Bypassing Synthesis: PLS for Face Recognition with Pose, Low-Resolution and Sketch

Abhishek Sharma
Institute of Advanced Computer Science
University of Maryland, USA
bhokaal@umiacs.umd.edu

David W Jacobs
Institute of Advanced Computer Science
University of Maryland, USA
djacobs@cs.umd.edu

Abstract

This paper presents a novel way to perform multi-modal face recognition. We use Partial Least Squares (PLS) to linearly map images in different modalities to a common linear subspace in which they are highly correlated. PLS has been previously used effectively for feature selection in face recognition. We show both theoretically and experimentally that PLS can be used effectively across modalities. We also formulate a generic intermediate subspace comparison framework for multi-modal recognition. Surprisingly, we achieve high performance using only pixel intensities as features. We experimentally demonstrate the highest published recognition rates on the pose variations in the PIE data set, and also show that PLS can be used to compare sketches to photos, and to compare images taken at different resolutions.

1. Introduction

In face recognition, one often seeks to compare gallery images taken under one set of conditions, to a probe image acquired differently. For example, in criminal investigations, we might need to compare mug-shots to a sketch drawn by a sketch artist based on the verbal description of the suspect. Similarly, mug-shots or passport photos might be compared to surveillance images taken from a different viewpoint. The probe image might also be of lower resolution (LR) compared to a gallery of high resolution (HR) images.

We propose a general framework that uses Partial Least Squares (PLS) [16] to perform recognition in a wide range of multi-modal scenarios. PLS has been used very effectively for face recognition, but in a different manner, with different motivation [17, 19, 20, 21, 22]; our contribution is to show how and why PLS can be used for cross-modal recognition. More generally, we argue for the applicability of linear projection to an intermediate subspace for multi-modal recognition, also pointing out the value of the Bilinear Model (BLM) [14] for face recognition, which also achieves state-of-the art results on some problems. Experimental evaluation of our framework

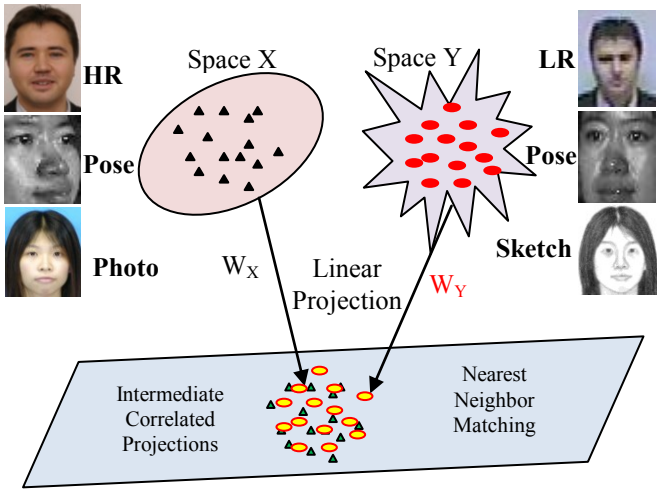

Figure 1: The basic over-view of the proposed method, W_X and W_Y are projection matrices learned using PLS on X and Y.

using PLS with pose variation has shown significant improvements in terms of accuracy and run-time over the state-of-art on the CMU PIE face data set [26]. For sketch-photo recognition, our method is comparable to the state of-art. We also illustrate the potential of our method to handle variation in resolution with a simple, synthetic example. In all three domains we apply exactly the same algorithm, and use the same, simple representation of images. Our generic approach performs either near or better than state-of-the-art approaches that have been designed for specific cross-modal conditions.

Our approach matches probe and gallery images by linearly projecting them into an intermediate space where images with the same identity are highly correlated (Figure 1). We argue that for a variety of cross-modality recognition problems, such projections will exist and can be found using PLS and BLM. One consequence of our approach is that we do not need to synthesize an artificial gallery image from the probe image.

1.1. Related Work

There has been a huge amount of prior work on comparing images taken in different modalities, which we

can only sample here. In much of this work, images taken in one modality are automatically converted to the second modality prior to comparison. For example a holistic mapping [1] is used to convert a photo image into a corresponding sketch image. In [2, 3, 5] the authors have used local patch based mappings to convert images from one modality to the other for sketch-photo recognition. Since the mapping from one modality to the other is generally non-linear, local patch based approaches generally perform better than the global ones because they can approximate the non-linearity in a better manner. [7] is a holistic and [6, 8, 9] are local patch-wise approaches to hallucinate a HR face image from a given LR face image and again a comparison reveals that local approaches performed better. For face recognition with pose and lighting variation [10, 23, 13] 3D knowledge of faces is used to warp an off-axis image to a frontal image, and to normalize lighting prior to comparison. These approaches may use representations that are specific to a domain, or may employ a more general, learning-based approach, that typically requires corresponding patches in the training set [13, 10, 18, 23]. Our approach does not attempt to synthesize images of one modality from another. While excellent work has been done on synthesis, this may in principle be an ill-posed problem that is more difficult than simply comparing images taken in two different modalities.

