Simultaneous and Orthogonal Decomposition of Data using Multimodal Discriminant Analysis

Terence Sim[†] Sheng Zhang[‡] Jianran Li[†] Yan Chen[†]

[†] School of Computing, National University of Singapore, Singapore 117417.
[‡] Department of Psychology, University of California, Santa Barbara, CA 93106-9560, USA.

Abstract

We present Multimodal Discriminant Analysis (MMDA), a novel method for decomposing variations in a dataset into independent factors (modes). For face images, MMDA effectively separates personal identity, illumination and pose into orthogonal subspaces. MMDA is based on maximizing the Fisher Criterion on all modes at the same time, and is therefore well-suited for multimodal and mode-invariant pattern recognition. We also show that MMDA may be used for dimension reduction, and for synthesizing images under novel illumination and even novel personal identity.

1. Introduction and Related Work

In computer vision, machine learning and data mining, one often deals with datasets that exhibit multiple "modes" *i.e*. independent variations that are simultaneously present in the data. For example, in face images, facial appearance depends not only on the person's identity, but also on illumination, and pose, each of which may be independently varied (Fig. 1). These multimodal (or multi-factor) variations cause severe difficulties for classical analyses and recogniton techniques, which implicitly assume single mode variation. For example, it is well-known that Eigenfaces [10] performs well for face recognition only under highly constrained conditions — usually frontal faces, neutral expressions and fixed lighting. This is equivalent to requiring that personal identity be the only mode that is varying.

Recognizing this shortcoming, researchers have over the years developed methods that can deal with datasets exhibiting more than one mode. For instance, Fisherfaces [1] model personal identity using the so-called *between-class* scatter matrix, and all other modes using the *within-class* scatter matrix. Likewise, Bayesfaces [8] model personal identity using *interpersonal* variation, and everything else using *extrapersonal* variation. A related problem is the discovery of hidden factors in image patterns, which Freeman and Tenenbaum attempt to solve in [4]. Although such techniques are useful, it would be ideal if they could be extended to handle multiple modes simultaneously.

Figure 1. Samples from the Multi-PIE dataset [6]: each row is one person under three different illuminations and two poses.

To this end researchers have recently turned to multilinear methods for multimodal decomposition. Ideally, after decomposition, each mode is also invariant to the other modes. This would then factor a face image into personal identity, illumination and pose, which is clearly useful when recognizing faces. Moreover, such a decomposition would also provide a means for compactly representing the image, as well as enable convenient re-rendering of the image by changing one mode independently of the others. This was in fact attempted in the early days of the Active Appearance Model (AAM) [3], when the authors Costen *et al*. tried to find what they called "orthogonal functional subspaces" [2] that could decouple the three modes of identity, illumination and pose. Unfortunately, they did not succeed. More recently, a method called Tensorfaces [11] was proposed to achieve multimodal decomposition by using Multilinear Principal Components Analysis (MPCA). Here, the authors devised an orthogonal iterative algorithm based on k-mode Singular Vector Decomposition (SVD), which extends conventional matrix SVD to tensors (higher-order matrices). The resulting decomposition succeeded in finding the modes due to identity, illumination and pose. However, because it is based on PCA, Tensorfaces achieved a decom-

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

position that was suited for pattern representation (just like PCA does), but not necessarily for pattern classification.

This raises the obvious question: can a multimodal decomposition based on the Fisher Linear Discriminant, or FLD, (also known as Linear Discriminant Analysis, or LDA) be achieved? If so, then such a method would be more suited for pattern classification than Tensorfaces, in the same way that the FLD is better for classification (at least in theory) than PCA.

In this paper, we propose a novel multimodal decomposition method based on the FLD, which we call Multimodal Discriminant Analysis (MMDA). Our method is conceptually easy to understand, and efficient to compute. The heart of our paper lies in Theorem 2 where we prove that under reasonable conditions, the different modes project onto distinct and orthogonal subspaces, which we term *Identity Spaces* (see Figure 2). Moreover, by retaining the *Residual Space* which is the orthogonal complement of all the Identity Spaces, our multimodal decomposition is fully invertible. Our paper makes the following contributions:

1. We show that MMDA simultaneously decomposes a dataset into different modes that occupy orthogonal subspaces. These are the "orthogonal functional subspaces" that eluded Costen *et al*. [2], and that make MMDA suitable for multimodal and mode-invariant pattern recognition.

2. We show how MMDA can be used to reduce the dimension of pattern vectors. In particular, the dimension is reduced in a way that preserves the Fisher Criterion, thereby ensuring that no discriminant information is lost after reduction. MMDA also allows for easy reconstruction, so that any dimension-reduced pattern vector can be mapped back to the original high-dimensionsal space. For images, this provides a useful way to visualize low-dimensional data.

3. We show that MMDA permits the synthesis of new pattern vectors, because each mode can be independently varied. We illustrate this using face images by synthesizing different illuminations and identities.

Arguably, the closest existing method is Tensorfaces, which has been used for multimodal decomposition, classification, dimension reduction, and synthesis [7, 11, 12]. MMDA can therefore be considered an alternative method. But as we will show, MMDA enjoys a number of advantages over Tensorfaces: it is easier to understand and implement because it is based on standard linear algebra, rather than multilinear algebra; it is more efficient to compute, and better for mode-invariant classification, dimension reduction, and synthesis. We demonstrate these advantages by proving MMDA's theoretical properties, and by running extensive experiments using face images.

