18]. However, we find that with simultaneous pose and lighting variations all three perform poorly. Performance for Gabor [14], Local Feature Hashing or LFH [26], PittPatt [26], Sparse coding [24] are taken directly from the papers. Since, all the implemented approaches lead to large eigenvalue problems, we use PCA to reduce the data dimension before feeding it to any of the feature extraction techniques. We kept the top p principal components that retained 95% of the variance. For GMA based approaches we fix $\alpha = 10$, $\mu = 1$, $\gamma = \frac{tr(B_1)}{tr(B_2)}$, $k1 = 50$, $k2 = 400$ (for GMMFA), and all samples are taken as exemplars for both GMMFA and GMLDA. Parameters for MFA ($k1$ and $k2$) were selected based on the guidelines given in [25]. For simple LDA and PCA, different illumination images in gallery and probe poses are used together to learn common projection directions. The dimension of the feature space is selected by choosing the top k eigenvectors that contain 98% of the total eigenvalues produced by the eigenvalue problems involved in finding projection directions. We tried similar approaches to automatically determine the dimension for PLS based classification but the results were very poor. So for PLS only, we did testing for all possible dimensions and report the best accuracy. While reporting the results from [14] we have considered results for the selected 18 illumination conditions only. PittPatt is a commercial face recognition software and its results were taken directly from [26]. LFH uses a hashing technique with SIFT features for face recognition and frontal, $45°$ and $90°$ in the gallery for pose robustness in contrast to our approach in which we have used only frontal pose in the gallery. Use of SIFT features provides some tolerance to pose, and a multi-pose gallery makes matching possible across different poses. The results for LFH and PittPatt are reported using the same 129 subjects from session 02 used in our testing set with gallery images in the left illumination condition, whereas, we have used frontal illumination as the gallery image. However, we found that using any of the 18 illuminations as gallery with GMLDA and GMMFA resulted in negligible differences in performance compared to those reported in Table 2. In [24], the authors have used a sparse representation for simultaneous registration and recognition. They have reported results for pose and lighting invariant face recognition for $15°$ probe pose only, under all illuminations with a gallery of 249 subjects and reported 77.5% accuracy whereas we have used a gallery of 129 subject and report 99.7%.

The results from the experiments are shown in Table2. It is clear that GMMFA and GMLDA outperformed other approaches except [14] for large pose differences but overall performance of the proposed GMA based approach is better than all the domain-specific as well as generic approaches. Surprisingly, LDA performance is better than CCA, which is not expected due to the large pose difference. This unexpected observation indicates the importance of using label information in training. It also explains the improvements offered by GMLDA, because GMLDA is a fusion of CCA and LDA. Unfortunately, LDA cannot be used in cases when the data dimensions are different in different views, for example - image-text or text-link cases.

3.2. Text-Image Retrieval

Text-image retrieval is yet another cross-view problem that requires a common representation. We show results on two publicly available datasets - Pascal VOC 2007 [12, 11, 5] and Wiki Text-Image data [15]. Pascal data consists of 5011/4952(training/testing) image-tag pairs collected by the authors in [12, 11, 5] and it has 20 different classes. We used the publicly available features [2] consisting of histograms of bag-of-visual-words, GIST and color for images and *relative* and *absolute* tag ranks for text with a Chi-square kernel (see [12] for details). Some images are

[1] http://www.cs.umd.edu/~djacobs/pubs_files/PLS_Bases.m

[2] http://www.cs.utexas.edu/~grauman/research/datasets.html

Table 2. Performance for MultiPIE pose and lighting invariant face recognition. The upper and lower blocks of the table show the results in Mode1 and Mode2 respectively. Some approaches from other published works have not reported results for all pose differences; the absence is indicated by '-'.