A second approach is to compare images using a representation that is insensitive to changes in modality. For example, Klare et al [4] used SIFT feature descriptors and multi-scale local binary patterns to represent grayscale and sketch images of faces then performed recognition based on this common representation. This approach worked well because both SIFT and LBP features extract gradient information that is approximately the same in both photo and sketch at corresponding positions. While some descriptors, such as SIFT, are robust across a range of variations in modalities, no single representation can be expected to handle all variations in modality.

Two prior methods are closer to our work in spirit, and have provided valuable inspiration. In [14] (BLM) the authors have used Singular Value Decomposition to derive a common content space for a set of different styles and [12] uses a probabilistic model to generate coupled subspaces for different poses. We discuss [14] further in the next section to provide motivation for our use of PLS, and we also compare experimentally to their representation. Recently, [24] used CCA to project images in different poses to a common subspace and compared them using probabilistic modeling. While related our approach is different in several ways: we achieve strong results using simple pixel intensities, without probabilistic modeling of patches; we show theoretically why projection methods can handle pose variation; and we show that PLS can outperform CCA with pose variation.

2. Bilinear Model

Tannenbaum and Freeman [14] proposed a bilinear model of *style* and *content*. In cross-modal face recognition, the two modes correspond to two styles, and subject identity corresponds to content. They suggest methods of learning BLMs and using them for a variety of tasks, such as identifying the style of a new image with unfamiliar content, or generating novel images based on separate examples of the style and content. However, their approach also suggests that their content-style models can be used to obtain style invariant *content* representation which can be used for classification of a sample in a different *style*.

Following their asymmetric model, modality matrices $\mathbf{A^m}$ can be learned by decomposing the matrix $\mathbf{Y}$ (which is a matrix in which the same subject's images under different modality are concatenated to make a long vector) using SVD as (see [14])**:**

$$\mathbf{Y} = \mathbf{USV^T} = (\mathbf{US})\mathbf{V^T} = (\mathbf{A})\mathbf{B} \quad (1)$$

$\mathbf{A}$ can be partitioned to give different modality models ($\mathbf{A^{m1}}$ and $\mathbf{A^{m2}}$) for our case m1 and m2 might represent two different poses or sketch and photo and so on. We know that matrix $\mathbf{U}$ has the eigenvectors of $\mathbf{YY^T}$ as its columns; denote the ith eigenvector and associated eigenvalue as λ_i and $\mathbf{u_i}$ respectively. So,

$$\mathbf{a_i} = \lambda_i \mathbf{u_i} = (\mathbf{YY^T})\mathbf{u_i} = \mathbf{Y}(\mathbf{Y^T u_i}) = \mathbf{Y}(\boldsymbol{\alpha_i}) \quad (2)$$

$\boldsymbol{\alpha_i}$ is a column vector with each element equal to the projection (inner product) of training images on eigenvectors $\mathbf{u_i}$ and $\mathbf{a_i}$ is the ith column of matrix A:

$$\alpha_{ik} = \mathbf{y_k^T u_i} \quad (3)$$

Hence, each eigenvector $\mathbf{u_i}$ and vector $\mathbf{a_i}$ can be defined as a linear combination of training images $\mathbf{y_k}$. To get the models for different modalities we need to partition the vectors $\mathbf{a_i}$ to yield $\mathbf{a_i^{m1}}$ & $\mathbf{a_i^{m2}}$ so from eqn (2) we get:

$$\mathbf{a_i^{m1}} = \mathbf{Y^{m1}}\boldsymbol{\alpha_i} \quad (4\text{-a})$$

$$\mathbf{a_i^{m2}} = \mathbf{Y^{m2}}\boldsymbol{\alpha_i} \quad (4\text{-b})$$

where, $\mathbf{Y^{m1}}$ and $\mathbf{Y^{m2}}$ are the matrices with images under modalities m1 and m2 as their columns. Now let's project a subject's face images under two different modalities m1 and m2 denoted as $\mathbf{f^{m1}}$ and $\mathbf{f^{m2}}$, on $\mathbf{a_i^{m1}}$ & $\mathbf{a_i^{m2}}$ to get the projection coefficients β_i^{mj} for j =1,2 as:

$$\beta_i^{mj} = (\mathbf{a_i^{mj}})^T \mathbf{f^{mj}} = (\mathbf{Y^{mj}}\boldsymbol{\alpha_i})^T \mathbf{f^{mj}} = \boldsymbol{\alpha_i^T}((\mathbf{Y^{mj}})^\mathbf{T}\mathbf{f^{mj}})$$

$$= \boldsymbol{\alpha_i^T}(\boldsymbol{\gamma^{mj}}) = \sum_{k=1}^{K} \alpha_{ik}\gamma_k^{mj} \quad (5)$$