2. Mathematical Background

In our previous work [13, 14], we showed that by prewhitening a set of pattern vectors, the FLD can achieve the best possible Fisher Criterion (Equation (3)) of $+\infty$.

We begin by letting $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$, denote a dataset of D-dimensional feature vectors. Each feature vector $\mathbf{x}_i$ belongs to exactly one of C classes $\{L_1, ..., L_C\}$. Let $\mathbf{m}_k$ denote the mean of class L_k, and suppose each class has the same number of vectors n, so that $N = nC$. Without loss of generality, we will assume that the global mean of $\mathbf{X}$ is zero, *i.e.* $(\sum_i \mathbf{x}_i)/N = \mathbf{m} = \mathbf{0}$. If not, we may simply subtract $\mathbf{m}$ from each $\mathbf{x}_i$.

2.1. Whitened Fisher Linear Discriminant (WFLD)

To whiten the data, first compute the total scatter matrix $\mathbf{S}_t = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top$, then eigen-decompose it to get $\mathbf{S}_t = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, retaining only non-zero eigenvalues in the diagonal matrix $\mathbf{D}$ and their corresponding eigenvectors in $\mathbf{U}$. Now compute the $(N-1) \times D$ matrix $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$, and apply it to the data to get the $(N-1) \times N$ matrix: $\widetilde{\mathbf{X}} = \mathbf{P}^\top \mathbf{X}$. The data is now whitened because the scatter matrix of $\widetilde{\mathbf{X}}$ is the identity matrix $\mathbf{I}$. Call the whitened class means $\widetilde{\mathbf{m}}_k$. On the whitened data we now define the between-class scatter matrix $\widetilde{\mathbf{S}}_b$, and the within-class scatter matrix $\widetilde{\mathbf{S}}_w$ in the usual way:

$$\widetilde{\mathbf{S}}_b = \sum_{k=1}^{C} n \widetilde{\mathbf{m}}_k \widetilde{\mathbf{m}}_k^\top \tag{1}$$

$$\widetilde{\mathbf{S}}_w = \sum_{i=1}^{C} \sum_{\widetilde{\mathbf{x}}_i \in L_k} (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k)(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k)^\top \tag{2}$$

We had proven [13, 14] that the Fisher Criterion,

$$J_F(\mathbf{\Phi}) = trace\{(\mathbf{\Phi}^\top \widetilde{\mathbf{S}}_w \mathbf{\Phi})^{-1} (\mathbf{\Phi}^\top \widetilde{\mathbf{S}}_b \mathbf{\Phi})\}, \tag{3}$$

is equal to the ratio $\frac{\lambda_b}{\lambda_w}$, where λ_b, λ_w are the eigenvalues of $\widetilde{\mathbf{S}}_b, \widetilde{\mathbf{S}}_w$, respectively. It may be shown that $\lambda_b + \lambda_w = 1$, so that by keeping the eigenvectors corresponding to $\lambda_b = 1$ in a matrix $\mathbf{V}$, the subspace spanned by $\mathbf{V}$ achieves $J_F = \frac{\lambda_b}{\lambda_w} = \frac{1}{0} = +\infty$, and is therefore the most discriminative subspace. Its dimension is $C - 1$.

2.2. Identity Space

We now extend our previous work by proving a property of the subspace spanned by $\mathbf{V}$, which we call *Identity Space*. We will show that (a) all points from the same class project onto the class mean; and (b) all within-class variation has been "projected out". That is, Identity Space reveals the class label (identity) of a data point, hence justifying its name. This also means that in Identity Space, all classes are *perfectly separated* (at least for training data).

Theorem 1. *In WFLD, if* $\mathbf{V}$ *is the set of eigenvectors of* $\widetilde{\mathbf{S}}_w$ *associated with* $\lambda_w = 0$*, then*

$$\mathbf{V}^\top \widetilde{\mathbf{x}}_i = \mathbf{V}^\top \widetilde{\mathbf{m}}_k, \qquad \forall \widetilde{\mathbf{x}}_i \in L_k. \tag{4}$$

Proof. Consider any $\mathbf{v} \in \mathbf{V}$. Since $\mathbf{v}$ is in the null space of $\widetilde{\mathbf{S}}_w$, we have $\widetilde{\mathbf{S}}_w \mathbf{v} = \mathbf{0}$, or $\mathbf{v}^\top \widetilde{\mathbf{S}}_w \mathbf{v} = 0$. So,

$$\begin{aligned} 0 &= \mathbf{v}^\top \widetilde{\mathbf{S}}_w \mathbf{v} && (5)\\ &= \sum_{k=1}^{C} \sum_{\widetilde{\mathbf{x}}_i \in L_k} \mathbf{v}^\top (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k)(\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k)^\top \mathbf{v} && (6)\\ &= \sum_{k=1}^{C} \sum_{\widetilde{\mathbf{x}}_i \in L_k} \|\mathbf{v}^\top (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k)\|^2 && (7) \end{aligned}$$

This is a sum of squared norms, which is zero if and only if each term is zero. Thus,

$$\begin{aligned} \forall i,k \quad \mathbf{v}^\top (\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k) &= \mathbf{0} && (8)\\ \Longrightarrow \mathbf{v}^\top \widetilde{\mathbf{x}}_i &= \mathbf{v}^\top \widetilde{\mathbf{m}}_k && (9) \end{aligned}$$

