Model-Driven Domain Adaptation on Product Manifolds for Unconstrained Face Recognition

Huy Tho Ho, and Raghuraman Gopalan

Abstract

Many classification algorithms see a reduction in performance when tested on data with properties different from that used for training. This problem arises very naturally in face recognition where images corresponding to the source domain (gallery, training data) and the target domain (probe, testing data) are acquired under varying degree of factors such as illumination, expression, blur and alignment. In this paper, we account for the domain shift by deriving a latent subspace or domain, which jointly characterizes the multifactor variations using appropriate image formation models for each factor. We formulate the latent domain as a product of Grassmann manifolds based on the underlying geometry of the tensor space, and perform recognition across domain shift using statistics consistent with the tensor geometry. More specifically, given a face image from the source or target domain, we first synthesize multiple images of that subject under different illuminations, blur conditions and 2D perturbations to form a tensor representation of the face. The orthogonal matrices obtained from the decomposition of this tensor, where each matrix corresponds to a factor variation, are used to characterize the subject as a point on a product of Grassmann manifolds. For cases with only one image per subject in the source domain, the identity of target domain faces is estimated using the geodesic distance on product manifolds. When multiple images per subject are available, an extension of kernel discriminant analysis is developed using a novel kernel based on the projection metric on product spaces. Furthermore, a probabilistic approach to the problem of classifying image sets on product manifolds is introduced. We demonstrate the effectiveness of our approach through comprehensive evaluations on constrained and unconstrained face datasets, including still images and videos.

H. T. Ho is with the Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742, U.S.A. (email: huytho@umd.edu).

R. Gopalan is with the Department of Video and Multimedia Technologies Research, AT&T Labs - Research, Middletown, NJ 07748 U.S.A. (email: raghuram@research.att.com).

I. INTRODUCTION

Face recognition has been one of the most active research topics in computer vision for several years. It has many applications in biometrics, law enforcement, surveillance and telecommunications [69]. However, face recognition is still a very difficult problem due to appearance variations between the probe and gallery images caused by multiple factors such as blur, expression, illumination, pose and resolution. As a result, face classifiers trained with the assumption that the training and testing data are drawn from similar distributions usually have very poor performance, especially when applied to uncontrolled environments. For instance, face recognition algorithms trained on samples from a source domain containing sharp, well-illuminated face images do not often perform well when used on a target domain containing blurred, poorly-illuminated face images [61]. The performance of these algorithms further degrades when only a limited number of images per subject is available due to the cost and other challenges in data acquisition.

While there have been several studies addressing pre-specified facial variations across source and target domains [69], such as the nine points of light study for illumination [31], analyzing domain shifts caused by multiple, unknown factors has not received much attention. Domain adaptation is a recent paradigm for addressing such transformations in a broader setting, where given labeled data from the source domain and few (resp. no) labeled data from target domain probe images, semi-supervised (resp. unsupervised) approaches have been devised to account for variations in data across domains [5], [54], [17]. Most of these techniques address domain shifts in a statistical sense as models causing variations in data are not known. This limits their application to the particular problem of face recognition where there is a rich literature on models for pose, lighting, blur, expression and aging. As a result, it is important to understand domain shifts with respect to the underlying constraints pertaining to models that generate the observed data. Such an analysis would necessitate the study of geometrical properties of the image space induced by these models.

Many traditional approaches, however, often either ignore the geometric structures of the space or naively treat the space as Euclidean [38]. While non-linear manifold learning algorithms such as ISOMAP [60] or Locally Linear Embedding (LLE) [53] offer alternatives, they require large amounts of training data to estimate the underlying non-linear manifold structure of the problem. Such a requirement on data may not always be satisfied in many real-world applications. One possible solution for handling facial variations due to multiple factors is by employing a mathematical framework called multilinear algebra - the algebra of higher-order tensors. As matrices represent linear operators over a vector space, their

generalization, tensors, define multilinear operators over a set of vector spaces [63]. While there have been studies using multilinear algebraic framework for face recognition [63], [64], such approaches ignore the curved geometry of the image space and resort to an Euclidean treatment. Attempts to incorporate non-linear geometrical structures into the tensor computing framework have been reported in [40], [50], [49], but they again need large training data.

We present a domain adaptive solution for face recognition using the tensor geometry corresponding to models explaining facial variations, with as few as a single image per subject in the source domain. Instead of finding linear transformations representing the shift across domains as in [54], [29], we propose a model-driven approach to construct a latent domain where multifactor facial variations across the source and target domains can be captured together. One main advantage of such an approach is even if data within the source domain and/or the target domain is heterogeneous, for instance when the domain shift is due to blur and both source and target data contain a mix of sharp and blurred faces, the process of accounting for domain shift remains unaltered unlike other techniques that expect the domains to be more or less homogeneous [54], [29], [17]. Furthermore, the proposed method overcomes the data requirement constraint for modeling domain variations by synthesizing multiple face images under different illumination, blur and 2D alignment from a *single* input image on the source or target domain, and uses them to formulate a multidimensional tensor unlike other methods like [40] that places more stringent data-requirement constraints. The tensor obtained from the set of synthesized images can then be represented on a product manifold by performing Higher-Order Singular Value Decomposition (HOSVD) and mapping each orthogonal factored matrix to a point on a Grassmann manifold. The order of the tensors is the number of factors used in the synthesis process. We then recognize the target domain face labels by performing computations pertaining to the tensor geometry for cases where the source domain either contains only one image per subject, or has multiple images per subject. We also address the problem of image set matching which is relevant to video-based face recognition where multiple frames in an video provide evidence related to the facial identity. An illustration of the proposed approach is shown in Figure 1.

Contributions:

- We propose a model-driven domain adaptation approach for face recognition with multiple factor variations, using multilinear algebraic principles. Unlike many other methods that require a large training set, the proposed algorithm uses as little as one face image per subject to characterize the underlying geometry of the latent domain as a product of Grassmannian manifolds.
- We then introduce a novel kernel derived from the projection metric on product spaces. When there

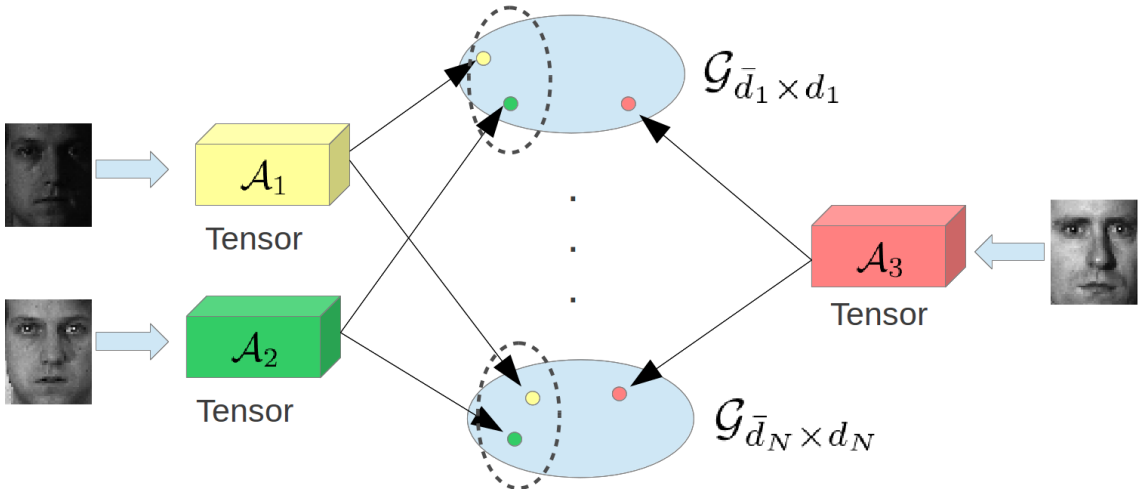


Fig. 1: An illustration of the approach. Face images from different domains are mapped to a latent domain using the multifactor analysis framework. First, a tensor $\mathcal{A}_i$ is obtained from each face image by synthesizing it under multifactor variations. The tensors are then mapped to a product manifold, the collection of $\mathcal{G}_{\bar{d}_j \times d_j}$'s ($j = 1, \ldots, N$), that acts as a latent domain. Subsequent computations are performed in the latent domain using geometric and statistical tools with which the identity of target domain faces are inferred. *(This figure is best viewed in color.)*

are sufficient samples available for each subject in the source domain, this projection kernel can be employed to extend any kernelized learning algorithms to product manifolds, which enables us to account for facial variations such as 3D pose and expression that are not explicitly modeled.

- We also present a probabilistic approach for performing image set classification. The classification is carried out in the projection space using the Kullback-Leibler divergence as a distance measure.

Organization of the paper: Section II discusses related works. The formulation of the proposed approach is given in Section III, along with an introduction to related mathematical details. Details about computations on product manifolds for performing face recognition are presented in Section IV. Section V focuses on the synthesis of face images under multiple factor variations. Experimental results for constrained and unconstrained face recognition on still-images as well as video datasets are provided in Section VI. Section VII concludes the paper.

II. Related Works

This section summarizes some previous works on domain adaptation as well as tensor and manifold learning that are relevant to the proposed method. More comprehensive surveys on general face recognition

as well as the use of matrix manifolds in computer vision are available from [69] and [38], respectively.