Here, K is the total number of subjects used in the training set to learn matrix $\mathbf{A}$. Each element of vector $\boldsymbol{\gamma^{mj}}$ is

the inner product of test images $\mathbf{f}^{mj}$ with the training set images $\mathbf{y}_k^{mj}$. For the BLM to work properly for recognition, it is required that the corresponding projection coefficients (β_i^{mj} for j = 1, 2) should be approximately the same. This requires that the projection vectors $\boldsymbol{\gamma}^{mj}$'s should be approximately the same for j = 1, 2 (Eqn 5) which demands that the projection coefficients for every training image pair should be the same across modalities. By using SVD, they capture the variation in the images, while their BLM ensures that images of the same content and different styles will project to the same coordinates in this basis. However, the BLM may not hold when the corresponding images are not well correlated. In such cases, it may create a representation that captures variation in the data, at the expense of capturing the features that account for the correlation between images in different styles, as show in Figure 2. In this toy problem, the x-coordinates of corresponding points in X and Y are the same and the y-coordinates are uncorrelated. Projection to the x-axis makes the data perfectly correlated but removes much of the variance. $BL_{x/y}$ and $PLS_{x/y}$ corresponds to the projection directions found using BLM and PLS on two different sets of correlated points X and Y. Note that PLS still finds directions which makes the projections correlated while the BLM mainly represents variance in Y and consequently fails to obtain the optimal X direction too.

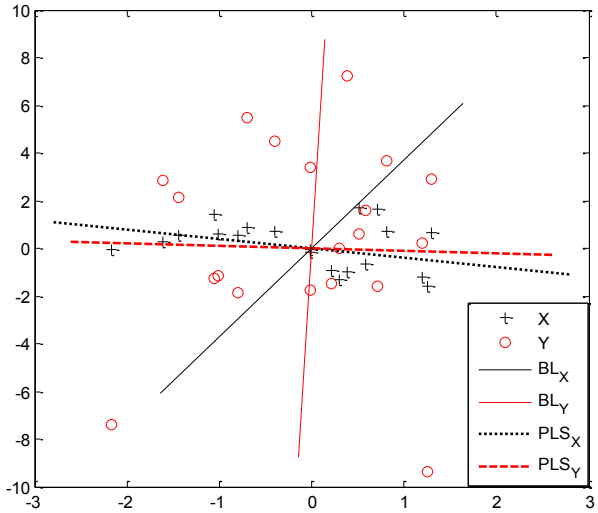

Figure 2: Comparison of PLS and BLM [14] for the case when the data X and Y are not correlated (see text for details). Note the different scales on x and y axes. This problem motivates our use of PLS for cross-modal recognition.

3. Partial Least Square

Partial Least Square analysis [16, 24 25] is a regression model that differs from Ordinary Least Square regression by first projecting the regressors (input) and responses (output) onto a low dimensional latent linear subspace. PLS chooses these linear projections such that the covariance between latent scores of regressors and responses is maximized and then it finds a linear mapping from the regressors' latent score to response's latent score. We apply PLS by using images from one modality as regressors and using corresponding images from a different modality as responses. In this way, we learn a linear projection for each modality that maps images into a common space in which they can be compared.

Partial Least Square has been previously used for face recognition [17, 19, 20, 21, 22]. We have been particularly motivated by the approach of [22], which achieves excellent experimental results. However, these results use PLS in a quite different way than we do. They used PLS to find a regression function from image feature space to a binary label space for performing one-vs-all classification. In [17, 19, 20, 21] PLS has been used to extract feature vectors in accordance with the label information. In this regard, it is very similar to Linear Discriminate Analysis (LDA), with the considerable advantage that given two classes, it can select an arbitrary number of linear features, rather than choosing a single linear projection. In contrast, we simply use pixel intensities as our features, and focus on PLS's ability to map images from different modalities into a common space.

There are several variants of PLS analysis based on the factor model assumption and the iterative algorithm used to learn the latent space [16, 24]. Some of these variants facilitate the intuition behind PLS while some are faster than others but the objective function for all of them is the same. In this paper, we have used the original NIPALS algorithm [12] to develop intuitions and a variant of NIPALS given in [25] to learn the latent space.

3.1. Description of PLS

Let us suppose that we have n observations (input space) and each of them is a p dimensional vector. In correspondence we have n observations lying in a q dimensional space as our output. Let **X** be the regressor matrix and **Y** be the response matrix where each row contains one observation so **X** and **Y** are (n×p) and (n×q) matrices respectively. PLS models **X** and **Y** such that:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \quad (6\text{-a})$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \quad (6\text{-b})$$

$$\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H} \quad (6\text{-c})$$

T and **U** are (n×d) matrices of the d extracted PLS scores or latent projections. The (p×d) matrix **P** and the (q×d) matrix **Q** represent matrices of loadings and the (n×p) matrix **E**, (n × q) matrix **F** and n×d matrix **H** are the residual matrices. **D** is a (d×d) diagonal matrix which relates the latent scores of **X** and **Y.** PLS works in a greedy way and finds a 1D projection of **X** and **Y** at each

iteration. That is, it finds normalized basis vectors **w** and **c** such that the covariance between the score vectors **t** and **u** (rows of **T** and **U**) is maximized:

$$max([cov(\mathbf{t},\mathbf{u})]^2) = max([cov(\mathbf{Xw},\mathbf{Yc})]^2) \quad (7)$$
$$s.t. \quad \|\mathbf{w}\| = \|\mathbf{c}\| = 1$$

PLS iterates this process with a greedy algorithm to find multiple basis vectors that project **X** and **Y** to a higher dimensional space.