This is true for any $\mathbf{v} \in \mathbf{V}$, and so Equation (4) follows. □

Remarks: (1) We see that the vector $\mathbf{y}_k = \mathbf{V}^\top \widetilde{\mathbf{m}}_k \in \mathbb{R}^{(C-1)}$, may be used to represent the identity of class L_k, since all samples of the class project onto it. We may therefore call it the *Identity Vector*. (2) We can see what makes the Theorem true: Equation (8) shows that Identity Space is orthogonal to the vector difference $\widetilde{\mathbf{x}}_i - \widetilde{\mathbf{m}}_k$. But this vector is nothing but the within-class variation of class L_k. Hence all within-class variation are projected out of Identity Space. (3) It may be proven that as long as class means are distinct, a vector from one class will not project onto the Identity Vector of a different class.

3. Multimodal Discriminant Analysis

3.1. Theory

The previous section derived the Identity Space based on the classes of one mode. Extending to multiple modes (*e.g.* illumination and pose for face images) is easy: simply compute the eigenvectors (with unity eigenvalues) of the scatter matrix of each mode p : $\widetilde{\mathbf{S}}_b^p \mathbf{V}^p = \mathbf{V}^p$. That is, the Identity Space $\mathbf{V}^p$ is the span of $\widetilde{\mathbf{S}}_b^p = n^p \sum \widetilde{\mathbf{m}}_k^p (\widetilde{\mathbf{m}}_k^p)^\top$. We need to prove that the Identity Spaces are mutually orthogonal.

Now consider a dataset $\widetilde{\mathbf{X}}$ with M modes, with the p^{th} mode having C^p classes, and each class (in that mode) having n^p data points. Further suppose that the dataset contains the full Cartesian product of the modes, *e.g.* the face dataset contains C^1 people, with each person under C^2 poses, and each pose under C^3 illuminations. The total number of data points (vectors) is simply the product of the number of classes from all modes:

$$N = \prod_{p=1}^{M} C^p \tag{10}$$

This also means that $N = n^p C^p$, $\forall p = 1, \ldots, M$. We may now state our key theorem.

Theorem 2. *If* $\mathbf{V}^p$ *and* $\mathbf{V}^q$ *are the Identity Spaces for modes* p *and* q $(p \neq q)$*, then*

$$(\mathbf{V}^p)^\top \mathbf{V}^q = \mathbf{0}. \tag{11}$$

See Appendix A for the proof.

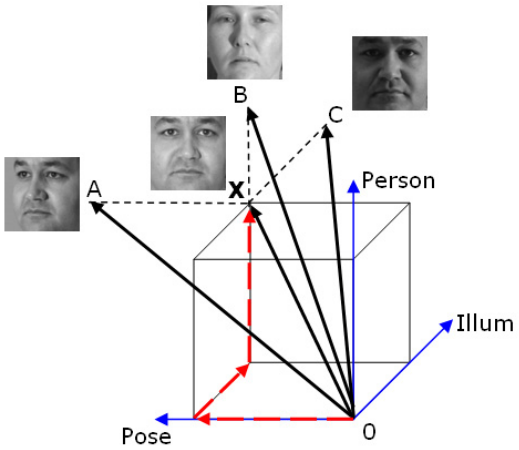


Figure 2. Illustrating MMDA: the face $\mathbf{X}$ (black vector) is decomposed into three orthogonal vectors (red), corresponding to the pose, illumination, and personal identity axes. Each axis (blue) is in fact a subspace (Identity Space) and all subspaces are mutually orthogonal. Moving along each axis (*i.e.* moving in each subspace) changes only one mode. Thus face A shows the same person in the same illumination but different pose; face B shows a different person in the same illumination and pose; while face C shows the same person in the same pose but different illumination.

Figure 2 illustrates the action of MMDA: a data vector (face) is resolved into orthogonal vectors corresponding to each mode present in the vector. Each orthogonal axis is in fact a subspace, whose dimension is $C^p - 1$. Moving along one axis changes only its corresponding mode; the other modes are unaffected.

Observe that all the Identity Spaces taken together span a subspace of dimension $\sum_{p=1}^{M} C^p - M$, but the whitened data vectors span a much larger subspace of dimension $N - 1 = \prod_{p=1}^{M} C^p - 1$. The remaining subspace therefore has dimension $r_0 = N - \sum_p C^p + M - 1$, which we will call *Residual Space*, denoted by $\mathbf{V}^0$.

Residual Space is simply the intersection of the (spans of the) within-class scatter matrices of all modes. It contains any residual variations that are outside of all the Identity

Spaces. It may be computed using the Gram-Schmidt procedure [5]: for each vector $\widetilde{\mathbf{x}}_i \in \widetilde{\mathbf{X}}$, subtract from it all its projections onto the existing set of orthonormal basis (initialized with the bases from all the Identity Spaces), and normalize the remaining component to form an orthonormal basis for Residual Space.