Method	Probe pose					Avg
	15°	30°	45°	60°	75°	
PCA	15.3	5.3	6.5	3.6	2.6	6.7
PLS [18]	39.3	40.5	41.6	41.1	38.7	40.2
BLM [18]	46.5	55.1	59.9	63.6	61.8	57.4
CCA [18]	92.1	89.7	88.0	86.1	83.0	83.5
LDA[a]	98.0	94.2	91.7	84.9	79.0	89.5
CCA+LDA	96.4	96.0	93.6	86.2	83.6	91.2
LDA+CCA	95.9	94.9	93.6	91.3	89.9	93.1
PittPatt [26] [a]	94	34.0	3.0	–	–	–
LFH [26][a]	63	58	61	41	43	53.2
Sparse [24][a]	77.5	–	–	–	–	–
GMLDA	**99.7**	**99.2**	**98.6**	**94.9**	**95.4**	**97.6**
GMMFA	**99.7**	**99.0**	**98.5**	**95.0**	**95.5**	**97.5**
PCA	14.0	4.9	6.1	3.3	2.4	6.2
PLS [18]	29.0	26.2	23.3	17.3	12.4	21.6
BLM [18]	53.9	44.6	34.3	22.5	20.8	35.3
CCA [18]	79.5	62.2	46.1	19.5	14.4	44.3
LDA[a]	88.5	68.9	56.2	21.7	21.0	51.3
CCA+LDA	79.5	58.0	44.6	21.0	20.1	44.6
LDA+CCA	74.9	54.7	37.8	13.4	11.0	38.4
Gabor [14][a]	77.9	74.5	58.1	**45.2**	**31.0**	57.4
GMLDA	**92.6**	**80.9**	**64.4**	32.3	28.4	**59.7**
GMMFA	**92.7**	**81.1**	**64.7**	32.6	28.6	**59.9**

[a] Domain-dependent for cross-view classification

Table 4. mAP scores on Pascal data.

Query	Others		Proposed	
	KPLS	KCCA	KGMMFA	KGMLDA
Image	0.279	0.298	0.421	**0.427**
Text	0.232	0.269	0.328	**0.339**
Average	0.256	0.283	0.375	**0.383**

multi-labeled so we selected images with only one object from the training and testing set, which resulted in 2808 training and 2841 testing data. The category of the object is used as the content so we have a 20 class problem. A second data set, Wiki Text-image, consists of 2173/693(training/testing) image-text pairs with 10 different classes. We have used the same data as supplied by the authors[3]. It has a 10 dimensional latent Dirichlet allocation model [2] based text features and 128 dimensional SIFT histogram image features (see [15] for more details). Both data sets have class labels that can be leveraged in our proposed GMA framework to achieve within-class invariance. The task is to retrieve images/text from a database for a given query text/image. A correct retrieval is one that belongs to the same class as the query. So we want more and more correct matches in the top k documents for a better retrieval.

Semantic Correlation Matching (SCM) with a linear kernel [15] has shown state-of-the-art performance for Wiki data, so we have compared the proposed GMLDA and GMMFA with CCA, PLS, BLM, SCM and Semantic Matching (SM) [15]. SM corresponds to using Logistic regression in the image and text feature space to extract semantically similar feature to facilitate better matching. SCM refers to the use of Logistic regression in the space of CCA projected coefficients (a two-stage learning process). Results for SM and SCM are directly taken from the paper [15]. The authors in [12] have shown the advantage of using a Chi-square kernel over a linear mapping so we have used a Chi-square kernel for Pascal data for all the methods resulting in KernelCCA (KCCA), KernelPLS (KPLS), KernelGMLDA (KGMLDA) and KernelGMMFA (KGMMFA). For GMA based approaches we fix $\alpha = 100$, $\mu = 1$, $\gamma = \frac{tr(B_1)}{tr(B_2)}$, $k1 = 500$, $k2 = 2200$ (for GMMFA) and all the samples belonging to a class are taken as exemplars for both GMMFA and GMLDA. We have kept same number of dimensions for all the methods as mentioned in [15] and [12] i.e. 10 for Wiki and 20 for Pascal. Precision at 11 different recall levels $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ is used to evaluate the performance. The Mean Average Precision (mAP) scores= $\frac{1}{11}\sum_{r=0}^{r=1} Precision_r$ for text and image query for Wiki and Pascal data are listed in Table3 and Table4, respectively. It is evident that GMLDA and GMMFA outperform CCA, PLS, BLM and SM on Wiki data. Surprisingly, our generic single-stage approach's performance is similar to the domain specific two-stage SCM approach. We also outperformed KCCA and KPLS on Pascal data. The improvement is more for Pascal data because there are more classes (20 vs 10) and more testing samples (2841 vs 693) as compared to Wiki data, which requires better union of within-class samples for better performance.