It is interesting to compare this to the objective function of Canonical Correlation Analysis (CCA) to emphasize the difference between PLS and CCA. CCA tries to maximize the correlation between the latent scores

$$max([corr(\mathbf{Xw},\mathbf{Yc})]^2) \quad (8)$$

where, $$corr(\mathbf{a},\mathbf{b}) = \frac{cov(\mathbf{a},\mathbf{b})}{var(\mathbf{a}).var(\mathbf{b})}. \quad (9)$$

putting the expression from (9) into (7) we get the PLS objective function as:

$$max([var(\mathbf{Xw})].[corr(\mathbf{Xw},\mathbf{Yc})]^2.[var(\mathbf{Yc})]) \quad (10)$$
$$s.t. \quad \|\mathbf{w}\| = \|\mathbf{c}\| = 1$$

It is clear from (10) that PLS tries to correlate the latent score of regressor and response as well as captures the variations present in the regressor and response space too. CCA only correlates the latent score hence CCA fails to generalize well to unseen testing points and even fails to differentiate between training samples in the latent space under some special conditions. BLM on the other hand as shown in the figure 2 attempts to capture variation in both spaces. One toy condition where PLS will succeed and both BLM and CCA will fail to obtain meaningful directions can be stated as follows - Suppose we have two sets of 3D points X and Y and x_i^j and y_i^j denote the j^{th} element of the i^{th} data point in X and Y Suppose that the first coordinates of all x_i and y_i are equal to a constant k i.e.

$$\forall i, x_i^1 = y_i^1 = k \Rightarrow Var(X^1) = Var(Y^1) = 0$$

The second coordinates are correlated with a coefficient ρ which is less than 1 and the variance present in the second coordinate is ψ i.e.

$$corr(X^2, Y^2) = \rho \ \ \& \ \ Var(X^2), Var(Y^2) \approx \psi$$

The third coordinate is almost uncorrelated and the variance is >> ψ i.e.

$$corr(X^3, Y^3) \approx 0 \ \ \& \ \ Var(X^3), Var(Y^3) >> \psi$$

Under this situation CCA will give the first coordinate as the principal direction which projects all the data points in sets X and Y to a common single point in the latent space, rendering recognition impossible. BLM will find a direction which is parallel to the third coordinate which preserves the inter-set variance but loses all the correspondence. PLS however, will opt for second coordinate which preserves variance (discrimination) as well as maintains correspondence which is crucial for our task of multi-modal recognition.

PLS therefore strikes a balance between the objectives of Principal Component Analysis (PCA) and CCA. It should be noted that the dimension of regressor and response score vectors is the same and is equal to the number of extracted PLS bases. Hence, the latent representation of both regressor and response lies in the same vector space. Moreover, since PLS bases are such that the latent scores are highly correlated it can be safely assumed that regressor and response latent scores are roughly embedded in a single linear manifold, thus a simple Nearest Neighbor metric will suffice for recognition.

3.2. Learning PLS bases

Consider the regressor and response data matrices **X** and **Y** (both column centered) defined in section 2.1. We define the regression model as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \Rightarrow \mathbf{(XW)Z}^T + \mathbf{E} \Rightarrow \mathbf{TZ}^T + \mathbf{E} \quad (11)$$

The detailed step by step algorithm to obtain these variables is given in [25]. The MATLAB code to obtain **W** and **Z** can be found here http://www.cs.umd.edu/~djacobs/pubs_files/PLS_Bases.m. Here, **B** is the (p×q) regression matrix from **X** to **Y**, **W** is the (p×d) projection matrix from **X** to the latent space, **T** is the latent score matrix of **X** and **Z** is a (q×d) matrix representing the linear transformation from the d dimensional latent space to **Y**. So essentially we can project **Y** into the latent space and calculate its latent score **U** as**:**

$$\mathbf{U} = \mathbf{YZ(Z}^T\mathbf{Z)}^{-1} \quad (12)$$

Please note that the matrices **T**, **U** and **W** are not the same matrices as in section 2.1 but can be scaled and columns of W and Z (Eqn 11) are equivalent to **w** and **c** (Eqn 7).

4. When can PLS work?

We will use PLS to find linear projections **w** and **c** that map images taken in two modes into a common subspace. Equation (10) shows that PLS will seek **w** and **c** that tend to produce high levels of correlation in the projection of corresponding images from different modalities. However, PLS cannot be expected to lead to effective recognition when such projections do not exist. In this section, we show some conditions in which projections of images from two modalities exist in which the projected images are perfectly correlated (and in fact equal). Then we show that these conditions hold for some interesting examples of cross-modality recognition.

We should note that the existence of such projections is

not sufficient to guarantee good recognition performance. We will assess the actual performance of PLS empirically, in the next section.

