Compositional Dictionaries for Domain Adaptive Face Recognition

Qiang Qiu, and Rama Chellappa, *Fellow, IEEE*.

***Abstract*—We present a dictionary learning approach to compensate for the transformation of faces due to changes in view point, illumination, resolution, etc. The key idea of our approach is to force domain-invariant sparse coding, i.e., design a consistent sparse representation of the same face in different domains. In this way, classifiers trained on the sparse codes in the source domain consisting of frontal faces can be applied to the target domain (consisting of faces in different poses, illumination conditions, etc) without much loss in recognition accuracy. The approach is to first learn a domain base dictionary, and then describe each domain shift (identity, pose, illumination) using a sparse representation over the base dictionary. The dictionary adapted to each domain is expressed as sparse linear combinations of the base dictionary. In the context of face recognition, with the proposed compositional dictionary approach, a face image can be decomposed into sparse representations for a given subject, pose and illumination respectively. This approach has three advantages: first, the extracted sparse representation for a subject is consistent across domains and enables pose and illumination insensitive face recognition. Second, sparse representations for pose and illumination can subsequently be used to estimate the pose and illumination condition of a face image. Finally, by composing sparse representations for subject and the different domains, we can also perform pose alignment and illumination normalization. Extensive experiments using two public face datasets are presented to demonstrate the effectiveness of the proposed approach for face recognition.**

***Index Terms*—Face Recognition, Domain Adaption, Sparse Representation, Pose Alignment, Illumination Normalization, Multilinear Image Analysis.**

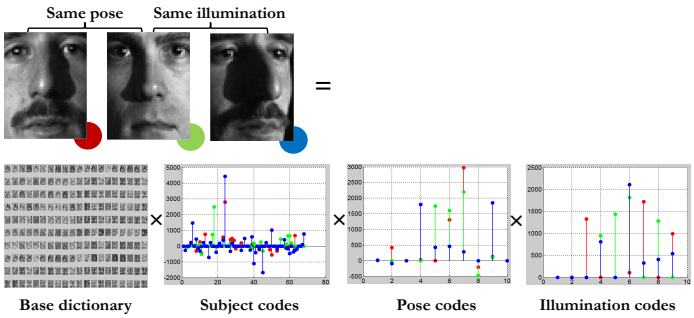


Fig. 1: Trilinear sparse decomposition. Given a domain base dictionary, an unknown face image is decomposed into sparse representations for each subject, pose and illumination respectively. The domain-invariant subject (sparse) codes are used for pose and illumination insensitive face recognition. The pose and illumination codes are also used to estimate the pose and lighting condition of a given face. Composing subject codes with corresponding domain codes enables pose alignment and illumination normalization. Note that the proposed domain-invariant sparse coding assigns similar subject codes to the first and third faces (red and blue), similar pose codes to the first and second faces (red and green), and similar illumination codes to the second and third faces (green and blue).

I. Introduction

Many image recognition algorithms often fail while experiencing a significant visual domain shift, as they expect the test data to share the same underlying distribution as the training data. A visual domain shift is common and natural in the context of face recognition. Such domain shift is due to changes in poses, illumination, resolution, etc.. Domain adaptation [1] is a promising methodology for handling the domain shift by utilizing knowledge in the source domain for problems in a different but related target domain. [2] is one of the earliest works on semi-supervised domain adaptation, where they model data with three underlying distributions: source domain data distribution, target domain data distribution and a distribution of data that is common to both domains. [3] follows a similar model in handling view point changes in the context of activity recognition, where they assume some activities are observed in both source and target domains, while some other activities are only in one of the domains. Under the above assumption, certain hyperplane-based features trained in the source domain are adapted to the target domain for improved classification. Domain adaptation for object recognition is studied in [4], where the subspaces of the source domain, the target domain and the potential intermediate domains are modeled as points on the Grassmann manifold. The shift between domains is learned by exploiting the geometry of the underlying manifolds. A good survey on domain adaptation can be found in [4].

Face recognition across domain, e.g., pose and illumination, has proved to be a challenging problem [5], [6], [7]. In [5], the eigen light-field (ELF) algorithm is presented for face recognition across pose and illumination. This algorithm operates by estimating the eigen light field or the plenoptic function of the subject's head using all the pixels of various images. In [6], [8], face recognition across pose is performed using stereo matching distance (SMD). The cost to match a probe image to a gallery image is used to evaluate the similarity of the two images. For near frontal faces, recent face alignment efforts such as [9], [10], [11] have been shown to be effective. For face recognition across severe pose and/or illumination variations, ELF and SMD methods still report state-of-the-art results. The proposed compositional dictionary learning approach shows comparable performance to these two methods for face recognition across domain shifts due to pose and illumination variations. In addition, our approach can also be used to classify the pose and lighting condition of a face, and perform pose alignment and illumination normalization.

Q. Qiu is at the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA. R. Chellappa is with the Center for Automation Research, UMIACS, University of Maryland, College Park, MD 20742 USA (e-mail: qiang.qiu@duke.edu, rama@umiacs.umd.edu)

The approach presented here shares some of the attributes of the Tensorfaces method proposed in [7], [12], [13], but significantly differs in many aspects. In the Tensorfaces method, face images observed in different domains, i.e., faces imaged in different poses under different illuminations, form a face tensor. Then a multilinear analysis is performed on the face tensor using the N-mode SVD decomposition to obtain a core tensor and multiple mode matrices, each for a different domain aspect. The N-mode SVD decomposition is similar to the proposed multilinear sparse decomposition shown in Fig. 1, where a given unknown image is decomposed into multiple sparse representations for the given subject, pose and illumination respectively. However, we show through experiments that our method based on sparse decomposition significantly outperforms the N-mode SVD decomposition for face recognition across pose and illumination. Another advantage of the proposed method approach over Tensorfaces is that, the proposed approach provides explicit sparse representations for each subject and each visual domain, which can be used for subject classification and domain estimation. Instead, Tensorfaces performs subject classification through exhaustive projections and matchings. Another work similar to Tensorfaces is discussed in [14], where a bilinear analysis is presented for face matching across domains. In [14], a 2-mode SVD decomposition is performed.

This paper makes the following main contributions:

- The proposed domain-invariant sparse coding enables a robust way for multilinear decomposition, and provides explicit sparse representations for out-of-training samples. Note that tensor-based methods obtain representations for out-of-training samples through exhaustive projections and matchings.
- The proposed domain base dictionary learning provides a base dictionary that is independent of subjects and domains, and we express the dictionary adapted to a specific domain as sparse linear combinations of base dictionary atoms using sparse representation of the domain under consideration.
- A face image is decomposed into sparse representations for subject, pose and illumination respectively. The domain-invariant subject (sparse) codes are used for pose and illumination insensitive face recognition. The pose and illumination codes are also used to estimate the pose and lighting condition of a given face. Composing subject codes with corresponding domain codes enables pose alignment and illumination normalization.

The remainder of the paper is organized as follows: Section II discusses some details about sparse decomposition and multilinear image analysis. In Section III, we formulate the compositional dictionary learning problem for face recognition. In Section IV, we present the proposed compositional dictionary learning approach, which consists of algorithms to learn a domain base dictionary, and perform domain-invariant sparse coding. Experimental evaluations are given in Section V on two public face datasets. Finally, Section VI concludes the paper.

II. Background

A. Sparse Decomposition

Sparse signal representations have recently drawn much attention in vision, signal and image processing research [15], [16], [17], [18]. This is mainly due to the fact that signals and images of interest can be sparse in some dictionary. Given an over-complete dictionary $\mathbf{D}$ and a signal $\mathbf{y}$, finding a sparse representation of $\mathbf{y}$ in $\mathbf{D}$ entails solving the following optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{x}, \tag{1}$$

where the ℓ_0 sparsity measure $\|\mathbf{x}\|_0$ counts the number of nonzero elements in the vector $\mathbf{x}$. Problem (1) is NP-hard and cannot be solved in a polynomial time. Hence, approximate solutions are usually sought [19], [20], [21].