With face recognition making a gradual transition from constrained acquisition scenarios that were prevalent until early 2000's, to the more recent unconstrained real-world settings, we are faced with the challenging problem of accounting for multiple facial variations across the source domain training data and the target domain testing data. Domain adaptation is one promising methodology for addressing such issues. While first investigated by the natural language processing community [12], adaptation in the context of visual object recognition has been receiving attention over the last three years. For instance, Saenko *et al.* [54] proposed a semi-supervised approach that leverages partially labeled data from the target domain to learn a domain shifting transformation on the labeled source domain data using metric learning. Kulis *et al.* [29] extended this work to handle asymmetric transformation across the source and target domains. Hoffman *et al.* [20] addressed multi-domain adaptation by using a hierarchical clustering type approach to select domains that are most informative to perform recognition. Unsupervised adaptation, where there is no availability of labels from the target domain, was addressed by Gopalan *et. al.* [17] through an incremental approach based on Grassmann manifold interpretation that could handle both single and multi-domain adaptation. Gong *et al.* [16] extended this approach by proposing an elegant solution to learn incremental information along the manifold by formulating a geodesic flow kernel. Independent of [16], a similar extension was developed by Zheng *et al.* [70]. Subsequently, Shi and Sha [58] proposed an information-theoretical approach for joint learning of domain shift features and classifiers, and Jhuo *et al.* [24] proposed a low-rank, sparsity-driven regularization approach that is robust to noise or outliers. Recent approaches such as [52], [43], [57] attempted to find domain shifts by using dictionary learning and sparse coding. While the above-mentioned techniques address the problem of domain adaptation by learning an appropriate domain shifting transformation, another class of techniques advocate a classifier-based approach that directly seeks to learn a target domain classifier from the classifiers trained on source domain(s) [68], [13], [14].

These techniques perform adaptation in a statistical sense by minimizing data-dependent mismatch in domain properties. Most facial variations, however, result from changes in the image formation mechanisms, and hence it is important to analyze domain shifts for face recognition by taking these imaging models, which often give rise to the notion of manifolds, into account. There have been some attempts in this direction. In order to directly model non-linear image manifolds, many approaches formed a set of synthesized face images from a single face [41], [35], [39], [1], [37]. However, in these methods, the synthesized images were simply generated by 2D perturbations [35], [39] or extracting patches from the original image [1], [37]. As a result, the manifolds constructed by these approaches may not capture

the variations introduced by multiple factors such as illumination, pose or blur. Although the approach in [1] tried to reduce the effect of illumination by performing photometric normalization on the image patches, this factor was not modeled explicitly on the image manifold.

To capture the variations created by multiple factors, the multilinear algebraic framework was introduced into the field of computer vision by Vasilescu and Terzopoulos [63]. In that paper, they proposed an extension of Principal Component Analysis (PCA) [27] called Multilinear PCA (MPCA) or Tensorfaces in order to handle multiple factor variations in face recognition. A kernel extension of the MPCA framework was developed in [34]. However, this approach ignored the curved geometry of the image space as it estimated the distance metric in the Euclidean space.

In order to incorporate the *geometrical* structures of the image space into the multilinear algebraic framework, Lui *et. al.* [40] characterized actions as tensors and mapped them to points on a product manifold for action classification from videos. In [48], Park and Savvides combined MPCA with ISOMAP [60] to preserve the local neighborhood structures. The drawback of this approach is that it required a dense sampling of the training dataset to construct the manifold. To avoid this drawback, the same authors proposed to use a Grassmannian instead of ISOMAP as the manifold representation [50]. However, as only a single Grassmann manifold was employed to model the non-linear structures, this may not capture the complex variations created by multiple factors. Another work by Park and Savvides [49] decomposed the manifold in the data space into factor-dependent sub-manifolds. However, this approach, together with [40] and [50], required multiple images for the manifold learning which may not be practical in the case of limited training samples. As a more systematic alternative, our method formulates a product manifold as a latent domain by analytically characterizing *multiple* factor variations with as few as *one* face image.

III. Problem Formulation

Given an input face image $\mathbf{I}$ of a subject, we analytically characterize domain shifts due to changes in illumination, blur and 2D perturbations of face images in different domains. First, the face is illuminated using the albedo estimated by applying the method of [6] and the universal configuration of lighting directions presented in [31]. The span of these relighted images approximates the subspace of illumination variation for this subject. Each relighted image is then blurred by convolving with a complete set of orthonormal basis functions in order to obtain a blur-invariant representation [18]. The relighted and blurred images are further perturbed by applying 2D similarity transformations in order to characterize the registration manifold. The set of synthesized images obtained after the last step are represented by a

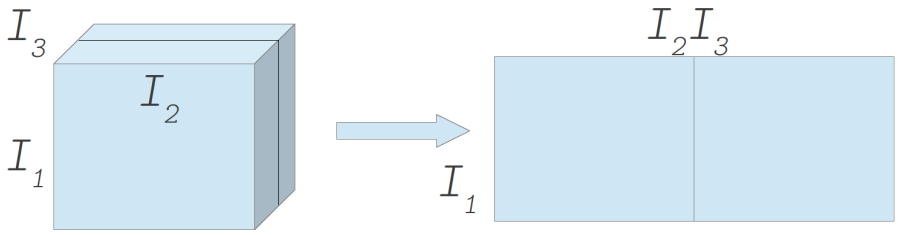


Fig. 2: Mode-1 flattening of a 3rd-order tensor.

4th-order tensor $\mathcal{A} \in \mathcal{R}^{d_1 \times d_2 \times d_3 \times d_4}$, where d_1 is the number of pixels in the face, d_2 is the number of light sources used for relighting, d_3 is the number of orthonormal basis vectors used to get the blur-invariant representation, and d_4 is the number of 2D similarity transformations. As a result of applying HOSVD on the tensor, we obtain a set of orthogonal matrices, where each matrix represents a variation factor and can be handled as a linear term. The tensor is then mapped to a point on a product of Grassmann manifolds using these orthogonal matrices. This product manifold acts as a latent domain for comparing projected data points from different domains. For instance, if only one sample is available per subject in the source domain, recognition of target domain faces is performed using the geodesic distance on the product manifold. In the case when there are multiple source domain samples per subject, a novel kernel on product spaces is proposed. This kernel can be used with any kernelized learning techniques in order to capture other variations such as 3D pose or expression that are not explicitly modeled.

Next, we will provide a brief review of the background mathematics used in the paper regarding tensors and how to represent tensors on product manifolds.

A. Tensors and Tensor Decomposition

Tensors are the natural generalization of matrices to multidimensional spaces. Let $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ be an N-order tensor, an element of $\mathcal{A}$ is denoted as $\mathcal{A}_{i_1 \ldots i_n \ldots i_N}$. The mode-$n$ *flattening* (or *unfolding*) of $\mathcal{A}$ maps the tensor to a 2D matrix $\mathbf{A}_{(n)} \in \mathbb{R}^{d_n \times \bar{d}_n}$ where $\bar{d}_n = d_1 \times \ldots \times d_{n-1} \times d_{n+1} \times \ldots \times d_N$. Each column vector of $\mathbf{A}_{(n)}$ is obtained by varying the n-th index i_n of $\mathcal{A}$ while keeping the other indices fixed. An example of mode-1 flattening of a 3rd-order tensor is shown in Figure 2.

Another important operation on tensors that is worth mentioning is the mode-n product of a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_n \times \ldots \times d_N}$ and a 2D matrix $\mathbf{M} \in \mathbb{R}^{l_n \times d_n}$. The product, denoted by $\mathcal{A} \times_n \mathbf{M}$, returns a tensor $\mathcal{B} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_{n-1} \times l_n \times d_{n+1} \times \ldots \times d_N}$ which can be computed in terms of flattened matrices as

$$\mathbf{B}_{(n)} = \mathbf{M}\mathbf{A}_{(n)}. \tag{1}$$

Similar to Singular Value Decomposition (SVD) for matrices, a tensor can be factorized using an extension of SVD, called HOSVD [40], as

$$\mathcal{A} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times \ldots \times_N \mathbf{U}_N, \tag{2}$$

where $\mathcal{Z} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is the core tensor, $\mathbf{U}_n \in \mathbb{R}^{d_n \times d_n}$, $1 \leq n \leq N$, are the mode-n orthogonal matrices spanning the column space of $\mathbf{A}_{(n)}$. $\mathbf{U}_n$ can be obtained by performing SVD on $\mathbf{A}_{(n)}$

$$\mathbf{A}_{(n)} = \mathbf{U}_n \Sigma_n \mathbf{V}_n^\top, \tag{3}$$

where $\Sigma_n \in \mathbb{R}^{d_n \times \bar{d}_n}$ is a rectangular diagonal matrix of singular values of $\mathcal{A}_{(n)}$, and $\mathbf{V}_n \in \mathbb{R}^{\bar{d}_n \times \bar{d}_n}$ is an orthogonal matrix spanning the row space of $\mathcal{A}_{(n)}$.

The core tensor $\mathcal{Z}$ captures the interaction between the mode matrices $\mathbf{U}_1, \ldots, \mathbf{U}_N$. It is analogous to the diagonal singular value matrix in conventional SVD. However, it is worth noting that $\mathcal{Z}$ does not have the diagonal structure [63].

B. Grassmann Manifolds

Given an n-dimensional real vector space $\mathcal{V}$, the *Grassmann manifold* (or simply *Grassmannian*) $\mathcal{G}_d(\mathcal{V})$ (with $0 \leq d \leq n$) is a set of all d-dimensional linear subspaces of $\mathcal{V}$ [15]. In the special case where $\mathcal{V} = \mathbb{R}^n$, the Grassmannian $\mathcal{G}_d(\mathbb{R}^n)$ is denoted as $\mathcal{G}_{n,d}$. Each point on $\mathcal{G}_{n,d}$ represents a subspace spanned by the column space of an $n \times d$ orthogonal matrix. Thus, all orthogonal matrices $\mathbf{Y} \in \mathbb{R}^{n \times d}$ spanning the same linear subspace are considered equivalent, i.e.

$$\lfloor \mathbf{Y} \rfloor = \{\mathbf{YR} | \mathbf{R} \in O(d)\}, \tag{4}$$

where $O(d) = \{\mathbf{R} \in \mathbb{R}^{d \times d} | \mathbf{R}^\top \mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}_d\}$ is the orthogonal group.