4.1. Existence of correlated projections

In a number of cases, images taken in two different modes can be viewed as different, linear transformations of a single ideal object. Let $\mathbf{I_k}$ and $\mathbf{J_k}$ denote column vectors containing the pixels of corresponding images, taken in two modalities. We denote by $\mathbf{R_k}$ a matrix (or column vector) that contains an idealized version of $\mathbf{I_k}$ and $\mathbf{J_k}$, such that we can write:

$$\mathbf{I_k} = \mathbf{A}\,\mathbf{R_k} \qquad \mathbf{J_k} = \mathbf{B}\,\mathbf{R_k} \tag{13}$$

for some matrices **A** and **B.** We would like to know when it will be possible to find vectors **w** and **c** that project sets of images into a 1D space in which they are highly correlated. We consider a simpler case, looking at when the projections can be made equal. That is, when we can find **w** and **c** such that for any $\mathbf{I_k}$ and $\mathbf{J_k}$ satisfying Equation (13) we have:

$$\mathbf{w}^T\mathbf{I_k} = \mathbf{c}^T\mathbf{J_k} \Rightarrow \mathbf{w}^T\mathbf{A}\mathbf{R_k} = \mathbf{c}^T\mathbf{B}\mathbf{R_k} \tag{14-a}$$

$$\Rightarrow \mathbf{w}^T\mathbf{A} = \mathbf{c}^T\mathbf{B} \tag{14-b}$$

Equation (14-a,b) can be satisfied if and only if the row spaces of **A** and **B** intersect, as the LHS of the Eqn (14-b) is a linear combination of the rows of **A,** while the RHS is a linear combination of the rows of **B**. We now give some examples in which this condition holds.

4.2. High resolution vs. low resolution

For this situation, we can assume that the ideal image is just the high resolution image, so that **A** is simply the identity matrix, and $\mathbf{I_k} = \mathbf{R_k}$. $\mathbf{J_k}$ then, can be obtained by smoothing $\mathbf{R_k}$ with a Gaussian filter, and subsampling the result. Both operations can be represented in matrix form. Any convolution can be represented as a matrix multiplication. For this, the *i*'th row of **B** contains a vectorized Gaussian filter centered at the image location of the *i*'th pixel in $\mathbf{R_k}$. **B** can subsample the result of this convolution by simply omitting rows corresponding to pixels that are not sampled. Now because **A** is the identity matrix, it has full rank, and its row space must intersect that of **B**.

4.3. Pose variation

We now consider the more challenging problem that arises when comparing two images taken of the same 3D scene from different viewpoints. This raises problems of finding a correspondence between pixels in the two images, as well as accounting for occlusion. To work our way up to this problem, we first consider the case in which there exists a one-to-one correspondence between pixels in the image, with no occlusion.

Permutations: In this case, we can again suppose that **A** is the identity matrix. In this case, **B** will be a permutation matrix, which changes the location of pixels without altering their intensities. In this case, **A** and **B** are both of full rank, and in fact have a common row space. So again, there exist **w** and **c** that will project $\mathbf{I_k}$ and $\mathbf{J_k}$ into a space where they are equal.

Stereo: We now consider a more general problem that is commonly solved by stereo matching. Suppose we represent a 3D object with a triangular mesh. Let $\mathbf{R_k}$ contain the intensities on all faces of the mesh that appear in either image (We will assume that each pixel contains the intensity from a single triangle. More realistic rendering models could be handled with slightly more complicated reasoning). Then, to generate images appropriately, **A** and **B** will be matrices in which each row contains one 1 and is 0 otherwise. **A** (or **B**) may contain identical rows, if the same triangle projects to multiple pixels. The rank of **A** will be equal to the number of triangles that create intensities in **I**, and similarly for **B**. The number of columns in both matrices will equal the number of triangles that appear in either image. So their row spaces will intersect, provided that the sum of their ranks is greater than or equal to the length of $\mathbf{R_k}$, which occurs whenever the images contain projections of any common pixels.

As a toy example, we consider a small 1D stereo pair showing a dot in front of a planar background. We might have $\mathbf{I_k}^T$ =[7 8 2 5] and $\mathbf{J_k}^T$ = [7 2 3 5]. In this example we might have $\mathbf{R}^T_{\mathbf{k}}$ = [7 8 2 3 5] and:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

It can be inferred from the example that row spaces of A and B intersect hence we expect PLS to work.

4.4. Comparing images to sketches

Finally, we note that our conditions may approximately hold in the relationship between images and sketches. This is because sketches often capture the edges, or high frequency components of an image. A filter such as a Laplacian of a Gaussian produces an output that is similar to a sketch (eg., Figure 1). Again, the ideal image can be the same as the intensity image, while the sketch image can be produced by a **B** that represents this convolution, satisfying our conditions.

5. PLS based multi-modal recognition

Given a problem of multi-modal recognition such as – Different pose face images, Sketch and Photo etc. we can learn the PLS bases on a training set using the iterative algorithm given in [25]. Then using equations (11) and (12) we can project a pair of images of the same subject seen under two different modalities to the latent space to generate a pair of latent scores. Once the latent space scores are obtained we can do simple NN recognition. For practical purposes, we will simply calculate and store the latent projections of gallery images and compute the latent projection of the probe image online. In the next few sections we present our results on face recognition across poses, sketch-photo pairs and Low and High Resolution pairs.