The dictionary $\mathbf{D}$ can be either based on a mathematical model of the data or it can be trained directly from the data [22]. It has been observed that learning a dictionary directly from training rather than using a predetermined dictionary (such as wavelet or Gabor) usually leads to better representation and hence can provide improved results in many practical applications such as restoration and classification [15], [16].

Various algorithms have been developed for the task of training a dictionary from examples. One of the most commonly used algorithms is the K-SVD algorithm [23]. Let $\mathbf{Y}$ be a set of N input signals in a n-dimensional feature space $\mathbf{Y} = [\mathbf{y}_1 ... \mathbf{y}_N],\ \mathbf{y}_i \in \mathbb{R}^n$. In K-SVD, a dictionary with a fixed number of K items is learned by finding a solution iteratively to the following problem:

$$\arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad s.t.\ \forall i, \|\mathbf{x}_i\|_0 \leq t \tag{2}$$

where $\mathbf{D} = [\mathbf{d}_1 ... \mathbf{d}_K],\ \mathbf{d}_i \in \mathbb{R}^n$ is the learned dictionary, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N],\ \mathbf{x}_i \in \mathbb{R}^K$ are the sparse codes of input signals $\mathbf{Y}$, and T specifies the sparsity that each signal has fewer than t items in its decomposition. Each dictionary item $\mathbf{d}_i$ is l_2-normalized.

B. Multilinear Image Analysis

Linear methods are popular in facial image analysis, such as principal components analysis (PCA) [24], independent component analysis (ICA) [25], and linear discriminant analysis (LDA) [26]. These conventional linear analysis methods work best when variations in domains, such as pose and illumination, are not present. When any visual domain is allowed to vary, the linear subspace representation above does not capture such variation well.

Under the assumption of Lambertian reflectance, Basri and Jacobs [27] showed that images of an object obtained under a wide variety of lighting conditions can be approximated accurately with a 9-dimensional linear subspace. [28] utilizes the fact that 2D harmonic basis images at different poses are related by close-form linear transformations [29], [30], and extends the 9-dimensional illumination linear space with additional pose information encoded in a linear transformation matrix. The success of these methods suggests the feasibility of decomposing a face image into separate representations for subject and individual domains, e.g. associated pose and illumination, through multilinear algebra.

A multilinear image analysis approach, called Tensorfaces, has been discussed in [7], [12], [13]. Tensor is a multidimensional generalization of a matrix. An N-th order tensor $\mathcal{D}$ is an N-dimensional matrix comprising N spaces. N-mode SVD, illustrated in Fig. 2, is an extension of SVD that decomposes the tensor as the product of N-orthogonal spaces, where Tensor $\mathcal{Z}$, the core tensor, is analogous to the diagonal singular value matrix in SVD. The mode matrix $\mathbf{U_n}$ contains the orthonormal vectors spanning the column space of mode-n flattening of $\mathcal{D}$, i.e., the rearranged tensor elements that form a regular matrix [7].

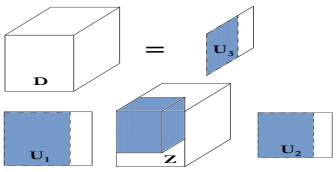


Fig. 2: An N-mode SVD (N=3 is illustrated) [7].

Consider the illustrative example presented in [7]. Given face images of 28 subjects, in 5 poses, 3 illuminations and 3 expressions, and each image contains 7943 pixels, we obtain a face tensor $\mathcal{D}$ of size $28 \times 5 \times 3 \times 3 \times 7943$. Suppose we apply a multilinear analysis to the face tensor $\mathcal{D}$ using the 5-mode decomposition as (3).

$$\mathcal{D} = \mathcal{Z} \times \mathbf{U}_{subject} \times \mathbf{U}_{pose} \times \mathbf{U}_{illum} \times \mathbf{U}_{expre} \times \mathbf{U}_{pixels} \tag{3}$$

where the $28 \times 5 \times 3 \times 3 \times 7943$ core tensor $\mathcal{Z}$ governs the interaction between the factors represented in the 5 mode matrices, and each of the mode matrix $\mathbf{U}_n$ represents subjects and respective domains. For example, the k^{th} row of the 28×28 mode matrix $\mathbf{U}_{subject}$ contains the coefficients for subject k, and the j^{th} row of 5×5 mode matrix $\mathbf{U}_{pose}$ contains the coefficients for pose j.

Tensorfaces perform subject classification through exhaustive projections and matchings. In the above examples, from the training data, each subject is represented with a 28-sized vector of coefficients to the $28 \times 5 \times 3 \times 3 \times 7943$ basis tensor in (4)

$$\mathcal{B} = \mathcal{Z} \times \mathbf{U}_{pose} \times \mathbf{U}_{illum} \times \mathbf{U}_{expre} \times \mathbf{U}_{pixels} \tag{4}$$

One can then obtain the basis tensor for a particular pose j, illumination l, and expression e as a $28 \times 1 \times 1 \times 1 \times 7943$ sized subtensor $\mathbf{B}_{j,l,e}$. The subject coefficients of a given unknown face image are obtained by exhaustively projecting this image into a set of candidate basis tensors for every j, l, e combinations. The resulting vector that yields the smallest distance to one of the rows in $\mathbf{U}_{subject}$ is adopted as the coefficients for the subject in the test image. In a similar way, one can obtain the coefficient vectors for pose and illumination associated with such a test image.

III. Problem Formulation

In this section, we formulate the compositional dictionary learning (CDL) approach for face recognition. It is noted that our approach is general and applicable to both image and non-image data. Let $\mathbf{Y}$ denote a set of N signals (face images) in an n-dim feature space $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$, $\mathbf{y}_i \in \mathbb{R}^n$. Given that face images are from K different subjects $[S_1, \cdots, S_K]$, in J different poses $[P_1, \cdots, P_J]$, and under L different illumination conditions $[I_1, \cdots, I_L]$, $\mathbf{Y}$ can be arranged in six different forms as shown in Fig. 4. We assume here that one image is available for each subject under each pose and illumination, i.e., $N = K \times J \times L$.

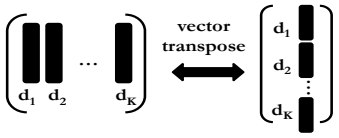


Fig. 3: The vector transpose operator.

$\mathbf{A}$ denotes the sparse coefficient matrix of J different poses, $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_J]$, where $\mathbf{a}_j$ is the sparse representation for the pose P_j. Let $dim(\mathbf{a}_j)$ denote the chosen size of sparse code vector $\mathbf{a}_j$, and $dim(\mathbf{a}_j) \leq J$. $\mathbf{B}$ denotes the sparse code matrix of K different subjects, $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_K]$, where $\mathbf{b}_k$ is the domain-invariant sparse representation for the subject S_k, and $dim(\mathbf{b}_k) \leq K$. $\mathbf{C}$ denotes the sparse coefficient matrix of L different illumination conditions, $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_L]$, where $\mathbf{c}_l$ is the sparse representation for the illumination condition I_l and $dim(\mathbf{c}_l) \leq L$. The domain base dictionary $\mathbf{D}$ contains $dim(\mathbf{a}_j) \times dim(\mathbf{b}_k) \times dim(\mathbf{c}_l)$ atoms arranging in a similar way as Fig. 4. Each dictionary atom is in the $\mathbb{R}^n$ space.

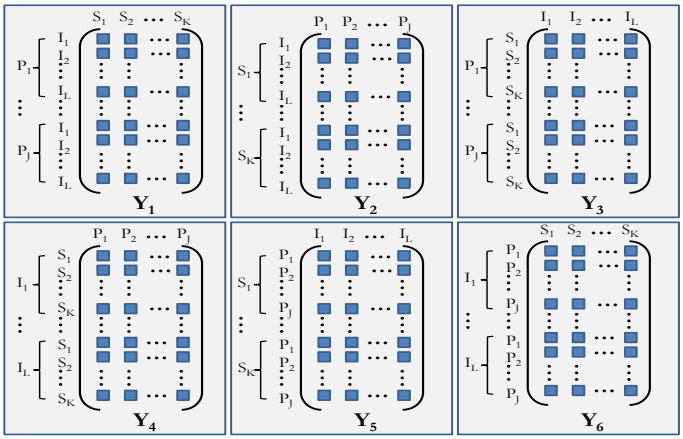


Fig. 4: Six forms of arranging face images of K subjects in J poses under L illumination conditions. Each square denotes a face image in a column vector form.