3.2. Discussion

Let us highlight the key issues in MMDA. To begin, note that there is no overlap between the different Identity Spaces. More precisely, $\mathbf{V}^p$ contains only the discriminant information for mode p, and not any other mode. That the Identity Spaces are mutually orthogonal means that the between-class variation for one mode is contained in the within-class variation for another mode. This is a desirable property. Moreover, each $\mathbf{V}^p$ is an orthogonal basis for its Identity Space. Note also that the Fisher Criterion is maximized to $+\infty$ in each Identity Space (for that mode), because $J_F = \frac{\lambda_b}{\lambda_w} = \frac{1}{0}$ in each Identity Space.

Classification: If we define $\mathbf{Q} = [\mathbf{V}^1 \cdots \mathbf{V}^M]$ and compute $\mathbf{t} = \mathbf{Q}^\top \widetilde{\mathbf{x}}$, then we would have simultaneously decomposed $\widetilde{\mathbf{x}}$ into all its modes. The first $(C^1 - 1)$ components of $\mathbf{t}$ may be used to classify $\widetilde{\mathbf{x}}$ in mode 1, the next $(C^2 - 1)$ components of $\mathbf{t}$ may be used to classify $\widetilde{\mathbf{x}}$ in mode 2, and so on. Because the class means are perfectly separated in each Identity Space, classification may be done using the nearest-neighbor rule with the Euclidean distance metric. In this way, MMDA permits *simultaneous* and *mode-invariant* classification in all modes. See also Section 4.1.

Furthermore, the components of $\mathbf{t}$ are statistically uncorrelated, because its scatter matrix (and hence covariance matrix) is diagonal: $\mathbf{Q}^\top \widetilde{\mathbf{S}}_\mathbf{t} \mathbf{Q} = \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$. This property is desirable, since many statistical learning techniques require feature vectors to be uncorrelated.

Invertibility: If we define $\mathbf{Q}' = [\mathbf{Q}\ \mathbf{V}^0]$, (*i.e.* if we include also the Residual Space), then $\mathbf{Q}'$ is an $(N-1) \times (N-1)$ orthogonal matrix, which makes the decomposition fully invertible. In other words, MMDA decomposes the whitened data space into a direct sum of the Identity Spaces and Residual Space: $\mathbb{R}^{N-1} = \mathbf{V}^0 \oplus \mathbf{V}^1 \oplus \cdots \oplus \mathbf{V}^M$. This property allows us to decompose a data vector, change some or all of its components, and reconstruct to synthesize new vectors. We will exploit this in Section 4.3.

Comparison with Tensorfaces: Note that in MMDA, the Identity Spaces are truly mutually orthogonal. This is not so with Tensorfaces. In [11] the authors claim that Tensorfaces "orthogonalizes ... and decomposes the tensor as the mode-n product of N-orthogonal spaces". This statement is misleading. The spaces are *not* orthogonal in the linear algebraic sense (*i.e.* their inner product is not zero). Rather, they are "orthogonal" simply because the data vectors are visualized as being arranged in a (hyper) cuboid, with each orthogonal face of the cuboid corresponding to one mode.

It follows then that the tensor decomposition does not orthogonalize one mode from another. To be sure, the modes are factored out as linear combinations of tensor bases, but the bases themselves are the tensor products (analogous to the matrix outer product) of non-orthogonal spaces. Therefore, it is unclear whether the modes are invariant to one another after tensor decomposition. It is also unclear whether the tensor decomposition yields components that are statistically uncorrelated.

Computational cost: In terms of computational efficiency, both Tensorfaces and MMDA have similar time complexity during the learning phase. Assuming that the dimension D of the data vectors is greater than the number of training samples N, the k-mode SVD algorithm in Tensorfaces requires $\mathcal{O}(kN^2D)$ time, while MMDA takes $\mathcal{O}(N^2D)$ time, being dominated by the calculation of the whitening matrix $\mathbf{P}$. However, when MMDA is used to decompose an input vector, it takes only $\mathcal{O}(ND)$ time. This contrasts sharply with Tensorfaces, which requires $\mathcal{O}(N^2D)$ time to perform the tensor decomposition. As for memory requirements, both methods require $\mathcal{O}(ND)$ of storage space.

What are modes? So far, we have been appealing to the reader's intuition by using the example of personal identity, illumination and pose in face images. If we tried to include gender (sex) as a mode, we would be wrong, because personal identity is intimately bound to gender, and it is not possible to independently vary the two. In fact, gender is a coarse grouping (superclass) of identity. However, it is legitimate to consider gender, illumination and pose as modes for face images, because personal identity is no longer included. In other words, modes must not exhibit subclass-superclass relationships. Likewise, ethnicity (race) cannot be a mode along with personal identity.

Technically speaking then, for MMDA, we may define *modes* as different sets of class labels that may be assigned to the data vectors, in such a way that the entire dataset satisfies Equation (10). Note that this constraint is crucial to Theorem 2, as can be seen in Equation (14). Without it, the Theorem would fail. Tensorfaces has the same constraint, and in practice this is not particularly onerous: for example, most face datasets satisfy this constraint.

Sufficient conditions: Due to the page limitation, we simply state without proof the sufficient conditions for MMDA to work: (a) $D \geq N - 1$; (b) all data vectors are linearly independent; and (c) the modes statisfy Equation (10). These conditions are sufficient to guarantee that the Identity Spaces exist for all modes, and that the multimodal decomposition can be performed.