We use the projection metric [15] as the measure of the geodesic distance between two points on a Grassmann manifold. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top$ be the principal angles between the two linear subspaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$, the geodesic distance based on the projection metric is computed as

$$d_c(\mathcal{Y}_1, \mathcal{Y}_2) = ||sin(\boldsymbol{\theta})||_2, \tag{5}$$

where $sin(\boldsymbol{\theta})$ is the vector of sines of the principal angles.

If $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{R}^{n \times d}$ are the orthogonal basis of $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively, the principal angles between the subspaces can be computed numerically by performing SVD on $\mathbf{Y}_1^\top \mathbf{Y}_2$ [7]. The singular values of this SVD are the cosines of the principal angles.

The projection metric can be understood as the Euclidean distance in $\mathbb{R}^{n\times n}$ by defining an embedding $\Psi_P(\mathcal{G}_{n,d})$ as

$$\Psi_P : \mathcal{G}_{n,d} \to \mathbb{R}^{n\times n}, \quad span(\mathbf{Y}) \mapsto \mathbf{Y}\mathbf{Y}^\top. \tag{6}$$

Thus, the corresponding inner product or projection kernel of the space can be obtained as

$$k_P(\mathbf{Y}_1, \mathbf{Y}_2) = tr\left[(\mathbf{Y}_1\mathbf{Y}_1^\top)(\mathbf{Y}_2\mathbf{Y}_2^\top)\right] = ||\mathbf{Y}_1^\top\mathbf{Y}_2||_F^2, \tag{7}$$

where tr is the matrix trace operator and $||.||_F$ is the Frobenius norm. As $k_P(\mathbf{Y}_1, \mathbf{Y}_2) = k_P(\mathbf{Y}_1\mathbf{R}_1, \mathbf{Y}_1\mathbf{R}_2)$ for any $\mathbf{R}_1, \mathbf{R}_2 \in O(d)$, this kernel is well defined. The proof that $k_P(\mathbf{Y}_1, \mathbf{Y}_2)$ is positive definite is given in [19].

C. Representing Tensors on Product Manifolds

As a result of performing SVD on the flattening matrix $\mathbf{A}_{(n)}$, we obtain two orthogonal matrices $\mathbf{U}_n$ and $\mathbf{V}_n$. The reason for not choosing $\mathbf{U}_n$ to represent the geometry of the tensor is that each $\mathbf{U}_n$ is a point on a special orthogonal group $SO(d_n)$. The geodesic distance on $SO(d_n)$ cannot be obtained as a closed form. Furthermore, if points on $SO(d_n)$ are mapped to a Grassmann manifold, the geodesic distance would always be zero [40].

The matrix $\mathbf{V}_n$ in (3) spans the row space of $\mathcal{A}_{(n)}$. As it is usually the case that $d_n < \bar{d}_n$, where $\bar{d}_n$ is defined in Section III-A, $\mathbf{V}_n$ can be substituted by an $\bar{d}_n \times d_n$ orthogonal matrix $\tilde{\mathbf{V}}_n$ by selecting the columns of $\mathbf{V}_n$ corresponding to the non-zeros singular values. Hence, the tensor $\mathcal{A}$ can be represented geometrically as a Cartesian product of the mappings of each $\tilde{\mathbf{V}}_n$ to a point on the Grassmann (factor) manifold $\mathcal{G}_{\bar{d}_n,d_n}$. Furthermore, it is known that the Cartesian product $\mathcal{M} = \mathcal{G}_{\bar{d}_1,d_1} \times \ldots \times \mathcal{G}_{\bar{d}_N,d_N}$ is also a smooth manifold with the manifold topology equivalent to the product topology [30]. Thus, the tensor $\mathcal{A}$ can be represented as a point on this product manifold.

IV. Computations on Product Manifolds

To account for domain shifts in face recognition, we characterize the tensor by synthesizing facial variations due to illumination, blur and 2D alignment from a single face image. While we defer the details on the synthesis process to the next section, here we focus on performing computations on the latent domain, the product of Grassmannians, where tensors corresponding to source domain face images are modeled to infer the identity of tensors derived from target domain faces. More specifically, we first present details on estimating the geodesic distance on product manifolds, which can accommodate cases where the source domain has only one face image per subject. We then derive a positive definite

kernel for product manifolds based on an extension of the projection metric to product spaces. With multiple images per subject in the source domain, this kernel can be used in any kernelized learning algorithm to account for domain shifts due to other factors, such as 3D pose and expression, that are not explicitly synthesized in Section V. As an illustration, by extending the kernel linear discriminant analysis (KLDA) on Grassmann manifolds [19] to product spaces using the proposed kernel, we can find projection directions maximizing inter-class variations (such as due to identities) while minimizing intra-class variations (such as due to pose, expression, occlusion, etc.). Finally, we present a probabilistic approach for performing classification of image sets on product spaces, which caters to video-based face recognition.

A. Geodesics and Projection Kernels on Product Manifolds

The geodesic in the product manifold $\mathcal{M} = \mathcal{G}_{\bar{d}_1,d_1} \times \ldots \times \mathcal{G}_{\bar{d}_N,d_N}$ is the Cartesian product of the geodesics in $\mathcal{G}_{\bar{d}_1,d_1}, \ldots, \mathcal{G}_{\bar{d}_N,d_N}$ [4]. As a result, the geodesic distance based on the projection metric on the product manifold can be estimated as

$$d_c^{\mathcal{M}}(\mathcal{A}^{(1)}, \mathcal{A}^{(2)}) = ||sin(\boldsymbol{\Theta})||_2, \tag{8}$$

where $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ are N-order tensors, and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_N^\top)^\top$ with $\boldsymbol{\theta}_n$ is the vector of principal angles computed on the factor manifold $\mathcal{G}_{\bar{d}_n,d_n}$.

In the case where there are only limited training samples, even with just one sample per subject, the above distance can be used to perform nearest-neighbor classification on the latent domain. As the geodesic distance between two points is the shortest distance on a curved space, it provides a meaningful similarity measure that takes into account the underlying geometry of the latent domain. Next, a positive definite kernel on product manifolds is introduced. When there are sufficient training samples, this kernel function can be used with any kernelized learning technique to statistically account for variations such as 3D pose and expression on the latent domain.

The extension of the embedding in (6) to the product of Grassmann manifolds $\mathcal{M}$ can be written as:

$$\Psi_P^{\mathcal{M}} : \mathcal{G}_{\bar{d}_1,d_1} \times \ldots \times \mathcal{G}_{\bar{d}_N,d_N} \rightarrow \mathbb{R}^{\bar{d}_1 \times \bar{d}_1} \times \ldots \times \mathbb{R}^{\bar{d}_N \times \bar{d}_N},$$
$$(span(\mathbf{Y}_1), \ldots, span(\mathbf{Y}_N)) \mapsto \left(\mathbf{Y}_1\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_N\mathbf{Y}_N^\top\right) \tag{9}$$

Thus, the projection kernel function on the product manifold can be defined as the inner product of this product space:

$$k_P^{\mathcal{M}}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = tr\left[\sum_{i=1}^{N}\left(\mathbf{Y}_i^{(1)}\mathbf{Y}_i^{(1)\top})(\mathbf{Y}_i^{(2)}\mathbf{Y}_i^{(2)\top}\right)\right], \tag{10}$$

where $\mathcal{C}^{(m)} = \left\{\mathbf{Y}_1^{(m)}, \ldots, \mathbf{Y}_N^{(m)}\right\}$, $\mathbf{Y}_i^{(m)} \in \mathbb{R}^{\bar{d}_i \times d_i}$ with $i = 1, \ldots, N$, and $\left(span(\mathbf{Y}_1^{(m)}), \ldots, span(\mathbf{Y}_N^{(m)})\right) \in \mathcal{M}$ with $m = 1, 2$.

This leads to the following proposition for the projection kernel on product manifolds:

Proposition IV.1. *The projection kernel $k_P^{\mathcal{M}}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})$, defined in* (10)*, is a positive definite kernel.*

Proof: For all $\mathbf{Y}_i^{(m)} \in \mathbb{R}^{\bar{d}_i \times d_i}$ with $i = 1, \ldots, N$ and $m = 1, 2$, we have

$$\begin{aligned} k_P^{\mathcal{M}}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) &= tr\left[\sum_{i=1}^{N}\left(\mathbf{Y}_i^{(1)}\mathbf{Y}_i^{(1)\top}\right)\left(\mathbf{Y}_i^{(2)}\mathbf{Y}_i^{(2)\top}\right)\right] \\ &= \sum_{i=1}^{N} tr\left[\left(\mathbf{Y}_i^{(1)}\mathbf{Y}_i^{(1)\top}\right)\left(\mathbf{Y}_i^{(2)}\mathbf{Y}_i^{(2)\top}\right)\right] \\ &= \sum_{i=1}^{N} k_P\left(\mathbf{Y}_i^{(1)}, \mathbf{Y}_i^{(2)}\right) \end{aligned}$$

Thus, $k_P^{\mathcal{M}}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})$ is the sum of the positive definite kernels defined in (7) on each factor manifold. Thus, it is a well-defined and positive definite kernel. ■

The proposed projection kernel between two tensors, $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$, can be computed by setting $\mathbf{Y}_i^{(m)} = \tilde{\mathbf{V}}_i^{(m)}$, for $i = 1, \ldots, N$ and $m = 1, 2$, where $\tilde{\mathbf{V}}_i^{(m)}$ are defined as in Section III-A.