Any of the six forms in Fig. 4 can be transformed into another through a sequence of vector transpose operations. As illustrated in Fig. 3, a vector transpose operation is to consider (stacked) image vectors in Fig. 4 as values and perform typical matrix transpose operation. For simplicity, we define six aggregated vector transpose operations $\{T_i\}_{i=1}^6$. For example, T_i transforms an input matrix, which is in any of the six forms, into the i-th form defined in Fig. 4 (note that three out of six operations are actually used).

Let $\mathbf{y}_k^{jl}$ be a face image of subject S_k in pose P_j under illumination I_l. The dictionary adapted to pose P_j and illumination I_l is expressed as

$$[[\mathbf{D}^{T_2}\mathbf{a}_j]^{T_3}\mathbf{c}_l]^{T_1}.$$

$\mathbf{y}_k^{jl}$ can be sparsely represented using this dictionary as,

$$\mathbf{y}_k^{jl} = [[\mathbf{D}^{T_2}\mathbf{a}_j]^{T_3}\mathbf{c}_l]^{T_1}\mathbf{b}_k,$$

where the subject sparse codes $\mathbf{b_k}$ are independent of both P_j and I_l. In this way, we can represent Fig. 4 in a compact matrix form as shown in (5).

$$\mathbf{Y}_1 = [[\mathbf{D}^{T_3}\mathbf{C}_1]^{T_2}\mathbf{A}_1]^{T_1}\mathbf{B}_1 \tag{5a}$$
$$\mathbf{Y}_2 = [[\mathbf{D}^{T_3}\mathbf{C}_2]^{T_1}\mathbf{B}_2]^{T_2}\mathbf{A}_2 \tag{5b}$$
$$\mathbf{Y}_3 = [[\mathbf{D}^{T_1}\mathbf{B}_3]^{T_2}\mathbf{A}_3]^{T_3}\mathbf{C}_3 \tag{5c}$$
$$\mathbf{Y}_4 = [[\mathbf{D}^{T_1}\mathbf{B}_4]^{T_3}\mathbf{C}_4]^{T_2}\mathbf{A}_4 \tag{5d}$$
$$\mathbf{Y}_5 = [[\mathbf{D}^{T_2}\mathbf{A}_5]^{T_1}\mathbf{B}_5]^{T_3}\mathbf{C}_5 \tag{5e}$$
$$\mathbf{Y}_6 = [[\mathbf{D}^{T_2}\mathbf{A}_6]^{T_3}\mathbf{C}_6]^{T_1}\mathbf{B}_6 \tag{5f}$$

We now provide the details of solutions to the following two problems

- How to learn a base dictionary that is independent of subject and domains.
- Given an input face image and the base dictionary, how to obtain the sparse representation for the associated pose and illumination, and the domain-invariant sparse representation for the subject.

IV. Compositional Dictionary Learning

In this section, we first show, given a domain base dictionary $\mathbf{D}$, sparse coefficient matrices $\{\mathbf{A}_i\}_{i=1}^6$, $\{\mathbf{B}_i\}_{i=1}^6$ and $\{\mathbf{C}_i\}_{i=1}^6$ are equal across different equations in (5). Then, we present algorithms to learn a domain base dictionary $\mathbf{D}$, and perform domain-invariant sparse coding.

A. Equivalence of Six Forms

To learn a domain base dictionary $\mathbf{D}$, we first need to establish the following proposition.

Proposition 1. *Given a domain base dictionary* $\mathbf{D}$*, matrices* $\{\mathbf{A}_i\}_{i=1}^6$ *in all six equations in (5) are equal, and so are matrices* $\{\mathbf{B}_i\}_{i=1}^6$ *and* $\{\mathbf{C}_i\}_{i=1}^6$.

Proof: First we show matrices $\mathbf{B}_i$ in (5a) and (5f) are equal. $\mathbf{Y}_1$ and $\mathbf{Y}_6$ in Fig. 4 are different only in the row order. We assume a permutation matrix $\mathbf{P}_{16}$ will permutate the rows of $\mathbf{Y}_1$ into $\mathbf{Y}_6$, i.e., $\mathbf{P}_{16}\mathbf{Y}_1 = \mathbf{Y}_6$. Through a dictionary learning process, e.g., k-SVD [23], we obtain a dictionary $\mathbf{D}_1$ and the associated sparse code matrix $\mathbf{B}_1$ for $\mathbf{Y}_1$. $\mathbf{Y}_1$ can be reconstructed as $\mathbf{Y}_1 = \mathbf{D}_1\mathbf{B}_1$. We change the row order of $\mathbf{D}_1$ according to $\mathbf{P}_{16}$ without modifying the actual atom value as $\mathbf{D}_6 = \mathbf{P}_{16}\mathbf{D}_1$. We decompose $\mathbf{Y}_6$ using $\mathbf{D}_6$ as $\mathbf{Y}_6 = \mathbf{D}_6\mathbf{B}_6$, i.e., $\mathbf{P}_{16}\mathbf{Y}_1 = \mathbf{P}_{16}\mathbf{D}_1\mathbf{B}_6$, and we have $\mathbf{B}_1 = \mathbf{B}_6$.

Then we show that matrices $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{C}_i$ in (5a) and (5b) are equal. If we stack all the images from the same subject under the same pose but different illumination as a single observation, we can consider $\mathbf{Y}_2 = \mathbf{Y}_1^T$. By assuming a bilinear model, we can represent $\mathbf{Y}_1$ as $\mathbf{Y}_1 = [\mathbf{D}_c\mathbf{A}_1]^T\mathbf{B}_1$, and we have $\mathbf{Y}_2 = \mathbf{Y}_1^T = [\mathbf{D}_c^T\mathbf{B}_1]^T\mathbf{A}_1$. As $\mathbf{Y}_2 = [\mathbf{D}_c^T\mathbf{B}_2]^T\mathbf{A}_2$, $\mathbf{A}_i$ and $\mathbf{B}_i$ are equal in (5a) and (5b). As both equations share a bilinear map $\mathbf{D}^{T_3}\mathbf{C}_i$, with a common base dictionary $\mathbf{D}$, matrices $\mathbf{C}_i$ are also equal in (5a) and (5b).

Finally, we show matrices $\mathbf{A}_i$ and $\mathbf{C}_i$ in (5a) and (5f) are equal. We have shown in (5a) and (5f) that matrices $\mathbf{B}_i$ are equal. $[[\mathbf{D}^{T_3}\mathbf{C}_1]^{T_2}\mathbf{A}_1]^{T_1}$ and $[[\mathbf{D}^{T_2}\mathbf{A}_6]^{T_3}\mathbf{C}_6]^{T_1}$ are different only in the row order. We can use the bilinear model argument made above to easily show that matrices $\mathbf{A}_i$ and $\mathbf{C}_i$ are equal in (5a) and (5f).