Condition (a) (the "small sample size" problem in the literature) is easily satisfied when one deals with images, which typically gives rise to very high-dimensional vectors, and fewer training images. It suggests that in the high-dimensional data space, there is always sufficient room for

MMDA to find orthogonal subspaces that cleanly decouple the modes. When this condition is not met (*i.e.* $D < N$), we can attempt to use kernel methods to map the data vectors into a higher dimension. Indeed, we have done so, and have run other experiments to validate this idea. Hence this condition is not really a problem.

Condition (b) is trickier. It may be violated in practice because having too much data (large N) could result in linear dependence among the vectors. This is the case for a number of our experiments in Section 4. We overcame this problem by limiting N. Another solution is to resort to the kernel trick, taking care to use a non-linear mapping function (because a linear mapping function preserves the linear dependence in the data).

By far, Condition (c) is the most problematic. This condition ensures that no matter which mode is being considered (say, personal identity), the other modes (say, illumination) are equally present among the classes, so that one class is not distinguishable from another due to these other modes, but only due to the mode being considered. For example, if one person is imaged under 5 illuminations, while another is imaged under 4 illuminations, then the discriminative information does not lie solely in the ID, but also in the illuminations. This difference in illumination will show up in the between-class scatter matrix, rather than in the within-class scatter matrix. If this condition is not satisfied, it is not immediately clear how to overcome it. We leave this as a future work item.

4. Experiments

4.1. Mode-invariant classification

MMDA is well-suited for multimodal and mode-invariant classification. We illustrate this with a number of face recognition experiments: across pose, illumination and facial expression, and focus just on the mode-invariant properties of MMDA. We use the classification procedure described in Section 3.2. For the face images, we use a subset of the CMU Multi-PIE dataset [6] (Figure 1). The Multi-PIE images exhibit four modes: personal identity, illumination, pose, facial expression, and were collected over four different recording sessions spanning several months. This makes the dataset very useful for validating MMDA.

Recognition across expression and illumination. The purpose of this experiment is to see how MMDA performs on three modes: personal identity (ID), illumination, and facial expression. The respective number of classes in each mode are: $C^1 = 64, C^2 = 19, C^3 = 3$. The three facial expressions are: neutral, surprise, squint. The pose is frontal and fixed, and the images come from Session 02. We vary n, the number of training images per person from 3 to 30 (out of a maximum of 57 images per person), and test on the rest. We run our experiment 10 times, randomly drawing n training images in each run. We report the mean and standard deviation of recognition accuracy (*i.e.* classification of ID) in Figure 3. The figure also compares the performance of PCA, LDA, Tensorfaces and MMDA. It is clear that more training images benefit all the four classification methods, and that MMDA (red curve) consistently outperforms the rest.

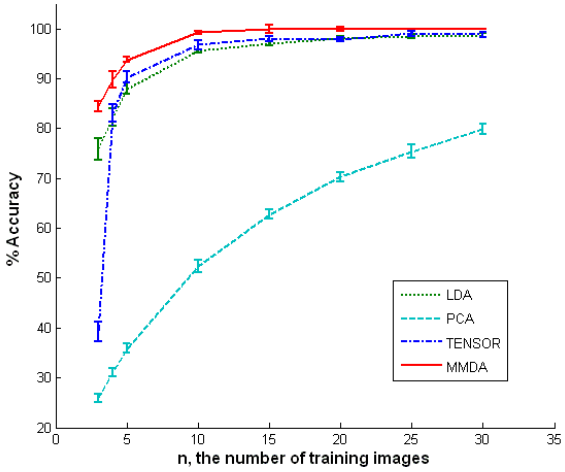


Figure 3. Face recognition under fixed pose, varying illumination and facial expression. The curves and errorbars show the mean and std. dev. of recognition accuracy (over 10 runs) for PCA, LDA, Tensorfaces, and MMDA.

Train:Test	PCA	Tensorfaces	MMDA
01 : 02	15.27	29.59	**41.26**
02 : 01	12.86	29.63	**40.92**

Table 1. Face recognition under varying pose and illumination, and across sessions. The numbers are the mean recognition accuracy over 10 runs. The best accuracy in each row is shown in bold. Overall performance is low across all classification methods, but MMDA is the top performer.

Recognition across pose and illumination. We now attempt a more difficult problem: person identification when both pose and illumination are changing. This is a three-mode problem, and we use a dataset comprising 70 people ($= C^1$) under 7 poses ($= C^2$) and 6 illuminations ($= C^3$). The problem is made harder because we are training on one session and testing on another session. Over two sessions, imaging conditions would have changed, and people would appear different due to changes in hairstyle, skin tone, and slight aging. We report our results in Table 1, where we compare PCA, Tensorfaces and MMDA. The numbers are the mean recognition accuracy over 10 runs. As expected, performance is low across all methods, due to the challenging conditions. But MMDA is the top performer.

Recognition across illumination. To better understand the previous experiment, we run it again, but this time keep-

Pose	PCA	Tensorfaces	MMDA
Pose 05_0	70.95	70.24	**83.57**
Pose 05_1	43.33	64.76	**82.14**
Pose 04_1	56.67	58.57	**65.95**
Pose 19_0	47.62	63.81	**65.24**
Pose 08_0	53.81	62.05	**63.57**
Pose 14_0	22.86	55.95	**60.48**
Pose 13_0	16.67	55.48	**56.19**

Table 2. Face recognition across illumination. Training images contain 7 illuminations under 1 pose from session 01; while testing images are from session 02 under the same pose. The highest recognition accuracy in each row is highlighted in bold, showing that MMDA is consistently the best.

ing the same pose for training (using session 01) and testing (using session 02). Table 2 shows that recognition rates improve significantly. This suggests that pose variation is the main cause of poor performance in the previous experiment. This is probably because pose variation is highly non-linear and cannot be adequately captured by a linear subspace — something that other researchers have also discovered.