Equipped with the above notation of the projection kernel on product manifolds, we can adapt any kernelized algorithms to perform the learning on the latent domain. In this work, we chose the KLDA algorithm on Grassmann manifolds [19] since its utility for face recognition has been demonstrated before. When there are sufficient training data available, the projection directions computed by using KLDA on product spaces help better separate samples from different people as they maximize inter-class variations due to identities, while minimize intra-class variations due to factors such as pose, expression, occlusion, etc. The extension is straight forward as we only need to replace the kernel function $k_P(\mathbf{Y}_1, \mathbf{Y}_2)$ with $k_P^{\mathcal{M}}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})$. The detailed implementation of KLDA on Grassmannians can be found in [19].

B. Image Set Classification on Product Manifolds

In this section, we present a probabilistic approach to perform domain adaptation for the problem of image set classification. Such a setting occurs naturally in video-based face recognition where several frames in a video sequence are representative of the facial identity. Given a set of images of a subject, it can be characterized as a set of points on a latent domain by projecting the points to a product manifold. For a classification problem with C different subjects, these points can be further mapped to vectors on

a $(C-1)$-dimensional space obtained by performing KLDA on the latent domain using the projection kernel proposed in Section IV-A.

Assume that the distribution of points in the set $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M | \mathbf{x}_i \in \mathbb{R}^{(C-1)}, i = 1, \ldots, M\}$ can be approximated by a multivariate Gaussian distribution $\boldsymbol{\pi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the maximum likelihood estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

Given two sets of points $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ represented by the distributions $\boldsymbol{\pi}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{\pi}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively, a distance measure between the sets can be estimated by using the Kullback-Leibler (KL) divergence, which can be obtained in closed form as [56], [2]:

$$d_{KL}(\boldsymbol{\pi}_1 || \boldsymbol{\pi}_2) = \frac{1}{2}\left(tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - ln\left(\frac{det(\boldsymbol{\Sigma_1})}{det(\boldsymbol{\Sigma_2})}\right) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - (C-1)\right). \quad (11)$$

where $det(\boldsymbol{\Sigma})$ denotes the determinant of $\boldsymbol{\Sigma}$. It is worth noting that the Kullback-Leibler divergence is a positive but non-symmetric measure. As a result, we estimate the KL divergences of the distribution of a probe set from the distributions of all the gallery sets, and select the gallery set corresponding to the minimum distance as the best match.

V. Multifactor Synthesis

Domain shifts caused by variations in factors such as illumination, blur, pose or expression can result in images of the same person having significantly different appearance in different domains. Furthermore, domain shifts can also be caused by localization errors of face detection algorithms when finding the facial bounding boxes and thus, reduce the accuracy of many existing face recognition algorithms. In this section, we discuss how to synthesize faces of the same subject with varying lighting and blur conditions from a single input image. We also present the details of how to characterize the registration manifold [39] using 2D perturbed images in order to account for the in-plane alignment issue. The synthesis process helps to characterizes domain shifts caused by factors such as illumination, blur and 2D alignment without the need for a large training dataset. Domain shifts due to other factors such as 3D pose and expression, that are not explicitly synthesized, are handled by the kernel learning technique on product manifolds presented in the previous section.

A. Illumination

By restricting to convex objects with the Lambertian reflectance model, the diffused component of the surface reflection is given by

$$I_{i,j} = \rho_{i,j} \max(\mathbf{n}_{i,j} \cdot \mathbf{s}, 0), \tag{12}$$

where $I_{i,j}$ is the pixel intensity at position (i,j), $\rho_{i,j}$ and $\mathbf{n}_{i,j}$ are the albedo and surface normal at the corresponding surface point, and $\mathbf{s}$ is the light source direction [6]. From (12), the initial estimate of the albedo $\rho_{i,j}^{(0)}$ can be obtained as

$$\rho_{i,j}^{(0)} = \frac{I_{i,j}}{\mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}}, \tag{13}$$

where $\mathbf{n}_{i,j}^{(0)}$ and $\mathbf{s}^{(\mathbf{0})}$ are the initial values of the surface normal and illuminant direction. The values of $\mathbf{n}_{i,j}^{(0)}$ are obtained from an average 3D face in the USF 3D database [8]. The initial lighting direction $\mathbf{s}^{(\mathbf{0})}$ is estimated using the approach presented in [9].

The initial estimate of the albedo $\rho_{i,j}^{(0)}$ can be related to the true albedo $\rho_{i,j}$ as:

$$\begin{aligned} \rho_{i,j}^{(0)} &= \rho_{i,j} \frac{\mathbf{n}_{i,j} \cdot \mathbf{s}}{\mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}} = \rho_{i,j} + \frac{\mathbf{n}_{i,j} \cdot \mathbf{s} - \mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}}{\mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}} \rho_{i,j} \\ &= \rho_{i,j} + w_{i,j}, \end{aligned} \tag{14}$$

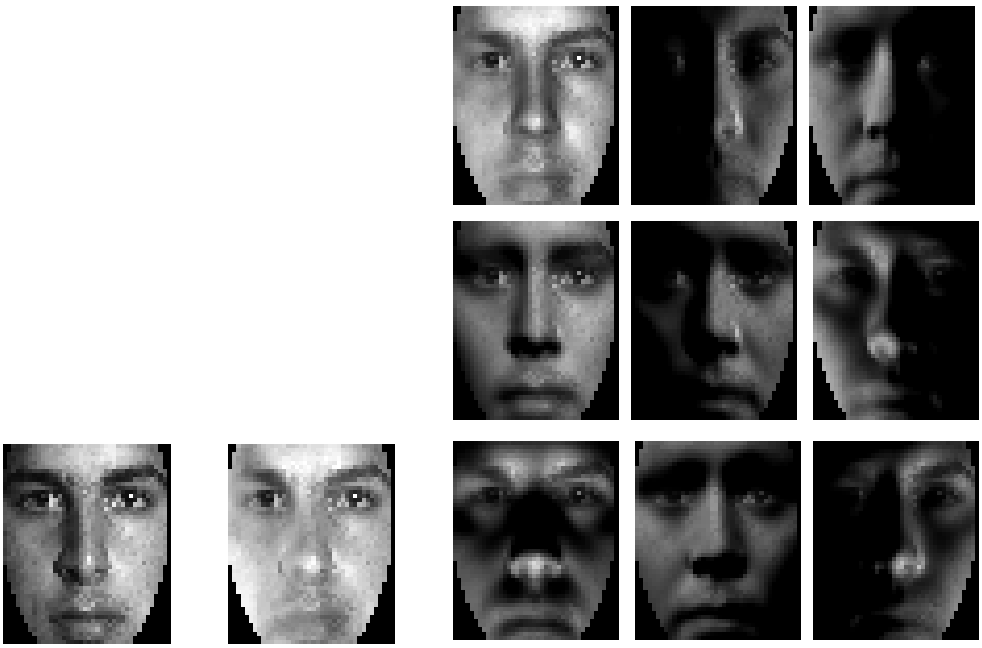
where $w_{i,j} = \frac{\mathbf{n}_{i,j} \cdot \mathbf{s} - \mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}}{\mathbf{n}_{i,j}^{(0)} \cdot \mathbf{s}^{(\mathbf{0})}} \rho_{i,j}$ is the signal-dependent additive noise. $\mathbf{n}_{i,j}$ is the true surface normal at (i,j) and $\mathbf{s}_{i,j}$ is the true lighting direction.

By considering (14) as a signal estimation problem where $\rho_{i,j}$ is the original signal and $\rho_{i,j}^{(0)}$ is the noisy observation, the albedo image can be solved by using the Linear Minimum Mean Square Error (LMMSE) method as in [6]. Figures 3a and 3b shows a face image and its albedo estimated using [6], respectively.

It has been shown that the set of all images of a convex, Lambertian object under different lighting conditions can be approximated by a nine-dimensional linear subspace [3]. This linear subspace can be characterized by illuminating the object using nine pre-specified light sources given in [31]

$$\begin{aligned} \phi &= \{0, 68, 74, 80, 85, 85, 85, 85, 51\}^\circ \\ \theta &= \{0, -90, 108, 52, -42, -137, 146, -4, 67\}^\circ. \end{aligned}$$

where ϕ and θ denote the azimuth and elevation angles, respectively. Figure 3c shows nine images of the same person illuminated by lights from the above configuration. The face image of this person at an



(a) (b) (c)

Fig. 3: From left to right : (a) input face image, (b) the albedo estimated using [6], and (c) images of the same person illuminated by using nine different light sources .

arbitrary illumination condition can be written as a linear combination of these nine basis images

$$\mathbf{I} = \sum_{i=1}^{9} \alpha_i \mathbf{I}_i. \tag{15}$$

As a result, given a single face image, we can estimate the albedo and relight the face at the nine different light sources using the approach in [6] in order to approximate the subspace of illumination variations of this person.

B. Blur

The blurring process can be modeled by the image formulation equation as [18]

$$\tilde{\mathbf{I}} = \mathbf{I} * \mathbf{k} + \boldsymbol{\eta}, \tag{16}$$

where $*$ denotes the 2D convolution between a clean image $\mathbf{I}_{(n_1 \times n_2)}$ and an unknown blur Point Spread Function (PSF) $\mathbf{k}_{(b_1 \times b_2)}$. $\tilde{\mathbf{I}}_{(n_1 \times n_2)}$ is the blurred image and $\boldsymbol{\eta}_{(n_1 \times n_2)}$ represents the noise introduced by the system (i.e. quantization or other sensor induced errors).