Through the transitivity of equivalence, we can further show matrices $\mathbf{A}_i$ in all six equations in (5) are equivalent, and so are matrices $\mathbf{B}_i$ and $\mathbf{C}_i$. We drop the subscripts in subsequent discussions and denote them as $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. ■

Input: signals $\mathbf{Y}$, sparsity level T_a, T_b, T_c
Output: domain base dictionary $\mathbf{D}$
begin
  Initialization stage:
  1. Initialize $\mathbf{B}$ by solving (5a) via k-SVD
  $\min_{\mathbf{D}_b,\mathbf{B}} \|\mathbf{Y}_1 - \mathbf{D}_b\mathbf{B}\|_F^2$, s.t. $\forall k\ \|\mathbf{b}_k\|_o \le T_b$, where $\mathbf{D}_b = [[\mathbf{D}^{T_3}\mathbf{C}]^{T_2}\mathbf{A}]^{T_1}$
  repeat
    2. apply $\mathbf{B}$ to (5a) and solve via k-SVD ($\mathbf{B}^\dagger = \mathbf{B}^T(\mathbf{B}^T\mathbf{B})^{-1}$)
    $\min_{\mathbf{D}_a,\mathbf{A}} \|(\mathbf{Y}_1\mathbf{B}^\dagger)^{T_2} - \mathbf{D}_a\mathbf{A}\|_F^2$, s.t. $\forall j\ \|\mathbf{a}_j\|_o \le T_a$, where $\mathbf{D}_a = [\mathbf{D}^{T_3}\mathbf{C}]^{T_2}$
    3. apply $\mathbf{A}$ to (5d) and solve via k-SVD
    $\min_{\mathbf{D}_c,\mathbf{C}} \|(\mathbf{Y}_4\mathbf{A}^\dagger)^{T_3} - \mathbf{D}_c\mathbf{C}\|_F^2$, s.t. $\forall l\ \|\mathbf{c}_l\|_o \le T_c$, where $\mathbf{D}_c = [\mathbf{D}^{T_1}\mathbf{B}]^{T_3}$
    4. apply $\mathbf{C}$ to (5e) and solve via k-SVD
    $\min_{\mathbf{D}_b,\mathbf{B}} \|(\mathbf{Y}_5\mathbf{C}^\dagger)^{T_1} - \mathbf{D}_b\mathbf{B}\|_F^2$, s.t. $\forall k\ \|\mathbf{b}_k\|_o \le T_b$, where $\mathbf{D}_b = [\mathbf{D}^{T_2}\mathbf{A}]^{T_1}$
  until *convergence*;
  5. Design the domain base dictionary:
  $\mathbf{D} \leftarrow [\mathbf{D}^{T_2}\mathbf{A}]\mathbf{A}^\dagger$;
  6. return $\mathbf{D}$;
end

Algorithm 1: Domain base dictionary learning.

B. Domain-invariant Sparse Coding

As matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are equal across all six forms in (5), we propose to learn the base dictionary $\mathbf{D}$ using Algorithm 1 given below. The domain dictionary learning in Algorithm 1 optimizes the following objective function,

$$\min_{\mathbf{D},\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathbf{Y}_1 - [[\mathbf{D}^{T_3}\mathbf{C}]^{T_2}\mathbf{A}]^{T_1}\mathbf{B}\|_F^2, \tag{6}$$
$$\text{s.t. } \forall j\ \|\mathbf{a}_j\|_o \le T_a, \forall k\ \|\mathbf{b}_k\|_o \le T_b, \forall l\ \|\mathbf{c}_l\|_o \le T_c,$$

where T_a, T_b, and T_c specify the sparsity level, i.e., the maximal number of non-zero values in a sparse vector. For simplicity and efficiency, we optimize (6) as a sequence of dictionary learning subproblems. More specifically, we first let $\mathbf{D}_b = [[\mathbf{D}^{T_3}\mathbf{C}]^{T_2}\mathbf{A}]^{T_1}$, and perform regular sparse dictionary learning to solve

$$\min_{\mathbf{D}_b,\mathbf{B}} \|\mathbf{Y}_1 - \mathbf{D}_b\mathbf{B}\|_F^2, \text{ s.t. } \forall k\ \|\mathbf{b}_k\|_o \le T_b.$$

We then use the obtained $\mathbf{B}$ to seek an update to $\mathbf{D}_b$ to minimize the same error $\|\mathbf{Y}_1 - \mathbf{D}_b\mathbf{B}\|_F^2$. Taking the derivative with respect to $\mathbf{D}_b$, we obtain $(\mathbf{Y}_1 - \mathbf{D}_b\mathbf{B})\mathbf{B}^T = 0$, leading to the updated $\mathbf{D}_b$ as $\mathbf{Y}_1\mathbf{B}^\dagger = \mathbf{Y}_1\mathbf{B}^T(\mathbf{B}^T\mathbf{B})^{-1}$. As $\mathbf{D}_b = [[\mathbf{D}^{T_3}\mathbf{C}]^{T_2}\mathbf{A}]^{T_1}$, we now use updated $\mathbf{D}_b$ to obtain $\mathbf{D}_a$ and $\mathbf{A}$ as

$$\min_{\mathbf{D}_a,\mathbf{A}} \|(\mathbf{Y}_1\mathbf{B}^\dagger)^{T_2} - \mathbf{D}_a\mathbf{A}\|_F^2, \text{ s.t. } \forall j\ \|\mathbf{a}_j\|_o \le T_a,$$

Input: an input face image $\mathbf{y}$, domain base dictionary $\mathbf{D}$, sparsity level T_a, T_b, T_c
Output: sparse representation vector for pose $\mathbf{a}$, illumination $\mathbf{c}$, subject $\mathbf{b}$
begin
Initialization stage:
1. Initialize domain sparse code vector $\mathbf{a}$ and $\mathbf{c}$ with random values;
Sparse coding stage:
repeat
2. apply $\mathbf{a}$ and $\mathbf{c}$ to (5a) and obtain $\mathbf{b}$ via OMP, $\min_{\mathbf{b}} \|\mathbf{y} - [[\mathbf{D}^{T_3}\mathbf{c}]^{T_2}\mathbf{a}]^T\mathbf{b}\|_2^2$, s.t. $\|\mathbf{b}\|_o \leq T_b$;
3. apply $\mathbf{b}$ and $\mathbf{c}$ to (5d) and obtain $\mathbf{a}$ via OMP, $\min_{\mathbf{a}} \|\mathbf{y} - [[\mathbf{D}^{T_1}\mathbf{b}]^{T_3}\mathbf{c}]^T\mathbf{a}\|_2^2$, s.t. $\|\mathbf{a}\|_o \leq T_a$;
4. apply $\mathbf{a}$ and $\mathbf{b}$ to (5e) and obtain $\mathbf{c}$ via OMP, $\min_{\mathbf{c}} \|\mathbf{y} - [[\mathbf{D}^{T_2}\mathbf{a}]^{T_1}\mathbf{b}]^T\mathbf{c}\|_2^2$, s.t. $\|\mathbf{c}\|_o \leq T_c$;
until *convergence*;
5. return
domain-invariant subject sparse codes: $\mathbf{b}$,
pose sparse codes: $\mathbf{a}$,
illumination sparse codes: $\mathbf{c}$;
end

Algorithm 2: Domain-invariant sparse coding.

where $\mathbf{D}_a = [\mathbf{D}^{T_3}\mathbf{C}]^{T_2}$. Then we fix $\mathbf{A}$ to update $\mathbf{D}_a$, and solve for $\mathbf{D}_c$ and $\mathbf{C}$, and so on.

Algorithm 1 is designed as an iterative method, and each iteration consists of several typical sparse dictionary learning problems. Thus, this algorithm is flexible and can rely on any sparse dictionary learning methods. We adopt the highly efficient dictionary learning method, k-SVD [23]. It is noted that we can easily omit one domain aspect through dictionary "marginalization". For example, after learning the base dictionary $\mathbf{D}$, we can marginalize over illumination sparse codes matrix $\mathbf{C}$ and adopt $[\mathbf{D}^{T_3}\mathbf{C}]^{T_2}$ as the base dictionary for pose domains only.

With the learned base dictionary $\mathbf{D}$, we can perform domain-invariant sparse coding as shown in Algorithm 2, which minimizes the following objective function for a fixed $\mathbf{D}$,

$$\min_{\mathbf{a},\mathbf{b},\mathbf{c}} \|\mathbf{y} - [[\mathbf{D}^{T_3}\mathbf{c}]^{T_2}\mathbf{a}]^{T_1}\mathbf{b}\|_F^2, \quad \text{s.t. } \|\mathbf{a}\|_o \leq T_a,\ \|\mathbf{b}\|_o \leq T_b,\ \|\mathbf{c}\|_o \leq T_c. \tag{7}$$

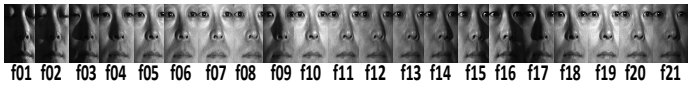
The l_0 norm minimization involved here is NP-hard and usually solved using greedy pursuit algorithms, such as basis pursuit, orthogonal matching pursuit (OMP) [20], [21], to represent a signal with the best linear combination of t atoms from a dictionary, where t is the sparsity. We adopt OMP in the paper. OMP is a greedy algorithm that iteratively selects the dictionary atom best correlated with the residual; and then it produces a new approximant by projecting the signal onto atoms already been selected.