4.2. Dimension Reduction

Dimension reduction comes naturally for MMDA: simply project the input vector $\mathbf{x} \in \mathbb{R}^D$ onto the required Identity Spaces. For example, to retain only the discriminant information for illumination, compute $\mathbf{t} = (\mathbf{V}^{\text{illum}})^\top \mathbf{P}^\top \mathbf{x}$. The dimension of $\mathbf{t}$ is $T-1$, where T is the dimension of the illumination subspace. Note that $\mathbf{t}$ contains sufficient discrminative information for illumination classification only. If we wish to classify in all modes, then we project $\mathbf{x}$ onto all Identity Spaces. The dimension is now $\sum_{p=1}^{M} C^p - M$.

In other words, MMDA may be considered a *supervised* dimension reduction method, which preserves the Fisher Criterion for all modes. Note that the dimension of $\mathbf{t}$ is fully determined by the dimensions of the Identity Spaces used for the projection, and cannot be chosen arbitrarily. This is an advantage over PCA and Tensorfaces, in which the reduced dimension is a free parameter that needs to be tuned. As an example, consider the dataset in our previous experiment. We could reduce the dimension of our face images from 9696 (pixels) to 80 (if we kept all discriminative information for personal identity, pose and illumination), or 11 (if we ignored ID), or 5 (if we wished to perform only illumination classification).

MMDA also permits easy reconstruction after dimension reduction. For image datasets, this provides a convenient way to visualize the reduced-dimension vector. Let $\mathbf{Q} = [\mathbf{V}^1 \dots \mathbf{V}^s]$ contain the desired Identity Spaces, and let the high-dimensional vector $\mathbf{x}$ be reduced by $\mathbf{t} = \mathbf{Q}^\top \mathbf{P}^\top \mathbf{x}$. Then we may reconstruct via $\mathbf{x}_r = \mathbf{P}\mathbf{Q}\mathbf{t}$. Using this, we may now visualize the Identity Vectors (*i.e.* class means) of 10 people under 14 illuminations. Figure 4(a) shows the Identity Vectors of the 10 people. Note that they are clearly recognizable, and all exhibit the average illumination (which happens to be frontal), *i.e.* there is no illumination variation. Figure 4(b) shows the illumination Identity Vectors. Note that the images are all of the average face (there is no ID variation), but under 14 different illuminations. This clearly shows that the between-class variation of one mode is "projected out" (suppressed) in another mode.

As another example, we select images of the same 10 people under 14 illuminations as in the previous example, but this time also under 3 expressions. We then reduce the dimension to 24 ($= 10 + 14 + 3 - 3$), by projecting onto the 3 modes of ID, illumination and expression. And we reconstruct. For comparison, we also use PCA to reduce the dimension to 24 and reconstruct. Figure 5 shows the result. It is clear that although PCA achieves the optimal RMSE, MMDA produces images of comparable quality.

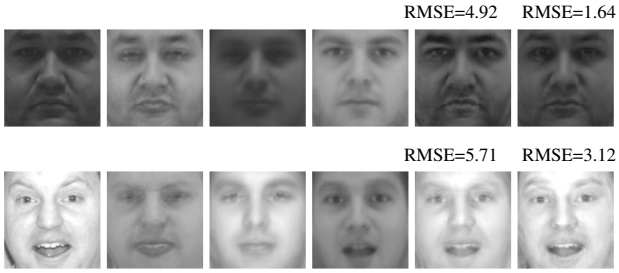


Figure 5. Comparing MMDA and PCA for dimension reduction and reconstruction. (Col. 1) Original images. (Col. 2) ID component. (Col. 3) Illumination component. (Col. 4) Expression component. (Col. 5) MMDA reconstruction. (Col. 6) PCA reconstruction after reducing to the same dimension as MMDA. Although PCA has a lower RMSE, MMDA reconstructs with comparable quality.

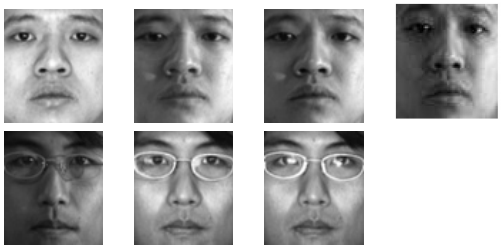
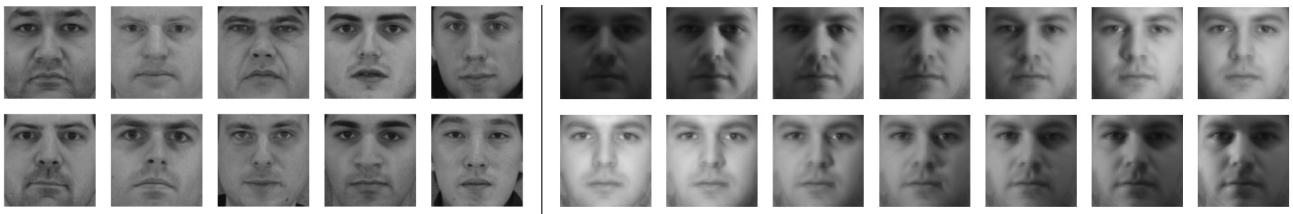


Figure 6. Illumination swap example. (Col. 1) Image pair: 2 persons under 2 illums. (Col. 2) Illumination swap using MMDA. (Col. 3) Ground truth illum. (Col. 4) Result from the Tensorfaces method of Lin *et al*. (for only one person).