It can be seen that, given $\{\boldsymbol{\phi}_i\}_{i=1}^{K}$ as a complete set of orthonormal basis functions for $\mathbb{R}^{b_1 \times b_2}$ with $K = b_1 \times b_2$, any square-integrable, shift-invariant kernel $\mathbf{k}_{(b_1 \times b_2)}$ can be written as

$$\mathbf{k} = \sum_{i=1}^{K} \alpha_i \boldsymbol{\phi}_i, \tag{17}$$

where $\{\alpha_i\}_{i=1}^{N}$ are the combining coefficients. Without noise, (16) can be rewritten as

$$\tilde{\mathbf{I}} = \mathbf{I} * \sum_{i=1}^{K} \alpha_i \boldsymbol{\phi}_i = \sum_{i=1}^{K} \alpha_i (\mathbf{I} * \boldsymbol{\phi}_i). \tag{18}$$

Let $\mathbf{D}(\mathbf{I}) = [(\mathbf{I} * \boldsymbol{\phi}_1)^v \quad (\mathbf{I} * \boldsymbol{\phi}_2)^v \ldots (\mathbf{I} * \boldsymbol{\phi}_K)^v]$ be a dictionary of size $d \times K$, where $d = n_1 \times n_2$ with $d > K$, and $(.)^v$ denotes the vectorization operation. The column span of $\mathbf{D}(\mathbf{I})$, i.e. $span(\mathbf{D}(\mathbf{I})) = \{\mathbf{I} * \mathbf{k} | \mathbf{k} \in \mathbb{R}^{b_1 \times b_2}\}$, is a subspace containing the set of convolutions of $\mathbf{I}$ with arbitrary kernels of maximum size $b_1 \times b_2$. Under certain assumptions, the $span(\mathbf{D}(\mathbf{I}))$ allows us to obtain a representation of the image $\mathbf{I}$ that is invariant to blurring with an arbitrary $\mathbf{k}$.

Proposition V.1. *Under three assumptions: (i) there is no noise in the system ($\boldsymbol{\eta} = 0$), (ii) the maximum size of the blur kernel $b_1 \times b_2 = K$ is known, and (iii) the $K \times K$ Block-Toeplitz-Toeplitz-Block (BTTB) matrix corresponding to the unknown blur PSF, under zero boundary conditions for convolution, is full rank, $span(\mathbf{D}(\mathbf{I}))$ is a blur-invariant of $\mathbf{I}$. In other words, $span(\mathbf{D}(\mathbf{I})) = span(\mathbf{D}(\tilde{\mathbf{I}}))$, where $\tilde{\mathbf{I}}$ is a blurred version of $\mathbf{I}$.*

Proof: See the proof of Proposition 2.1 in [18]. ■

The main advantage of this representation is that there are no constraints on the shape of the blur kernels that can be handled, as long as the blur kernels satisfy the above assumptions. In the paper, all the face images are resized to 40×48 and the maximum kernel size is set at $b_1 \times b_2 = 7 \times 7$. As a result, for each relighted face image, a set of $K = 49$ basis vectors is obtained by convolving the image with $\{\boldsymbol{\phi}_i\}_{i=1}^{49}$. In our experiments, the set of basis vectors $\{\boldsymbol{\phi}_i\}_{i=1}^{49}$ is selected as the columns of a 49×49 identity matrix.

C. 2D Registration

In practice, it may be unrealistic to expect that a face detection system can locate faces with many appearance variations with high precision. Thus, the extracted bounding boxes of faces with varying illumination or blur conditions may not align perfectly. In order to account for the alignment errors during the face localization process, a set of perturbed images using 2D similarity transformations is obtained for each face image in order to characterize the registration subspace [39].

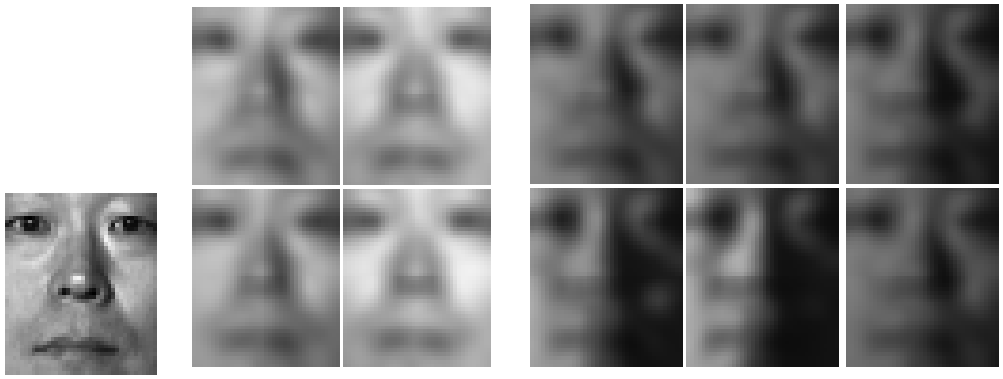
A similarity transformation mapping an image coordinate (i, j) to the new coordinate (u, v) can be written in the homogeneous form as

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} i \\ j \\ 1 \end{pmatrix} \tag{19}$$

where θ and s are the rotation and isometric scaling parameters, respectively. t_x and t_y are the translation parameters. In our experiments, we set the values of θ as $\{-4, -2, 0, 2, 4\}^\circ$, s as $\{0.9, 0.95, 1, 1.05, 1.1\}$, and t_x and t_y as $\{-3, 0, 3\}$ as they provide a reasonable coverage for possible alignment errors between the probe and gallery images. Bilinear interpolation is employed to sample the transformed images. As a result, a total of two hundreds and twenty five perturbed images are synthesized for each relighted and blurred face image.

VI. Experiments

We first test our approach for face identification where the goal is to estimate the subject label of a probe image, and then for face verification, where given a pair of probe images, the goal is determine if they correspond to the same subject or not. For face identification we consider four public datasets namely, the CMU-PIE [59] and AR [42] datasets that contain still faces captured under constrained settings, the UMD remote face dataset [18] comprising of unconstrained still faces, and the Honda/UCSD video dataset [32]. For face verification we use the recent, unconstrained Labeled Faces in the Wild (LFW) dataset [23]. Most of these datasets contain facial variations that are not explicitly synthesized by our method. We compare our approach with several other techniques that were evaluated on these datasets. It is also worth noting that the existing works on domain adaptation such as [54], [29], [17] may not be applicable to these experimental settings. One of the reasons is that they impose data requirement constraints that are not often satisfied as there may be as little as one image available per individual per the source and target domains. Another reason is that the requirement for the source and target domains to be more or less homogeneous may not hold in unconstrained face recognition as domain shifts can be caused by multiple factor variations such as illumination, blur, expression and alignment. In all these experiments, we use the algorithmic parameters that were discussed in Section V. Given a cropped face image of size 40×48, it takes about 4 seconds to generate the synthesized images and perform the tensor decomposition on an Intel Core i7 computer. It takes less than 0.05 second to estimate the geodesic distance between two tensors on a product manifold using the same machine.



(a) (b) (c)

Fig. 4: Example images of a subject from the CMU-PIE used in the experiments: (a) clear and well-illuminated gallery image, (b) Good Illumination (GI) probe images, and (c) Bad Illumination (BI) probe images. A 7×7 Gaussian kernel with $\sigma = 3$ is used to blur the probe images.

A. CMU-PIE Dataset

First, experimental results on the *illumination subset* of the CMU-PIE dataset [59] are presented. Facial bounding boxes are obtained using the Viola-Jones object detection algorithm [65] without performing any pre-processing alignment step. We apply the same experiment settings as in [61]. The source domain which contains the frontal images (c_{27}) with good illumination (f_{21}) of all 68 subjects is used as our gallery. The target domain containing the remaining frontal images with 10 different illumination conditions is used as probe. The probe set is further divided into two subsets: a) Good Illumination (GI) consisting of f_{09}, f_{11}, f_{12}, and f_{20}, and b) Bad Illumination (BI) consisting of $f_{13}, f_{14}, f_{15}, f_{16}, f_{17}$ and f_{22}. The probe faces are blurred by convolving with Gaussian kernels of $\sigma \in (0.5, 1.0, 1.5, 2, 2.5, 3)$ and size $(2\sigma+1) \times (2\sigma+1)$ for each σ. Figure 4 shows some examples face images from the CMU-PIE database used in the experiments.

We compare the proposed method with the algorithms discussed in [47], [44], [18], and [61]. The Local Phase Quantization (LPQ) method in [47] utilized phase information computed locally for every image position in order to perform blur insensitive face classification. On the other hand, Nishiyama *et al.* [44] proposed a method called FAcial DEblur INference (FADEIN) that attempted to infer a PSF representing the process of blur on faces. Gopalan *et al.* [18] performed blur robust face recognition by comparing subspaces created from a clean image and its blurred version on the Grassmann manifold. The Illumination-Robust Recognition of Blurred Faces (rIRBF) algorithm [61] handled blur and illumination

variations in face recognition by comparing bi-convex sets formulated from face images at different blur and illumination conditions. Recognition rates of different approaches across domain shifts caused by illumination and (synthetic) blur variations on the CMU-PIE dataset are shown in Table I. $\sigma = 0$ means that the recognition rates are obtained with only illumination variations and without blurring the probe faces. As there is only a single gallery image per subject, the nearest-neighbor classification based on the geodesic distance on the latent domain is used in our approach. It can be seen from the figure that our method achieves consistently higher recognition rates compared to other algorithms in all combinations of illumination and blur. When the size of the blur kernel increases, the performance of all algorithms decreases. However, even at the worst scenario (kernel size of 7×7 at $\sigma = 3$, bad illumination), the proposed method still achieves the highest recognition rate at 92.4%, which is 11% higher than the next best result obtained by [61]. In the case of bad illumination and blur, the assumptions on the Lambertian model and quantization noise used in obtaining the synthesized blurred images may be violated, and thus lead to the reduction in the performance of our algorithm.