With a base dictionary $\mathbf{D}$ learned using Algorithm 1, a face image can be decomposed using Algorithm 2 into sparse representations $\mathbf{a}$ for the associated pose and $\mathbf{c}$ for illumination, and a domain-invariant sparse representation $\mathbf{b}$ for the subject. While minimizing (6) in Algorithm 1, we obtain the learned domain dictionary $\mathbf{D}$, and model codes $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. Each column of $\mathbf{A}$ denotes the sparse representation assigned to a particular pose in the training. When a training pose shown at testing, the decomposed pose code $\mathbf{a}$ using Algorithm 2 converges to the respective column in $\mathbf{A}$. As shown later, testing poses unseen at training are converged to a sparse linear combination of known poses in a consistent way. We can observe similar convergence for both subject codes and illumination codes.

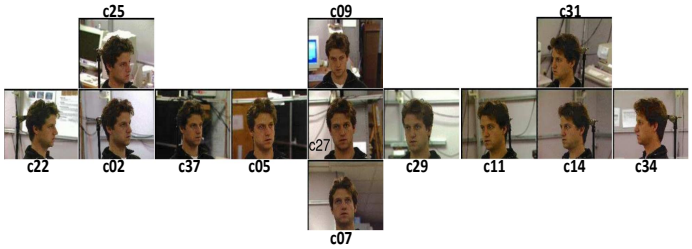
Convergence of Algorithms 1 and 2 can be established using the convergence results of k-SVD discussed in [23]. Although both algorithms optimize a single objective function, the convergence depends on the success of greedy pursuit algorithms involved in each iteration step. We have observed empirical convergence for both Algorithm 1 and 2 in all the experiments reported below.

During training, the domain dictionary learning consists of multiple k-SVD procedures, and the complexity of k-SVD is analyzed in details in [31]. The complexity of Algorithm 2 is more critical, as domain-invariant sparse coding is usually performed at testing. As shown later, it usually takes about 10 iterations for Algorithm 2 to converge, and each iteration consists of three OMP operations. As discussed in [31], [32], different implementations of OMP have different complexities. Considering a signal of dimension m with assumed sparsity t, and a dictionary of N atoms, OMP implemented using the QR Decomposition has the complexity of $Nt + mt + t^2$.

V. Experimental Evaluation



(a) Illumination variation



(b) Pose variation

Fig. 5: Pose and illumination variation in the PIE dataset.

This section presents experimental evaluations on two public face datasets: the CMU PIE dataset [33] and the Extended YaleB dataset [34]. The PIE dataset consists of 68 subjects imaged simultaneously under 13 different poses and 21 lighting conditions, as shown in Fig. 5. The Extended YaleB dataset contains 38 subjects with near frontal pose under 64 lighting conditions. 64×48 sized images are used in the domain composition experiments in Section V-C for clearer visualization. In the remaining experiments, all the face images are resized to 32×24. The proposed Compositional Dictionary learning method is refereed to as CDL in subsequent discussions.

Experimental evaluation is summarized as follows: Section V-A provides the learning configurations for all base dictionaries used. The convergence of domain-invariant sparse

coding is illustrated in Section V-B. Section V-C demonstrates how domain composition is used for pose alignment and illumination normalization. Domina-invariant subject representation is adopted for cross-domain recognition in Section V-D. Section V-E shows that the proposed method is more robust for multilinear decompostion than the tensor-based method. Domain estimation is demonstrated in Section V-F.

A. Learned Domain Base Dictionaries

In our experiments, four different domain base dictionaries $\mathbf{D}_{10}$, $\mathbf{D}_4$, $\mathbf{D}_{34}$, and $\mathbf{D}_{32}$ are learned. We explain here the configurations for each base dictionary.

- $\mathbf{D}_4$: This dictionary is learned from the PIE dataset by using 68 subjects in 4 poses under 21 illumination conditions. The four training poses to the dictionary are $c02$, $c07$, $c09$ and $c14$ poses shown in Fig. 5. The dimensions of coefficient vectors for subject, pose and illumination are 68, 4 and 9. The respective coefficient sparsity values, i.e., the maximal number of non-zero coefficients, are 20, 4 and 9. Note that there is no defined way to specify the sparsity value, and we manually specify sparsity to make each coefficient vector around $\frac{1}{3}$ to $\frac{1}{2}$ full.
- $\mathbf{D}_{10}$: This dictionary is learned from the PIE dataset by using 68 subjects in 10 poses under all 21 illumination conditions. The three unknown poses to the dictionary are $c27$ (frontal), $c05$ (side) and $c22$ (profile) poses. The dimensions of coefficient vectors for subject, pose and illumination are 68, 10 and 9. The respective coefficient sparsity values are 20, 8 and 9.
- $\mathbf{D}_{34}$: This dictionary is learned from the PIE dataset by using the first 34 subjects in 13 poses under 21 illumination conditions. The dimensions of coefficient vectors for subject, pose and illumination are 34, 13 and 9. The respective coefficient sparsity values are 12, 8 and 9.
- $\mathbf{D}_{32}$: This dictionary is learned from the Extended YaleB dataset by using 38 subjects under 32 randomly selected lighting conditions. The dimensions of coefficient vectors for subject and illumination are 38, and 32. The respective coefficient sparsity values are 20 and 20.

The choice of the above 4 dictionary configurations is explained as follows: usually, a common challenging setup for the PIE dataset is to classify subjects in three poses: *frontal* ($c27$), *side* ($c05$) and *profile* ($c22$). Given 13 poses in PIE, we keep the remaining 10 poses to learn $\mathbf{D}_{10}$; We further experiment with fewer samples, e.g., a subset of the remaining 10 poses, to learn $\mathbf{D}_4$; Given 68 subjects in PIE, we learn $\mathbf{D}_{34}$ using half of the subjects; Given 64 illumination conditions in the Extended YaleB data, we learn $\mathbf{D}_{32}$ using half of the lighting conditions.

B. Convergence of Domain-invariant Sparse Coding

We demonstrate here the convergence of the proposed domain-invariant sparse coding in Algorithm 2 over a base dictionary learned using Algorithm 1. We first learn the domain base dictionary $\mathbf{D}_{10}$ using Algorithm 1, and also obtain the associated domain matrices (learned model codes) $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. The matrix $\mathbf{A}$ consists of 10 columns and each column is a unique sparse representation for one of the 10 poses. The matrix $\mathbf{B}$ consists of 68 columns, and each column describes one of the 68 subjects. The matrix $\mathbf{C}$ consists of 21 columns, and each column describes one of the 21 illumination conditions. We observe no significant reconstruction improvements from the learned base dictionary after 2 iterations of Algorithm 1.

Given a face image $(s43, c29, f05)$, i.e., subject $s43$ in pose $c29$ under illumination $f05$, Fig. 6a, 6b and 6c show the decomposed sparse representations for subject $s43$, pose $c29$ and illumination $f05$ after 1, 2 and 100 iterations of Algorithm 2 respectively. We can notice that the decomposed sparse codes (color red) converge to the learned model codes (color blue) in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. As shown in Fig. 8a, we observe convergence after 4 iterations.