5.1. Pose invariant face recognition

To demonstrate the efficacy of the proposed algorithm we have used it for face recognition across poses. We have used the CMU PIE face database [26] to evaluate and compare the performance of our method with other approaches. This database consists of 13 poses with large pose variation. In the past, many researchers have used this dataset to evaluate their algorithms. The dataset is divided into training (subject 1 to 34) and testing (subject 35 to 68) subsets. PLS bases corresponding to each of the different pose pairs are learned using the training set and recognition performance is evaluated on the testing set. For all the pose pairs we have used 30 PLS bases for our proposed method and 25 eigenvectors for the Bilinear Model; these values produce the best results. Since, there are 13 poses there are 126 galley and probe pairs. Table 1 reports the accuracy for all of these cases. For the purpose of comparison with other methods we have adopted two different protocols. Some methods have reported the accuracy in the form of Table 1 (34 face gallery with all possible pose pairs); for those the comparison has been done in Table 2. For others, comparison is done in Table 3, in accordance with their protocol. We are citing the results of performance by other methods directly from the papers except BLM and CCA for which we have done all the experiments. It should be noted that unlike [24] we have used CCA with simple pixel intensities without probabilistic modeling i.e as with PLS to compare the strength of PLS and CCA under equal conditions.

It is clear from the comparison that the proposed method is a significant improvement over prior methods. Note that on the two pose pairs reported in [12], we perform somewhat less well than their method. However, it is notable that their method requires 14 hand-clicked points by a human operator. They then compare responses of Gabor filters in the area of these points. Our method requires alignment of face thumbnails using the eyes and mouth. Moreover, when they have used simple intensity as the feature their accuracy dropped significantly.

Table 2 – Comparison of proposed method with others on 34 face gallery on CMU PIE dataset.

Methods	Accuracy	Time per comparison
Eigenfaces [10]	16.6	< 0.005 seconds
FaceIt [10]	24.3	> 5 minutes
ELF [10]	66.3	> 5 minutes
Bilinear Model [14]	79.6	< 0.005 seconds
4ptSMD [11]	86.8	0.35 Seconds
CCA	87.35	<0.005 Seconds
Proposed	**90.12**	**0.0046 seconds**

Table 3 – Results under different settings as per as the results reported by different authors.

Method	Gallery	Probe	Accuracy /proposed
PGFR [15]	c27	c05/37/25/22/29/11/14/34	86/**93.4**
TFA [12]	c27	c05/22	95/**90**
LLR [13]	c27	c05/29/37/11/07/09	94.6/**100**
ELF [10]	c27	c05/29/37/11/07/09	89.8/**100**

In addition, some authors that do not use a training set have reported results using a gallery of 68 individuals. In particular [18] has reported strong results in this setting. While we cannot compare directly to their results, we note that [11] reports results for galleries of 68 and 34 faces. With a gallery of 68 faces results in [11] are considerably better than those of [18] (82.4% vs. 74.3%) and with a gallery of 34 faces, our results are substantially better than those of [11] (90.1% vs. 86.8%). We note that our approach does require prior knowledge of the pose of the probe image, and a training set that contains example faces taken in a similar pose. A similar assumption is made in the ELF [10] algorithm. [18] makes use of hand-clicked points and a morphable model to compute face pose, while [11] uses hand-clicked points to compute the epipolar geometry relating the two images. Research and commercial systems have shown impressive performance in automatically computing pose. Some preliminary experiments on proposed method showed that recognition performance does not decrease drastically with slight change in pose between PLS bases and gallery/probe faces. Exploring this aspect thoroughly will be our future effort for evaluation using automatic pose identifiers.

5.2. Low resolution face recognition

This problem is yet another multi-modal problem because probe images from a surveillance camera are generally low resolution (LR) with slight motion blur and noise. The gallery generally contains high resolution (HR) faces. To verify the applicability of our method we have

Table 1: Accuracy for all the possible pose-pairs on CMU PIE dataset using proposed method overall accuracy for all pose pairs is **90.12%**