We also compare our algorithm with the results obtained by applying the Euclidean nearest-neighbor (NN) classification based on ℓ_2 norm directly on synthesized (blurred, relighted and transformed) images from the training set. It is clear from Table I that the performance of the method based on direct NN on synthesized images degrades by a large margin when the blur kernel size increases and is significantly lower than the recognition rates obtained by our algorithm. This can be explained by noting that the direct NN method only searches for the closest discrete point in the image space rather than modeling domain shifts due to multiple factor variations as in our approach.

Recognition results using the proposed approach without synthesizing images at different illuminations are also included. It can be seen from Table Ia that when the lighting component is held out, the performance of the proposed method remains approximately the same with good illuminated faces. However, in the case of bad illumination in Table Ib, the recognition rates reduces significantly, especially when the size of the blur kernel is large. This shows the importance of modeling illumination variations in our approach when the lighting condition is bad.

B. AR Dataset

In this section, experimental results for face images in the AR face dataset [42] are presented, which contain expression variations and real occlusions. It is worth mentioning that the proposed approach is not explicitly designed to handle domain shifts due to occlusion and expression. Hence, this offers a test case to analyze the robustness of our method to variations that are not synthesized.

TABLE I: Recognition rates (in %) of different approaches across illumination and (synthetic) blur variations on the CMU-PIE dataset. σ is the standard deviation of the Gaussian kernel used for blurring. The results for [47], [44] and [61] are obtained from [61].

(a) Good Illumination (f_{09}, f_{11}, f_{12} and f_{20})

σ	0	0.5	1	1.5	2	2.5	3
LPQ [47]	99.63	99.63	99.63	99.63	97.05	79.42	46.32
FADEIN [44] + LPQ	98.53	95.6	93.6	91.2	89.8	88.60	87.13
Grassmannian [18]	99.63	99.63	99.63	99.63	99.63	96.32	93.38
rIBRF [61]	99.7	99.7	99.63	99.63	99.63	99.63	97.45
NN_2 with synthesized images	95.59	93.75	91.91	91.91	77.2	58.82	52.21
Our approach (without illumination)	100	100	100	99.63	99.63	99.26	98.53
Our approach	100	100	100	100	100	100	99.26

(b) Bad Illumination ($f_{13}, f_{14}, f_{15}, f_{16}, f_{17}$ and f_{22})

σ	0	0.5	1	1.5	2	2.5	3
LPQ [47]	99.1	97.79	96.08	88.97	73.04	58.08	27.7
FADEIN [44] + LPQ	91.5	87.7	81.8	69.11	62.74	56.37	44.61
Grassmannian [18]	85.71	84.66	84.24	79.2	71.01	67.23	60.92
rIBRF [61]	95.1	92.7	92.7	91.6	88.2	84.78	81.36
NN_2 with synthesized images	92.89	86.27	81.37	68.87	66.42	54.65	35.05
Our approach (without illumination)	98.77	98.77	98.77	96.08	96.08	92.65	88.48
Our approach	100	100	100	99.26	98.77	96.64	92.4

(a) (b) (c) (d) (e) (f)

Fig. 5: Six facial images of a subject from the first session in the AR dataset [42]. The faces are detected and cropped using the OpenCV implementation of the Viola-Jones object detection algorithm [65].

The AR face database contains frontal images of more than 100 individuals taken over two sessions separated by two weeks time. Following the experimental setups in [25], [26], a total of 12 images per person are used in the experiments. Face detection is also performed using the Viola-Jones object

TABLE II: Recognition rates (%) with real occlusion on the AR dataset for a variety of training and testing sets. Results for other methods were obtained from [26]

Training Set	Testing Set	PSVM [26]	PWCM [25]	NN_2	NN_1	**Our approach**
$[e, f]$	$[a]$	96.0	89.0	45.0	79.0	91.0
$[e, f]$	$[a']$	79.4	71.0	31.0	50.0	76.0
$[e, f]$	$[b, c, d]$	80.0	72.0	31.7	59.7	66.3
$[e, f]$	$[b', c', d']$	58.7	47.3	20.3	32.7	52.7
$[e, f]$	$[e', f']$	57.0	55.0	25.5	29.0	66.5
$[e, f, e', f']$	$[b, c, d, b', c', d']$	86.6	76.2	31.3	56.5	79.8
$[e, f, e', f']$	$[a, a']$	96.4	95.0	48.5	83.0	95.0

detection algorithm. However, unlike [25], [26], we do not perform any facial alignment step and instead, use the bounding boxes returned by the face detection algorithm directly in the recognition. The detected faces from six images of an individual in the first session of the AR dataset are shown in Figure 5. They are labeled a through f, and the corresponding images in the second session are labeled a' through f'. In addition to occlusions, there are also expression variations between the images.

Table II compares the recognition rates for different approaches on the AR dataset with a variety of training and testing sets. These are very challenging experiments as in many cases, there are approximately 50% occlusions in both the training and testing images. As a large part of the face images is occluded, the albedo cannot be reliably estimated and thus, we do not perform the synthesis for illumination. Furthermore, because the number of training samples is limited, the nearest-neighbor classification based on the geodesic distance is employed in our approach. Recognition rates using the simple Euclidean nearest-neighbor classification based on the ℓ_2 and ℓ_1 norms, NN_2 and NN_1, are also included. We also compare our algorithm with two methods, Partial Within-Class Match (PWCM) [25] and Partial Support Vector Machines (PSVM) [26], that are specifically designed to handle occlusion in face recognition. PWCM performs classification by reconstructing a test sample as a linear combination of the training samples from each class. The reconstruction is solely based on the visible data in the face images. On the other hand, PSVM extends SVM to handle occlusion by deriving a criterion that can handle the case of missing entries in the feature vectors.

It can be seen from the table that our method significantly outperforms both the Euclidean nearest-neighbor classification algorithms. Although our results are not as good as the ones obtained by PSVM, it is encouraging to see that the proposed algorithm is better than PWCM in most cases. Especially in

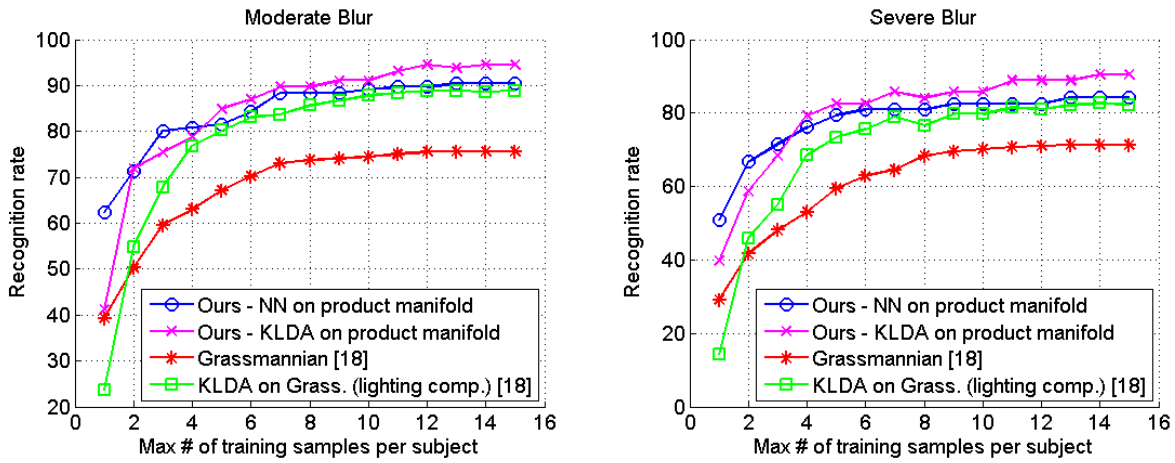
Fig. 6: Example images of six subjects from the UMD remote face dataset. First row: source domain containing clean face images. Second row: target domain containing moderately blurred face images. Third row: target domain containing severely blurred face images.

the case where there are occlusions in both the training ($[e, f]$) and testing sets ($[e', f']$), the proposed algorithm outperforms both PSVM and PWCM by a large margin (66.5% compared to 57.0% and 55.0%, respectively). These experiments show that our method is robust to domain shifts caused by variation factors such as occlusion and expression even if they are not explicitly modeled.

C. UMD Remote Face Dataset

We then present recognition results on the UMD remote face dataset using the same data partitioning reported in [18]. This is an unconstrained dataset used for surveillance consisting of cropped faces of 17 subjects. In addition to moderate-to-severe blur, face images in the dataset also contain moderate variations in other factors such as pose, illumination, expression and occlusion. In this experiment, the source domain contains face images with variations such as illumination, occlusion and pose but without much blur. The target domains contain faces with moderate and severe amount of blur as well as other variations. Figure 6 shows some example images of the UMD remote face dataset with respect to different domains.

The comparisons between our approach and the method discussed in [18] are shown in Figure 7.



(a) (b)

Fig. 7: Recognition rates for moderate and severe blurred probe images on the UMD remote face dataset. (*This figure is best viewed in color*).

It can be seen that our approach achieves better results for both moderate and severe blur conditions regardless of the number of training samples. When only a single training image per subject is available, our approach based on the nearest-neighbor classification using geodesic distance on the latent domain still obtains significantly higher recognition rates compared to [18], for both moderate and severe blur cases. This is a result of accounting for domain shift due to not only blur variations as in [18], but also for illumination and 2D alignment in our approach. The performance of the proposed KLDA algorithm on the latent domain increased significantly when more data are used in the training. The main reason is that it is able to learn the structure of the image space better by capturing domain shift due to other factors such as 3D pose and expression that are not explicitly modeled.