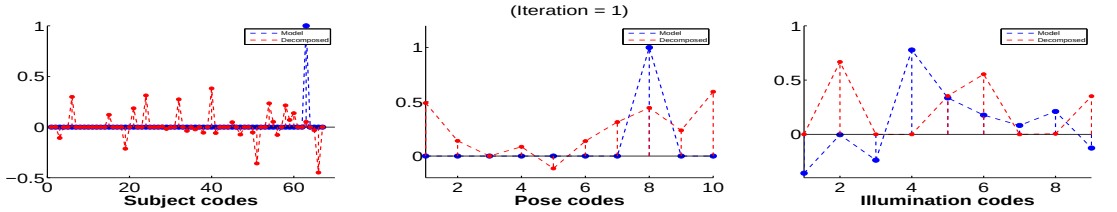
[35] proposed a *Tensor k-SVD* method, which is similar to Tensorfaces but replaces the N-mode SVD with k-SVD to perform multilinear sparse decomposition. Using the Tensor k-SVD method, we are able to learn a Tensor k-SVD dictionary and the associated domain matrices. As the Tensor k-SVD method is designed for data compression, it is not discussed in [35] how to decompose a single image into separate sparse coefficient vectors over such learned Tensor k-SVD dictionary. We adopt a learned Tensor k-SVD dictionary as the base dictionary for domain-invariant sparse coding using Algorithm 2. As shown in Fig.6d, the decomposed sparse codes do not converge well to the learned model codes. It indicates that Algorithm 2 performs an inconsistent decomposition over the Tensor k-SVD dictionary. Therefore, a base dictionary learned from Algorithm 1 is required by the proposed domain-invariant sparse coding in Algorithm 2 to enforce a consistent multilinear sparse decomposition.

We further decompose face images $(s43, c27, f13)$ and $(s01, c27, f13)$ over $\mathbf{D}_{10}$. As shown in Fig. 7, even when pose $c27$ is unknown to $\mathbf{D}_{10}$, the decomposed sparse codes for subjects $s43$ and $s01$, and illumination $f13$ still converge to the learned models. By comparing the pose codes in Fig. 7a and 7b, we notice that the unknown pose $c27$ is represented as a sparse linear combination of known poses in a consistent way. Given the non-optimality of the greedy OMP adopted in each iteration [21], we still observe convergence after about 10 iterations for both cases, as shown in Fig.8b and Fig.8c.

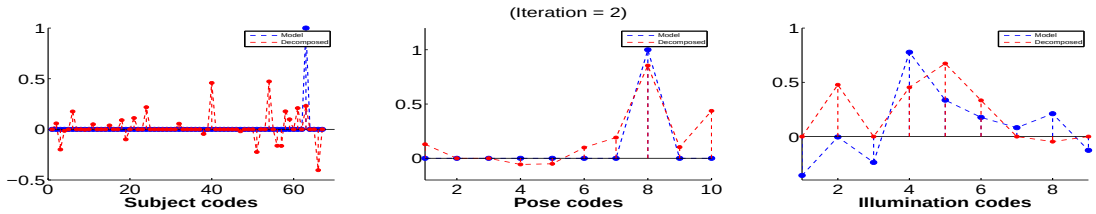
C. Domain Composition

Using the proposed trilinear sparse decomposition over a base dictionary as illustrated in Algorithm 2, we extract from a face image the respective sparse representations for subject, pose and illumination. We can then translate a subject to a different pose and illumination by composing the corresponding subject and domain sparse codes over the base dictionary. As discussed in Sec. II-B, Tensorfaces also enable the decomposition of a face image into separate coefficients for the subject, pose and illumination through exhaustive projections and matchings. We adopt the Tensorfaces method here for a fair comparison in our domain composition experiments.

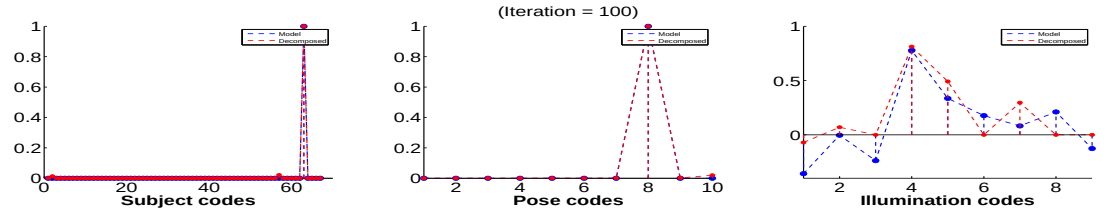
1) Pose Alignment: In Fig. 9a, the base dictionary $\mathbf{D}_{34}$ is used in the CDL experiments. To enable a fair comparison, we adopt the same training data and sparsity values for $\mathbf{D}_{34}$ in



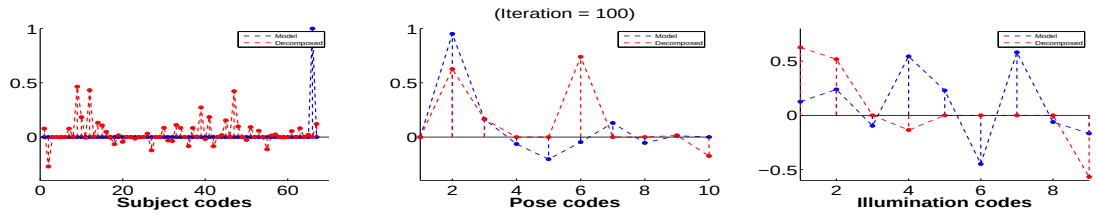
(a) Sparse codes decomposed over $\mathbf{D}_{10}$ after 1 iteration.



(b) Sparse codes decomposed over $\mathbf{D}_{10}$ after 2 iterations.



(c) Sparse codes decomposed over $\mathbf{D}_{10}$ after 100 iterations.



(d) Sparse codes decomposed over a *Tensor k-SVD* dictionary [35] after 100 iterations.

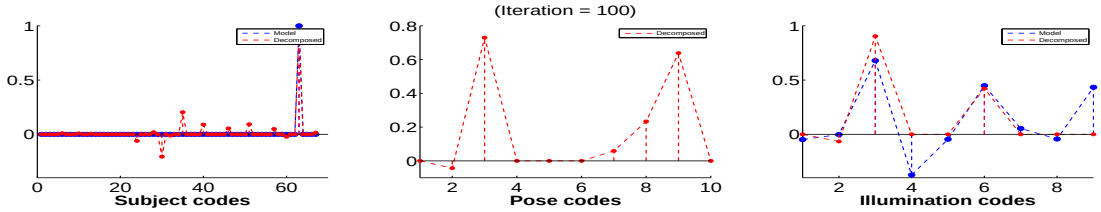
Fig. 6: Domain-invariant sparse coding (Algorithm 2) of the face image $(s43, c29, f05)$, i.e., subject $s43$ in pose $c29$ and illumination $f05$, over a domain base dictionary. $c29$ is a known pose to $\mathbf{D}_{10}$. In (a)-(c), the decomposed sparse codes (color red) converge to the model codes (color blue) when the base dictionary is learned using Algorithm 1; and such convergence is not warranted in (d), when the base dictionary is not from Algorithm 1.

the corresponding Tensorfaces experiments. Given faces from subject s01 under different poses, where both the subject and poses are present in the training data, we extract the subject (sparse) codes for $s01$ from each of them. Then we extract the pose codes for $c27$ (frontal) and the illumination codes for $f05$ from an image of subject $s43$. It is noted that, for such *known subject* cases, the composition $(s01, c27, f05)$ through both CDL and Tensorfaces provides good reconstructions to the ground truth image. The reconstruction using CDL is clearer than the one using Tensorfaces.

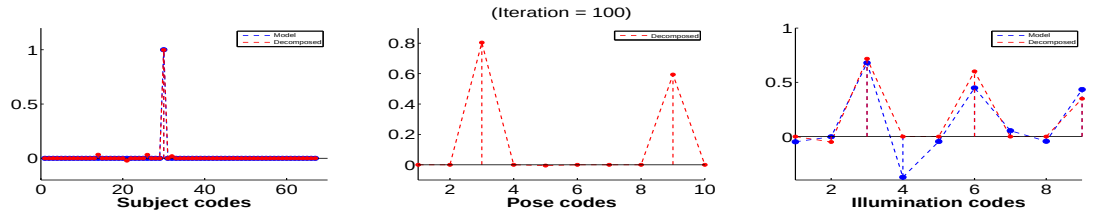
In Fig. 9b, we first extract the subject codes for $s43$, which is an unknown subject to $\mathbf{D}_{34}$. Then we extract the pose codes and the illumination codes from the set of images of $s01$ in Fig. 9a. In this *unknown subject* case, the composition using our CDL method provides significantly more accurate reconstruction to the groundtruth images than the Tensorfaces method. The central assumption in the literature on sparse representation for faces is that the test face image should be represented in terms of training images of the same subject [36], [37]. As $s43$ is unknown to $\mathbf{D}_{34}$, therefore, it is expected that the reconstruction of the subject information is through a linear combination of other known subjects, which is an approximation but not exact.