4.3. Synthesis

MMDA can be used to synthesize new data as follows: (a) project the data vector $\mathbf{x}$ using $\mathbf{t} = (\mathbf{Q}')^\top \mathbf{P}^\top \mathbf{x}$; (b) alter some of the components in $\mathbf{t}$; and (c) reconstruct via $\mathbf{x}_r = \mathbf{P}\mathbf{Q}'\mathbf{t}$. See Section 3.2 for the definition of $\mathbf{Q}'$.

Swapping illumination: To illustrate this, we select a pair



(a) (b)

Figure 4. Reconstruction of the Identity Vectors of (a) 10 people under (b) 14 illuminations. In (a) all faces exhibit the average illumination, *i.e*. there is no illumination variation; while in (b) all faces are of the average person, *i.e*. there is no identity variation.

of face images from the CMU PIE dataset [9] of different people under different illuminations, and proceed to swap their illuminations. That is, given two vectors $\mathbf{t}_1, \mathbf{t}_2$ that are decomposed by MMDA into illumination and ID modes, we swap their components in the illumination Identity Space. However, we need to take care of their Residual Space components also. This is because Residual Space also contains residual ID and illumination information.

In general, information to discriminate between classes may be found in the class means (first-order statistics), as well as in their covariance matrices (second-order statistics). MMDA separates such discriminative information into Identity Spaces (which capture the class means) and Residual Space (which captures the covariance matrices). Classical FLD uses only the class means for discrimination and ignores the second-order statistics. This is why it will not work if class means are equal.

But the Identity Space is no longer sufficient for synthesizing new data. We also need to account for the second-order statistics present in Residual Space. The technical details are lengthy, so we simply sketch the basic idea: we need to align (rotate) the orthogonal basis $\mathbf{V}^0$ in Residual Space so that the coefficients of one class (in one mode) in Residual Space "has the same meaning" as those of another class from the same mode. In this example, we want person A's illumination coefficients (in Residual Space) to "mean the same" as those of person B. Once we align the basis, the coefficients may be swapped. Figure 6 shows the result, and compares it with the method of Lin *et al*. [7], which is based on Tensorfaces. At least for this example, MMDA handles shadows better (observable from the nose region).

Synthesizing identity: Our next example synthesizes novel identity. To do this, we first use the AAM [3] to fit contour lines on key facial features on a frontal face, *e.g*. chin, eyebrows, eyes, nose, mouth. These lines capture the shape $\mathbf{s}$ of the individual. We then normalize the face shape so that all faces are warped onto a reference face (Figure 7). This shape-normalized face contains only intensity (shading) information, which is re-arranged into a vector $\mathbf{y}$. Finally, we concatenate $\mathbf{s}$ and $\mathbf{y}$ to make our data vector $\mathbf{x}$.

We select 12 people under 6 illuminations from the CMU PIE dataset (Figure 7), compute their data vectors $\mathbf{x}$, and apply MMDA to decompose into ID and illumination. To create novel identities, we simply take linear combinations of the Identity Vectors in the ID space, reconstruct the vector, and unwarp the image to undo the shape normalization. Figure 8 clearly shows that the new identities are mixtures of those in Figure 7.

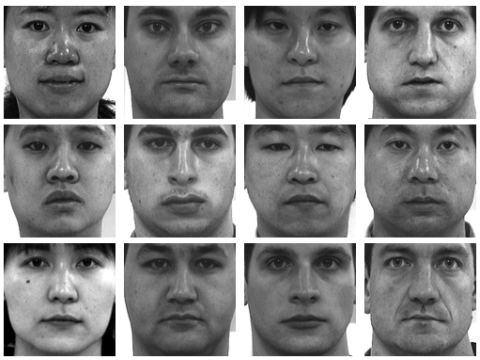


Figure 7. Applying AAM to normalize the shape of 12 people from the CMU PIE dataset.

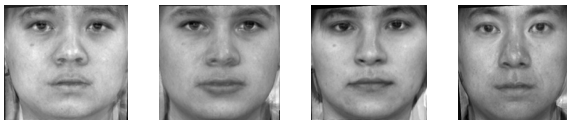


Figure 8. Identity synthesis: these novel "persons" are a mixture of those in Figure 7.

Synthesizing illumination: Finally we show an example of illumination synthesis. Taking the person in the 1st row, 1st column of Figure 7, we generate different linear combinations of the 6 illumination Identity Vectors. As can be seen from Figure 9, the identity in all images is preserved, while the illumination variation looks realistic.

5. Conclusion

To conclude, we have shown in this paper that MMDA (a) is easy to understand and implement; (b) efficient to compute; (c) permits multimodal and mode-invariant classification; (d) reduces dimension while preserving the Fisher Criterion; and (e) provides a convenient way to synthesize new data. We hope that these properties will make

Figure 9. Illumination synthesis: the person's identity is well-preserved, while illumination is varied to uniformly sample the illumination subspace (mode).