4. Conclusion

We have proposed a novel generic framework for multi-view feature extraction by extending several unsupervised and supervised feature extraction techniques to their multi-view counterpart. We call the proposed framework Generalized Multiview Analysis or GMA. It is a first step towards unified multi-view feature extraction. The proposed approach is general and kernelizable, simultaneously learns multi-view projection directions and generalizes across unseen classes. We have shown that any feature extraction technique in the form of a generalized eigenvalue problem can be extended to its multi-view counterpart and we have

[3] http://www.svcl.ucsd.edu/projects/crossmodal/

Table 3. mAP scores for image and text query on Wiki text-image data.

Query	Others					Proposed	
	PLS	BLM	CCA	SM	SCM	GMMFA	GMLDA
Image	0.207	0.237	0.182	0.225	**0.277**	0.264	0.272
Text	0.192	0.144	0.209	0.223	0.226	0.231	**0.232**
Average	0.199	0.191	0.196	0.224	0.252	0.248	**0.253**

used GMA to obtain multi-view counterparts of PCA, LDA, LPP, NPE and MFA. We have also unified CCA, PLS, BLM as specific instances of Generalized Multiview PCA. Using LDA and MFA in our framework we have significantly outperformed all generic and most of the domain specific approaches for pose and lighting invariant face recognition. Using the same general framework we have also shown state-of-the-art results on text-image retrieval on Wiki data and outperformed generic approaches on Pascal data. GMA has outperformed CCA for all tasks when label information is available therefore, proving to be a superior alternative for CCA under similar conditions.

5. Acknowledgement

Work supported by US Office of Naval Research under MURI Grant N00014-08-10638.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEE TPAMI*, 19(7):711–720, 1997. 2, 3, 4

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003. 7

[3] Q. Chen and S. Sun. Hierarchical multi-view fisher discriminant analysis. ICONIP '09, pages 289–298, 2009. 2

[4] T. Diethe, D. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. In *ECML PKDD*, pages 328–343. Springer-Verlag, 2010. 2

[5] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. http://www.pascalnetwork.org/challenges/voc/voc2007/workshop/index.html. 6

[6] J. D. R. Farquhar, H. Meng, S. Szedmak, D. R. Hardoon, and J. Shawe-taylor. Two view learning: Svm-2k, theory and practice. In *NIPS*. MIT Press, 2006. 2

[7] R. Gros, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multipie. *Image and Vision Computing*, 28:807–813, 2010. 5

[8] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*, 16:2639–2664, 2004. 2, 4

[9] X. He, D. Cai, S. Yan, and H. Zhang. Neighborhood preserving embedding. In *ICCV*, volume 2, pages 1208 –1213, 2005. 3, 4, 5

[10] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE TPAMI*, 27(3):328–340, 2005. 2, 3, 4, 5

[11] S. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *TPAMI, IEEE*, 2011. 6

[12] S. J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, pages 1–12, 2010. 2, 6, 7

[13] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *TPAMI, IEEE*, 29(6):1005 –1018, 2007. 2

[14] A. Li, S. Shan, and W. Gao. Coupled bias-variance trade off for cross pose face recognition. *IEEE TIP*, 2011. 2, 6, 7

[15] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM Multimedia*, pages 251–260, 2010. 2, 3, 6, 7

[16] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *LNCS*, pages 34–51, 2006. 2, 4

[17] A. Sharma, A. Dubey, P. Tripathi, and V. Kumar. Pose invariant virtual classifiers from single training image using novel hybrid-eigenfaces. *Neurocomputing*, pages 1868–1880, 2010. 2

[18] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR*, pages 593–600. IEEE, 2011. 2, 4, 6, 7

[19] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. 2, 4

[20] L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *TPAMI*, 33(1):194 – 200, 2011. 2

[21] T. Sun and S. Chen. Locality preserving cca with applications to data visualization and pose estimation. *Image and Vision Computing*, 25(5):531 –543, 2007. 2, 5

[22] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000. 2, 4

[23] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 2, 3, 4

[24] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination by sparse representation. *CVPR*, pages 597–604, 2009. 6, 7

[25] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE TPAMI*, 29(1):40–51, 2007. 2, 3, 4, 5, 6

[26] Z. Zeng, T. Fang, S. Shah, and I. Kakadiaris. Local feature hashing for face recognition. In *IEEE BTAS*, pages 1–8, 2009. 2, 6, 7