In Fig. 9c, the base dictionary $\mathbf{D}_{10}$ is used in the CDL experiments, and the same training data and sparsity values for $\mathbf{D}_{10}$ are used in the corresponding Tensorfaces experiments. We first extract the subject codes for $s43$. Then we extract the pose codes for pose $c22$, $c05$ and $c27$, which are unknown poses to the training data. Through domain composition, for such *unknown pose* cases, we obtain more acceptable reconstruction to the actual images using CDL than Tensorfaces. This indicates that, using the proposed CDL method, an unknown pose can be much better approximated in terms of a set of observed poses.

2) Illumination Normalization: In Fig. 10a, we use frontal faces from subject $s28$, which is known to $\mathbf{D}_{34}$, under different illumination conditions. For each image, we first isolate the codes for subject, pose and illumination, and then replace the illumination codes with the one for $f11$. If $f11$ is observed in the training data, the illumination codes for $f11$ can be

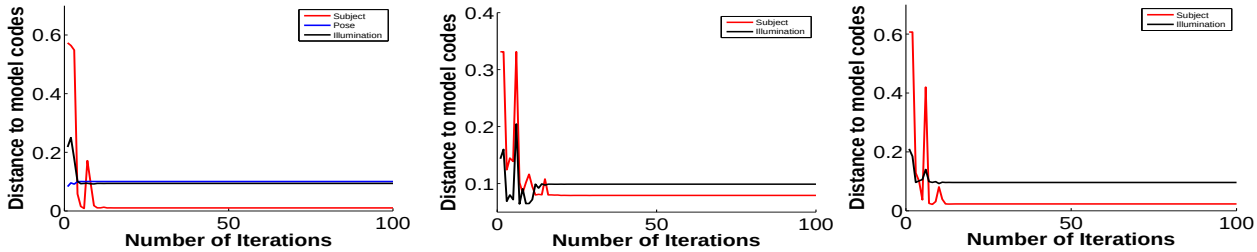


(a) Sparse codes of $(s43, c27, f13)$ decomposed over $\mathbf{D}_{10}$ after 100 iterations.



(b) Sparse codes of $(s01, c27, f13)$ decomposed over $\mathbf{D}_{10}$ after 100 iterations.

Fig. 7: Domain-invariant sparse coding (Algorithm 2) of face images $(s43, c27, f13)$ and $(s01, c27, f13)$ over the domain base dictionary $\mathbf{D}_{10}$. $c27$ is an unknown pose to $\mathbf{D}_{10}$. The decomposed subject and illumination codes (color red) converge to the learned model codes (color blue). Note that the unknown pose $c27$ is represented as a sparse linear combination of known poses in a consistent way.



(a) Sparse decomposition of $(s43, c29, f05)$. (b) Sparse decomposition of $(s43, c27, f13)$. (c) Sparse decomposition of $(s01, c27, f13)$.

Fig. 8: Convergence of domain-invariant sparse coding in Algorithm 2.

obtained during training. Otherwise, the illumination codes for f11 can be extracted from a face image of any subjects under $f11$ illumination. It is shown in Fig. 10a that, for such *known subject* cases, after removing the illumination variation, we can obtain a reconstructed image close to the ground truth image using both CDL and Tensorfaces.

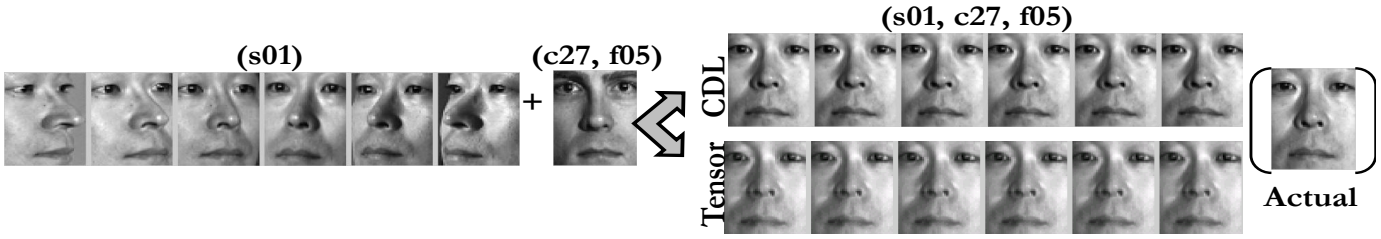
Subject $s43$ in Fig. 10b is unknown to $\mathbf{D}_{34}$. The composed images from CDL exhibit significantly more accurate subject, pose and illumination reconstruction than Tensorfaces. As discussed before, the reconstruction to the subject here is only an approximation but not exact.

D. Pose and Illumination Invariant Face Recognition

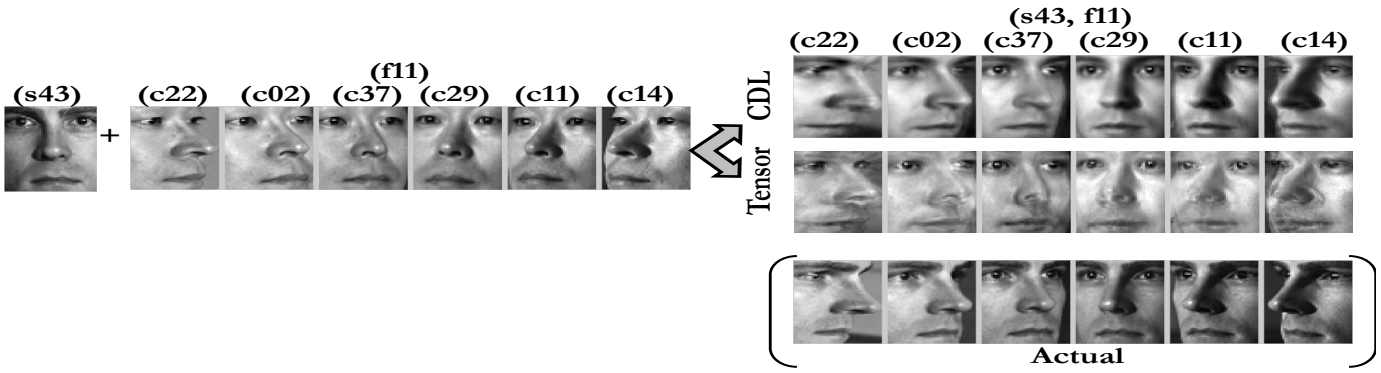
1) Classifying PIE 68 Faces using $\mathbf{D}_4$ *and* $\mathbf{D}_{10}$*:* Fig. 11 shows the face recognition performance under combined pose and illumination variation for the CMU PIE dataset. To enable the comparison with [8], we adopt the same challenging setup as described in [8]. In this experiment, we classify 68 subjects in three poses, frontal ($c27$), side ($c05$), and profile ($c22$), under all 21 lighting conditions. We select one of the 3 poses as the gallery pose, and one of the remaining 2 poses as the probe pose, for a total of 6 gallery-probe pose pairs. For each pose pair, the gallery is under the lighting condition $f11$ as specified in [8], and the probe is under the illumination indicated in the table. Methods compared here include Tensorface[7], [12], SMD [8], and the proposed method CDL. CDL-4 uses the dictionary $\mathbf{D}_4$ and CDL-10 uses $\mathbf{D}_{10}$. In both CDL-4 and CDL-10 setups, three testing poses $c27$, $c05$, and $c22$ are unknown to the training data. It is noted that, to the best of our knowledge, SMD reports the best recognition performance in such experimental setup. As shown in Fig. 11, among 4 out of 6 Gallery-Probe pose pairs, the proposed CDL-10 is better or comparable to SMD.