D. Honda/UCSD Video Dataset

Experiments on face recognition from videos were also conducted on the Honda/UCSD dataset [32]. This dataset contains 59 videos sequences of 20 different subjects. The number of frames in each video sequence varies from 12 to 645. Variations in illumination, pose, occlusion and expression appear across different sequences of each subject. The faces are also detected and cropped from the video frames using the Viola-Jones algorithm.

The proposed approach is compared with different algorithms such as [28], [66], [10], [21], [11]. Kim *et al.* [28] presented a discriminative learning method based on canonical correlations (DCC) and applied

TABLE III: Recognition rates (%) on the Honda/UCSD dataset for different values of the maximum set length. The results for other methods were obtained from [11].

Set Length	DCC [28]	MDA [66]	AHISD [10]	CHISD [10]	SANP [21]	DFRV [11]	**Our approach**
50 frames	76.92	74.36	87.18	82.05	84.62	89.74	**97.44**
100 frames	84.62	94.87	84.62	84.62	92.31	**97.44**	**97.44**
Full Length	94.87	97.44	89.74	92.31	**100**	97.44	**100**
Average	85.47	88.89	87.18	86.33	92.31	94.87	**98.29**

it to image set classification. Another discriminative learning technique proposed Wang and Chen, called Manifold Discriminant Analysis (MDC) [66], modeled each image set as a manifold and tried to find an embedded space to better separate manifolds from different classes. On the other hand, Cevikalp and Triggs [10] developed methods called AHISD (Affine Hull based Image Set Distance) and CHISD (Convex Hull based Image Set Distance) that characterized each image set by a convex geometric region (the affine or convex hull). The Sparse Approximated Nearest Points (SANP) algorithm [21] introduced a between-set distance defined as the nearest distance between sparse approximated points in the two sets. The last method used in the comparison is the Dictionary-based Face Recognition from Video (DFRV) algorithm [11] that extracted joint appearance and behavioral features from facial videos using dictionary learning.

We follow the experiment procedure in [21]: 20 sequences were used for training and the remaining 39 sequences for testing. Table III shows the recognition results obtained by using the algorithm presented in Section IV-B. The set length is the maximum number of cropped face images per video sequence. If the number of images in a sequence is less than the set length, all the images are used for classification. It can be seen from the table that our algorithm consistently outperforms all other methods in the comparison. This shows that when multiple video frames are available, the proposed KLDA on product manifolds is able to find an embedded space that separated face images from different individuals well.

E. Face Verification

In order to apply the proposed approach to face verification, given a pair of face images $\mathbf{I}_1$ and $\mathbf{I}_2$, two 4-th order tensors $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ are formulated from the synthesized images at different illumination, blur and 2D transformations as presented in Section V. The vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_4^\top)^\top$, with $\boldsymbol{\theta}_j$ is the vector of principal angles on the factor manifold $\mathcal{G}_{\bar{d}_j, d_j}$ $(j = 1, \ldots, 4)$, is computed from the pair of

(a) Same (b) Different

Fig. 8: Examples of same and different pairs of face images from the LFW dataset.

tensors and used as the feature in the training and testing based on Support Vector Machines (SVM) [62]. In this experiment, SVM with RBF kernel was used for classifying whether the pair of images was from the same individual or not. The optimal values of the parameters for training SVM are determined by performing cross validation on the training set.

We use the Labeled Faces in the Wild (LFW) [23] dataset for our experiments. It is a challenging dataset containing more than 13000 unconstrained face images from 5749 individuals. Face images in the dataset have large variations in pose, illumination, age, expression, etc. Figure 8 shows examples of same and different pairs of face images from this dataset.

We follow the "image restricted" protocol [23]: only binary "same" or "different" labels are available for pairs of image, the identities of the images are unknown. The performance is measured by using 10-fold cross validation. In order to evaluate the robustness of the proposed algorithm to face cropping quality, we report our results on both unaligned and aligned cropped faces of the dataset [55].

Table IV compares the face verification rates of different approaches on the LFW dataset on this image restricted protocol. The method proposed by Nowak and Jurie [45] used randomized binary trees to quantize the differences between local descriptors sampled from "same" and "different" image pairs. On the other hand, Wolf *et. al.* [67] developed a new patch-based descriptor based on Local Binary Patterns (LBP) [46]. Another method used in the comparison was proposed by Pinto *et. al.* [51] that combined V1-like models and multiple kernel learning (MKL) for face verification. All of these above methods were evaluated on LFW images aligned using an unsupervised technique called funneling [22]. We also include the results obtained using a recent pose-invariant algorithm based on adaptive probabilistic elastic matching (APEM) [33] on both aligned and unaligned images. It can be seen from the table that the proposed approach outperforms other methods used in the comparison, except the

TABLE IV: Performance comparison for different methods on the most restricted LFW. Both mean classification rates and standard errors of the mean are reported.

Method	Accuracy $\pm$ Error (%)
Nowak (unaligned) [45]	72.45 ± 0.40
Nowak (aligned) [45]	73.93 ± 0.49
Hybrid descriptor-based (aligned) [67]	78.47 ± 0.51
V1-like/MKL (aligned) [51]	79.35 ± 0.55
APEM (fusion, unaligned) [33]	81.70 ± 1.78
APEM (fusion, aligned) [33]	84.08 ± 1.20
Our approach (unaligned)	82.67 ± 1.14
Our approach (aligned)	82.94 ± 0.83

APEM on aligned images. This is understandable as our algorithm does not explicitly synthesize pose and expression variations. Furthermore, a data-driven method such as KLDA on product manifolds cannot be applied as only pairs of face images are available without any identity information. However, the result is encouraging as we are able to outperform the APEM method on unaligned images, even though that method combines multiple features such as LBP [46] and SIFT [36] and is designed to handle pose variations. The verification rate obtained using the proposed approach on aligned images is only slightly better than when using unaligned images. This shows that our algorithm is not as sensitive to 2D face alignment as in other approaches, since this factor is explicitly accounted for using 2D perturbations. The Receiver Operating Characteristic (ROC) curves of different approaches are shown in Figure 9 in order to better evaluate their performances. The ROC curve of the proposed algorithm is obtained by thresholding the probability estimates computed using kernel SVM.

VII. Conclusions

We have shown that the underlying geometry of a set of face images of a person under multiple factor variations plays an important role in the recognition of face images from different domains. We showed that such a geometry can be studied by representing this set of images as a tensor and mapping the tensor to a point on a product manifold. The product manifold served as a latent domain where domain shifts due to multifactor variations such as illumination, blur and 2D alignment were jointly modeled. For cases where only a single gallery image per subject was available, geodesic distance was used to perform nearest-neighbor classification on the latent domain. Furthermore, a novel positive definite

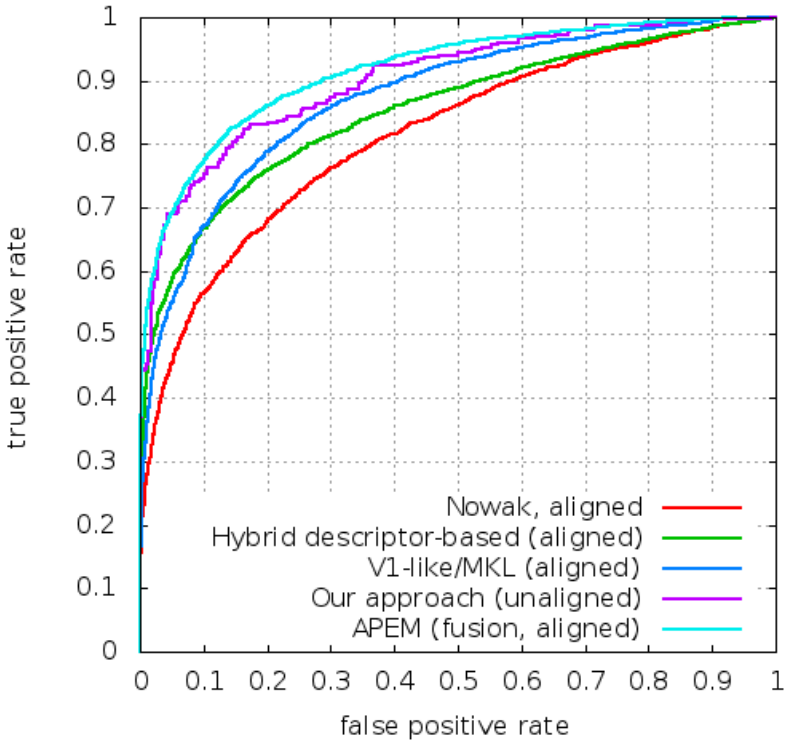


Fig. 9: ROC curves of different approaches on the LFW dataset.

kernel based on an extension of the projection metric to the product space was proposed. When there were sufficient samples available from the source domain, this projection kernel could be employed in any kernelized learning techniques to account for domain shifts due to other facial variations such as 3D pose and expression that were not explicitly modeled. Finally, a probabilistic method for classifying image sets on the latent domain using the KL divergence was also introduced. Competitive experimental results on different datasets showed the effectiveness of the approach in handling domain shifts caused by multifactor variations.

REFERENCES

[1] O. Arandjelović. Unfolding a Face: from Singular to Manifold. In *Proc. ACCV*, pages 203–213, 2009.

[2] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face Recognition with Image Sets using Manifold Density Divergence. In *Proc. CVPR*, pages 581–588, 2005.