MMDA an attractive method for multimodal data analysis. We would like to acknowledge the generous support of NUS research grant R-252-000-261-422 for this work.

A. Proof of Theorem 2

Proof. Define $\mathbf{H}_b^p = [\cdots \widetilde{\mathbf{m}}_k^p \cdots]$, *i.e.* a matrix whose columns are the class means. In turn, the class means $\widetilde{\mathbf{m}}_k^p$ and $\widetilde{\mathbf{m}}_l^q$ for modes p and q can be written as

$$\widetilde{\mathbf{m}}_k^p = \frac{1}{n^p} \sum_{\widetilde{\mathbf{x}}_i \in L_k^p} \widetilde{\mathbf{x}}_i \text{ and } \widetilde{\mathbf{m}}_l^q = \frac{1}{n^q} \sum_{\widetilde{\mathbf{x}}_j \in L_l^q} \widetilde{\mathbf{x}}_j \tag{12}$$

Now consider the inner product between $\mathbf{H}_b^p$ and $\mathbf{H}_b^q$:

$$(\mathbf{H}_b^q)^\top \mathbf{H}_b^p = \left[(\widetilde{\mathbf{m}}_l^q)^\top \widetilde{\mathbf{m}}_k^p\right] = \frac{1}{n^p n^q}\left[\sum_i \sum_j \widetilde{\mathbf{x}}_j^\top \widetilde{\mathbf{x}}_i\right] \tag{13}$$

where $k = 1, \cdots, C^p$ and $l = 1, \cdots, C^q$. Before we compute $(\widetilde{\mathbf{m}}_l^q)^\top \widetilde{\mathbf{m}}_k^p$, recall that:

1. The inner product between $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ is given by Lemma 3.
2. The set $\{\widetilde{\mathbf{x}}_i \mid \widetilde{\mathbf{x}}_i \in L_k^p\}$ and $\{\widetilde{\mathbf{x}}_j \mid \widetilde{\mathbf{x}}_j \in L_l^q\}$ share $\frac{N}{C^p C^q}$ samples in common. Note that $\frac{N}{C^p C^q} = \frac{n^p n^q}{N}$ because $n^p C^p = n^q C^q = N$.

$$\begin{aligned} \text{So, } \sum_{i,j} \widetilde{\mathbf{x}}_j^\top \widetilde{\mathbf{x}}_i &= (n^p n^q - \frac{n^p n^q}{N})(-\frac{1}{N}) + \frac{n^p n^q}{N}(1 - \frac{1}{N}) \\ &= n^p n^q \left[(1 - \frac{1}{N})(-\frac{1}{N}) + \frac{1}{N}(1 - \frac{1}{N})\right] \\ &= 0 \end{aligned} \tag{14}$$

In other words, for any $p \neq q$, $(\mathbf{H}_b^q)^\top \mathbf{H}_b^p = \mathbf{0}$, *i.e.* $\mathbf{H}_b^p$ and $\mathbf{H}_b^q$ are mutually orthogonal. Since $\widetilde{\mathbf{S}}_b^p = n^p \sum_{k=1}^{C^p} \widetilde{\mathbf{m}}_k^p (\widetilde{\mathbf{m}}_k^p)^\top = n^p \mathbf{H}_b^p (\mathbf{H}_b^p)^\top$, and $\mathbf{V}^p$ is its span (with similar remarks for $\mathbf{V}^q$), the Theorem follows. □

B. A Lemma

The following Lemma is required for the proof of Theorem 2. It says that (a) the whitened data vectors all have the same length; and (b) the angle between any two vectors is the same.

Lemma 3. *For the whitened data matrix $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \cdots, \widetilde{\mathbf{x}}_N]$, the inner product between any two columns is given by:*

$$\widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j = \begin{cases} 1 - \frac{1}{N} & \text{if } i = j; \\ -\frac{1}{N} & \text{if } i \neq j. \end{cases} \tag{15}$$

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 1997.

[2] N. Costen, T. Cootes, G.J.Edwards, and C.J.Taylor. Simultaneous extraction of functional face subspaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 492–497, 1999.

[3] G. Edwards, T. Cootes, and C. Taylor. Face Recognition Using Active Appearance Models. In *European Conference on Computer Vision*, 1998.

[4] W. Freeman and J. Tenenbaum. Learning Bilinear Models for Two-Factor Problems in Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 554–560, 1997.

[5] G. Golub and C. V. Loan. *Matrix Computations, 3rd Edition*. Johns Hopkins Univ. Press, 1996.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[7] D. Lin, Y. Xu, X. Tang, and S. Yan. Tensor-based Factor Decomposition for Relighting. In *International Conference on Image Processing*, 2005.

[8] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):780–788, June 2002.

[9] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[10] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[11] M. Vasilescu and D. Terzopoulos. Multilinear Subspace Analysis of Image Ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[12] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005.

[13] S. Zhang and T. Sim. When Fisher Meets Fukunaga-Koontz: A New Look at Linear Discriminants. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 323–329, June 2006.

[14] S. Zhang and T. Sim. Discriminant Subspace Analysis: A Fukunaga-Koontz Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2007.