Gallery \ Probe	c34	c31	c14	c11	c29	c09	c27	c07	c05	c37	c25	c02	c22	**Avg**
c34	--	0.88	0.94	0.94	0.91	0.88	0.91	0.97	0.85	0.88	0.70	0.85	0.61	**0.862**
c31	0.85	--	1	1	1	0.88	0.85	0.91	0.85	0.88	0.76	0.85	0.76	**0.884**
c14	0.97	1	--	1	0.97	0.91	0.97	1	0.91	1	0.82	0.91	0.67	**0.928**
c11	0.79	0.97	1	--	1	0.88	1	1	0.97	0.97	0.85	0.88	0.67	**0.916**
c29	0.76	0.94	1	1	--	1	1	1	1	1	0.85	0.91	0.73	**0.933**
c09	0.76	0.88	0.91	0.94	0.94	--	0.97	0.94	0.91	0.88	0.82	0.79	0.70	**0.872**
c27	0.85	0.91	0.97	1	1	1	--	1	1	1	0.85	0.88	0.79	**0.939**
c07	0.79	0.91	0.97	1	1	0.97	1	--	1	0.97	0.85	0.91	0.76	**0.929**
c05	0.79	0.97	0.97	0.94	1	0.94	1	1	--	0.97	0.91	0.91	0.82	**0.936**
c37	0.79	0.94	1	0.94	0.94	0.88	0.94	0.94	0.97	--	1	1	0.94	**0.941**
c25	0.67	0.82	0.76	0.79	0.88	0.88	0.88	0.91	0.94	0.97	--	0.97	0.76	**0.855**
c02	0.76	0.88	0.88	0.94	0.94	0.88	0.97	0.94	1	1	1	--	0.97	**0.931**
c22	0.64	0.70	0.64	0.79	0.76	0.67	0.82	0.82	0.85	0.91	0.85	0.91	--	**0.784**

synthetically generated low resolution images for frontal face images in a subset of FERET face dataset and performed recognition. The original HR images were chosen to be 76×66 and different size LR images were tested for recognition. Fig. 3 shows the recognition accuracy of the proposed method. Note that a direct comparison of HR and LR face images with as low a resolution as 5×4 resulted in 60% recognition accuracy. Moreover, the number of PLS bases used in any case for optimal performance are not greater than 20 and for some cases just 3 PLS bases gave 95% accuracy. We have used 90 faces for training and 100 for testing. Due to lack of space we have not shown the results for BLM but it should be noted that it performed similarly. However, performance of CCA was very poor ranging between 30-50% only.

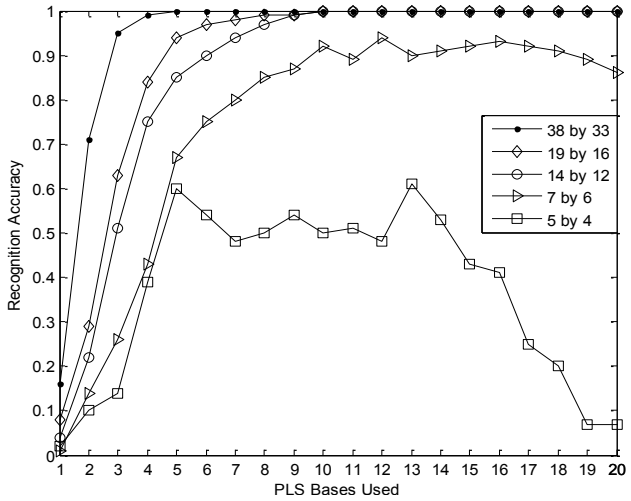

Figure 3: Accuracy for Low Resolution face recognition vs. the number of PLS bases used with different size LR images used

5.3. Sketch-Photo recognition

To demonstrate the generality of our proposed approach we have also tested it on a sketch vs. photo recognition problem. To test the performance of our method we have used a subset of the CUHK sketch – face dataset [3]. We used a subset containing 188 subjects' face images and corresponding hand drawn sketch pairs. 88 sketch-photo pairs were used as the training sample and the remaining 100 were used as the testing set. We formed 5 random partitions of the dataset to generate different sets of training and testing data and report the average accuracy. In this case, we have used 70 PLS bases and 50 eigenvectors for the Bilinear Model. A comparison of our method with other reported results is shown in Table 4.

From the comparison it is clear that in spite of being holistic in nature, the proposed method achieves respectable accuracy. We feel that this is encouraging because our method is completely general; we have used exactly the same algorithm for pose, LR face recognition and sketch. The table also reflects the trend that accuracy is increasing continuously as we move down from holistic to pixel level representation. So it may be possible that using patch-wise features with our method will improve the accuracy. It should be noted that in [5] and [4] the authors have used strong classifiers after extracting patch-wise and pixel based features, whereas we have simply used the NN metric after latent score extraction.

Table – 4 Sketch – Photo pair recognition accuracy

Method	Testing set	Type	Accuracy
Wang [1]	100	Holistic	81
Liu [5]	300	Patch-wise	87.67
Klare [4]	300	Pixel-wise	99.47
Proposed	100	**Holistic**	**93.6**
Bilinear	100	Holistic	94.2
CCA	100	Holistic	94.6

6. Conclusions

We have demonstrated a general latent space framework for cross-modal recognition and the relevance of PLS to cross-modal face recognition. Theoretically, we have shown that in principle, there exist linear projections of images taken in two modalities that map them to a space in which images of the same individual are equal. This is

true for images taken in different poses, at different resolutions, and approximately, for sketches and intensity images. Experimentally, we show that PLS and BLM can be used to achieve strong face recognition performance in these domains. Of particular note, we show that PLS has outperformed the best reported performance on the problem of face recognition with pose variation with impressive margin both in terms of accuracy as well as run-time and that Bilinear Models in all three domains outperformed many existing approaches. Moreover, using the exact same method we have also achieved comparable performance for sketch-photo and cross resolution face recognition.