[3] R. Basri and D. W. Jacobs. Lambertian Reflectance and Linear Subspaces. *IEEE Trans. PAMI*, 25(2):218–233, 2003.

[4] E. Begelfor and M. Werman. Affine Invariance Revisited. In *Proc. CVPR*, pages 2087–2094, 2006.

[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

[6] S. Biswas, G. Aggarwal, and R. Chellappa. Robust Estimation of Albedo for Illumination-Invariant Matching and Shape Recovery. *IEEE Trans. PAMI*, 31(5):884–899, 2009.

[7] A. Björck and G. H. Golub. Numerical Methods for Computing Angles between Linear Subspaces. *Mathematics of Computations*, pages 579–594, 1973.

[8] V. Blanz and T. Vetter. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH*, pages 187–194, 1999.

[9] M. J. Brooks and B. K. P. Horn. Shape and Source from Shading. In *Proc. IJAI*, pages 932–936, 1985.

[10] H. Cevikalp and B. Triggs. Face Recognition Based on Image Sets. In *Proc. CVPR*, pages 2567–2573, 2010.

[11] Y. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-Based Face Recognition from Video. In *Proc. ECCV*, pages 766–779, 2012.

[12] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126, 2006.

[13] L. Duan, I. Tsang, D. Xu, and T.-S. Chua. Domain Adaptation from Multiple Sources via Auxiliary Classifiers. In *Proc. ICML*, pages 289–296, 2009.

[14] L. Duan, I. Tsang, D. Xu, and T.-S. Chua. Domain Transfer Multiple Kernel Learning. *IEEE Trans. PAMI*, 34(3):465–479, 2012.

[15] A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM J. Matrix Analysis and Applications*, 20:303–353, 1999.

[16] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In *Proc. CVPR*, pages 2066–2073, 2012.

[17] R. Gopalan, R. Li, and R. Chellappa. Domain Adaptation for Object Recognition: An Unsupervised Approach. In *Proc. ICCV*, pages 999–1006, 2011.

[18] R. Gopalan, S. Taheri, P. Turaga, and R. Chellappa. A Blur-Robust Descriptor with Applications to Face Recognition. *IEEE Trans. PAMI*, 34(6):1220–1226, 2012.

[19] J. Hamm and D. Lee. Grassmann Discriminant Analysis: A Unifying View on Subspace-based Learning. In *Proc. ICML*, pages 376–383, 2008.

[20] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering Latent Domains for Multisource Domain Adaptation. In *Proc. ECCV*, pages 702–715, 2012.

[21] Y. Hu, A. S. Mian, and R. Owens. Sparse Approximated Nearest Points for Image Set Classification. In *Proc. CVPR*, pages 27–40, 2011.

[22] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *Proc. ICCV*, 2007.

[23] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[24] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust Visual Domain Adaptation with Low-Rank Reconstruction. In *Proc. CVPR*, pages 2168–2175, 2012.

[25] H. Jia and A. M. Martinez. Face Recognition with Occlusions in the Training and Testing Sets. In *Proc. FG*, pages 1–6, 2008.

[26] H. Jia and A. M. Martinez. Support Vector Machines in Face Recognition with Occlusions. In *Proc. CVPR*, pages 136–141, 2009.

[27] I. T. Joliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[28] T. K. Kim, J. Kittler, and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes using Canonical Correlations. *IEEE Trans. PAMI*, 29(6):1005–1018, 2007.

[29] B. Kulis, K. Saenko, and T. Darrell. What You Saw is Not What You Get: Domain Adaptation using Asymmetric Kernel Transforms. In *Proc. CVPR*, pages 1785–1792, 2011.

[30] J. M. Lee. *Introduction to Topological Manifolds*. Springer, 2 edition, 2010.

[31] K. C. Lee, J. Ho, and D. J. Kriegman. Acquiring Linear Subspaces for Face Recognition under Variable Lighting. *IEEE Trans. PAMI*, 27(5):684–698, 2005.

[32] K. C. Lee, J. Ho, M. H. Yang, and D. J. Kriegman. Visual Tracking and Recognition using Probabilistic Appearance Manifolds. *CVIU*, 99(3):303–331, 2005.

[33] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic Elastic Matching for Pose Invariant Face Recognition. In *Proc. CVPR*, 2013.

[34] Y. Li, Y. Du, and X. Lin. Kernel-based Multifactor Analysis for Image Synthesis and Recognition. In *Proc. ICCV*, pages 114–119, 2005.

[35] J. Liu, S. Chen, Z. Zhou, and X. Tan. Single Image Subspace for Face Recognition. In *Proc. AMFG*, pages 205–219, 2007.

[36] D. J. Lowe. Distinctive Image Features From Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004.

[37] J. Lu, Y. P. Tan, and G. Wang. Discriminative Multi-Manifold Analysis for Face Recognition from a Single Training Sample per Person. In *Proc. ICCV*, pages 1943–1950, 2011.

[38] Y. M. Lui. Advances in Matrix Manifolds for Computer Vision. *Imag. Vis. Comp.*, 30:380–388, 2012.

[39] Y. M. Lui and J. R. Beveridge. Grassmann Registration Manifolds for Face Recognition. In *Proc. ECCV*, volume 2, pages 44–57, 2008.

[40] Y. M. Lui, J. R. Beveridge, and M. Kirby. Action Classifications on Product Manifolds. In *Proc. CVPR*, pages 833–839, 2010.

[41] A. M. Martinez. Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class. *IEEE Trans. PAMI*, 24(6):748–763, 2009.

[42] A. M. Martinez and R. Benavente. The AR Face Database. *CVC Technical Report*, 24, 1998.

[43] J. Ni, Q. Qiu, and R. Chellappa. Subspace Interpolation via Dictionary Learning for Unsupervised Domain Adaptation. In *Proc. CVPR*, 2013.

[44] M. Nishiyama, A. Hadid, H. Takeshima, J. Shotton, T. Kozakaya, and O. Yamaguchi. Facial Deblur Inference using Subspace Analysis for Recognition of Blurred Faces. *IEEE Trans. PAMI*, 33(4):838–845, 2011.

[45] E. Nowak and F. Jurie. Learning Visual Similarities Measures for Comparing Never Seen Objects. In *Proc. CVPR*, 2007.

[46] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. PAMI*, 24(7):971–987, 2002.

[47] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification using Local Phase Quantization. In *Proc. ICISP*, pages 236–243, 2008.

[48] S. W. Park and M. Savvides. An Extension of Multifactor Analysis for Face Recognition based on Submanifold Learning. In *Proc. CVPR*, pages 2645–2652, 2010.

[49] S. W. Park and M. Savvides. Multifactor Analysis based on Factor-Dependent Geometry. In *Proc. CVPR*, pages 2817–2824, 2011.

[50] S. W. Park and M. Savvides. The Multifactor Extension of Grassmann Manifolds for Face Recognition. In *Proc. FG*, pages 464–469, 2011.

[51] N. Pinto, J. J. Dicarlo, and D. D. Cox. How Far Can You Get With a Modern Face Recognition Test Set Using Only Simple Features. In *Proc. CVPR*, 2009.

[52] Q. Qiu, V. Patel, P. Turaga, and R. Chellappa. Domain Adaptive Dictionary Learning. In *Proc. ECCV*, pages 631–645, 2012.

[53] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, pages 2323–2326, 2000.

[54] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting Visual Category Models to New Domains. In *Proc. ECCV*, pages 213–226, 2010.

[55] C. Sanderson and B. C. Lovell. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In *Proc. ICB*, pages 199–208, 2009.

[56] G. Shakhnarovich, J. W. Fisher, and T. Darell. Face Recognition from Long Term Observations. In *Proc. ECCV*, pages 851–868, 2002.

[57] S. Shekhar, V. Patel, H. Nguyen, and R. Chellappa. Generalized Domain-Adaptive Dictionaries. In *Proc. CVPR*, 2013.

[58] Y. Shi and F. Sha. Information-Theoretical Learning of Discriminative clusters for Unsupervised Domain Adaptation. In *Proc. ICML*, 2012.

[59] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Trans. PAMI*, 25(12):1615–1618, 2003.

[60] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Non-Linear Dimensionality Reduction. *Science*, pages 2319–2323, 2000.

[61] P. Vageeswaran, K. Mitra, and R. Chellappa. Blur and Illumination Robust Face Recognition via Set-Theoretic Characterization. *IEEE Trans. Image Processing*, 22(4):1362–1372, 2013.

[62] V. N. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[63] M. A. O. Vasilescu and D. Terzopoulos. Multilinear Analysis of Image Ensembles. In *Proc. ECCV*, pages 447–460, 2002.

[64] M. A. O. Vasilescu and D. Terzopoulos. Multilinear Projection for Appearance-based Recognition in the Tensor Framework. In *Proc. ICCV*, pages 1–8, 2007.

[65] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proc. CVPR*, pages 511–518, 2001.

[66] R. Wang and X. Chen. Manifold Discriminant Analysis. In *Proc. CVPR*, pages 429–436, 2009.

[67] L. Wolf, T. Hassner, and Y. Taigman. Descriptor Based Methods in the Wild. In *Faces in Real-Life Images Workshop in ECCV*, 2008.

[68] J. Yang, R. Yan, and A. Hauptmann. Cross-Domain Video Concept Detection using Adaptive SVMs. In *Proc. ACM MM*, pages 188–197, 2007.

[69] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face Recognition: A Literature Survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[70] J. Zheng, M.-Y. Liu, R. Chellappa, and J. Phillips. A Grassmann Manifold-Based Domain Adaptation Approach. In *Proc. ICPR*, pages 2095–2099, 2012.