Acknowledgements

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government. Authors would also like to thanks Raghuraman for helping with the cropped images of CMU PIE data. First author is grateful to Christine and Marcello for helping out with TA stuff giving him ample of time for research.

References

[1] X. Tang and X. Wang Face Sketch Recognition, IEEE Transactions on Circuits Systems for Video Technology, 14(1), 50-57, 2004.

[2] B. Xiao, X. Gao, D. Tao, Y Yuan and J. Li, Photo-sketch synthesis and recognition based on subspace learning, Neurocomputing 73, 840-852, 2010.

[3] X. Wang and X. Tang, Face photo-sketch synthesis and recognition, IEEE Transactions Pattern Analysis Machine Intelligence, 31(11), 1955-1967, 2009.

[4] B. Klare, Z. Li and A. K. Jain, Matching forensic sketches to mugshot photos, IEEE Pattern Analysis and Machine Intelligence, 29 Sept. 2010.

[5] Q. Liu, X. Tang, H. Jin, H Lu, S. Ma, Nonlinear Approach for Face Sketch Synthesis and Recognition, IEEE CVPR, 1005-1010, 2005.

[6] C. Liu, H. Y. Shum and W. T. Freeman, Face hallucination: theory and practice, IJCV 75(1), 115–134, 2007.

[7] J. Yang, H. Tang, Y. Ma and T. Huang, Face hallucination via sparse coding, IEEE Int. Conf Image Processing'08, 1264-1267, 2008.

[8] B. Li, H. Chang, S. Shan and X. Chen, Aligning coupled manifolds for face hallucination, IEEE Signal Processing Letters, 16(11), 957-960, 2009.

[9] Y. Zhuang, J. Zhang, and F. Wu, Hallucinating faces: LPH super-resolution and neighbor reconstruction for residue compensation, Pattern Recognition, 40, 3178-3194, 2007.

[10] R. Gross, I. Matthews, S. Baker, Appearance-based face recognition and light - fields, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4), 449–465, 2004.

[11] C.D. Castillo, D.W. Jacobs, Using stereo matching with general epipolar geometry for 2d face recognition across pose, Pattern Analysis and Machine Intelligence, 31(1), 2298 - 2304, 2009.

[12] S.J.D. Prince, J.H. Elder, J. Warrell, F.M. Felisberti, Tied Factor Analysis for Face Recognition across Large Pose Differences, IEEE Patt. Anal. Mach. Intell, 30(6), 970-984, 2008.

[13] X. Chai, S. Shan, X. Chen and W. Gao, Locally linear regression for pose invariant face recognition, IEEE Tran. Image Processing, 16(7), 1716-1725, 2007.

[14] J. B. Tenenbaum, W. T. Freeman, Separating style and content with bilinear models. Neural Computation 12 (6), 1247-1283, 2000.

[15] X. Liu, T. Chen, Pose-robust face recognition using geometry assisted probabilistic modeling, IEEE CVPR, vol. 1, 2005, pp. 502–509.

[16] R. Rosipal & N. Kräämer, Overview and recent advances in partial least squares, In Subspace, latent structure and feature selection techniques, Lecture Notes in Computer Science, Springer, 34-51, 2006.

[17] C. Dhanjal, S. R. Gunn and J. S. Taylor, Efficient sparse kernel feature extraction based on partial least squares, IEEE Patt. Anal. Mach. Intell. 31(8), 1947-1961, 2009.

[18] S. Romdhani, V. Blanz, and T. Vetter, Face Identification by Fitting a 3d Morphable Model Using Linear Shape and Texture Error Functions, Proc. ECCV, 4, 3-19, 2002.

[19] J. Baeka and M. Kimb, Face recognition using partial least squares components, Pattern Recognition, 37, 1303-1306, 2004.

[20] V. Struc, N. Pavesic, Gabor-based kernel partial-least-squares discrimination features for face recognition, Informatica, 20(1), 2009.

[21] X. Li, j Ma and S. Lia, Novel face recognition method based on a principal component analysis and kernel partial least square, IEEE ROBIO 2007, 1773-1777.

[22] W.R. Schwartz, H. Guo, L.S. Davis. A Robust and Scalable Approach to Face Identification. ECCV 2010.

[23] S. Romdhani, T. Vetter, D. J. Kriegman, Face recognition using 3-D models: pose and illumination, proc. of *IEEE*, 94(11), 1977 – 1999, 2006.

[24] A. Li, S. Shan, X. Chen, W Gao, Maximizing Intra-individual Correlations for Face Recognition Across Pose Differences, IEEE CVPR, 2009, pp. 605-611.

[25] Partial Least Square Tutorial, http://www.statsoft.com/textbook/partial-least-squares/#SIMPLS.

[26] T. Sim, S. Baker, and M. Bsat, The CMU Pose, Illumination, and Expression Database, IEEE Trans. Patt. Anal. Machine Intelligence, 25(12), 1615-1618, 2003.