The stereo matching distance method performs classification based on the stereo matching distance between each pair of gallery-probe images. The stereo matching method can be seen as an example of a zero-shot method as no training is involved. The stereo matching distance becomes more robust when the pose variation between such image pair decreases. However, the proposed CDL classifies faces based on subject codes extracted from each image alone. The robustness of the extracted subject codes only depends on the capability of the base dictionary to reconstruct such a face. This explains why our CDL method significantly outperforms SMD for more challenging pose pairs, e.g., *Profile-Frontal* pair with 62^o pose variation; but performs worse than SMD for easier pairs, e.g., *Frontal-Side* with 16^o pose variation.

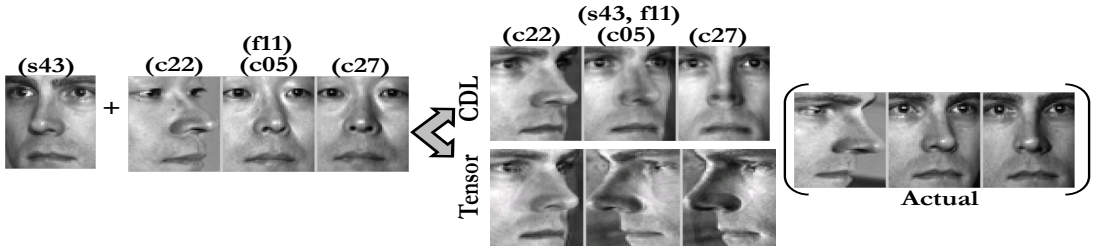
It can be observed in Fig. 9c that an unknown pose can be approximated in terms of a set of observed poses. By representing three testing poses through four training poses in $\mathbf{D}_4$, instead of ten poses in $\mathbf{D}_{10}$, we obtain reasonable



(a) Composition using base dictionary $\mathbf{D}_{34}$. s01 is a known subject to $\mathbf{D}_{34}$. c27 and f05 are extracted from an unknown subject s43.



(b) Composition using base dictionary $\mathbf{D}_{34}$. s43 is an unknown subject to $\mathbf{D}_{34}$.



(c) Composition using base dictionary $\mathbf{D}_{10}$. c22, c05 and c27 are unknown poses to $\mathbf{D}_{10}$.

Fig. 9: Pose alignment through domain composition. In each corresponding Tensorfaces experiment, we adopt the same training data and sparsity values used for the CDL base dictionary for a fair comparison. When a subject or a pose is unknown to the training data, the proposed CDL method provides significantly more accurate reconstruction to the ground truth images.

performance degradations but with 60% less training data.

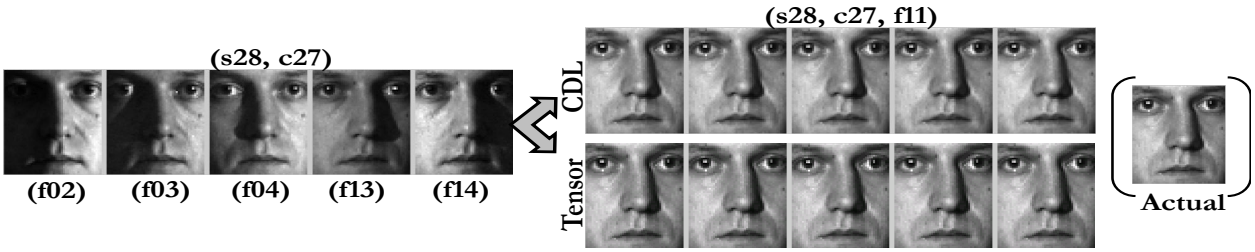
Though the Tensorface method shares a similar multilinear framework to CDL, as seen from Fig. 11, it only handles limited pose and illumination variations.

2) Classifying Extended YaleB using $\mathbf{D}_{32}$*:* We adopt a similar protocol as described in [38]. In the Extended YaleB dataset, each of the 38 subjects is imaged under 64 lighting conditions. We split the dataset into two halves by randomly selecting 32 lighting conditions as training, and the other half for testing. Fig. 12 shows the illumination variation in the testing data. When we learn $\mathbf{D}_{32}$ using Algorithm 1, we also obtain one unique domain-invariant sparse representation for each subject. During testing, we extract the subject codes from each testing face image and classify it based on the best match in unique sparse representation of each subject learned during training. As shown in Table I, the proposed CDL method outperforms other state-of-the-art sparse representation methods (The results for other compared methods are taken from [38]). When the extreme illumination conditions are included, we obtain an average recognition rate 98.91%. By excluding the extreme illumination condition $f35$, we obtain an average recognition rate 99.7%.

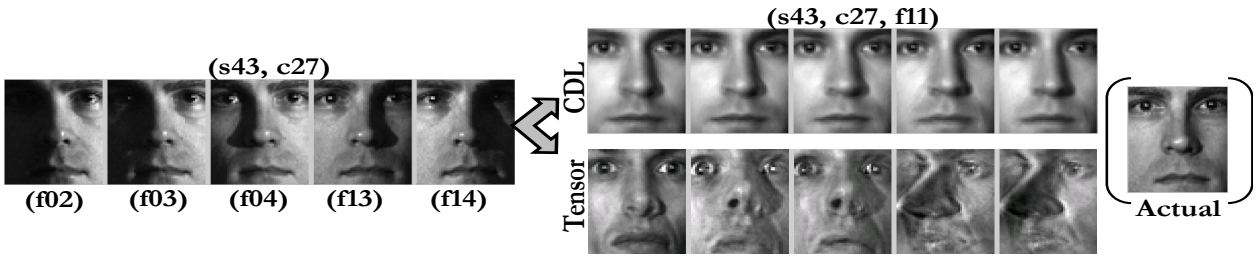
Fig. 12: Illumination variation in the Extended YaleB dataset.

3) Comparisons with More Face Recognition Methods: In this section, we present comparisons with more state-of-the-art face recognition methods to further evaluate the effectiveness of our approach for face recognition across domains. In [18], a different approach for realizing domain-adaptive face recognition is presented. We adopt the same experimental conditions in [18] by learning a domain base dictionary using five training poses $c11$, $c22$ $c27$, $c34$, and $c37$. Fig. 13 shows the classification accuracy on 68 subjects 5712 face images in four testing poses $c02$, $c05$, $c14$, and $c29$ over 21 different lighting conditions. The proposed method (color *red*) significantly outperforms state-of-the-art methods DADL [18] and SRC [36] for face recognition across domains.

We further compare with several techniques designed for illumination robust face representation, including Gradient-



(a) Composition using base dictionary $\mathbf{D}_{34}$. s28 is a known subject to $\mathbf{D}_{34}$.



(b) Composition using base dictionary $\mathbf{D}_{34}$. s43 is an unknown subject to $\mathbf{D}_{34}$.

Fig. 10: Illumination normalization through domain composition. In each corresponding Tensorfaces experiment, we adopt the same training data and sparsity values used for the CDL base dictionary for a fair comparison. When a subject is unknown to the training data, the proposed CDL method provides significantly more accurate reconstruction to the ground truth images.

TABLE I: Face recognition rate (%) on the Extended YaleB face dataset across 32 different lighting conditions. By excluding the extreme illumination condition f35, we obtain an average recognition rate 99.7%

CDL	D-KSVD [39]	LC-KSVD [38]	K-SVD [23]	SRC [36]	LLC [40]
98.91	94.10	95.00	93.1	97.20	90.7

faces [41], LTV [42], SQI [43], and MSR [44]. Following the experiments described in [41], we use 68 subjects with 1428 frontal ($c27$) face images, each with 21 different illuminations for testing. We use one image per subject as the reference images, the other images as the query images. It is noted that some of the compared methods here are unsupervised, and the proposed method requires an additional base dictionary learning step. We adopt here the domain base dictionary $\mathbf{D}_4$ learned from four other training poses. As shown in Table II, the proposed method outperforms compared methods for face recognition under varying illumination.

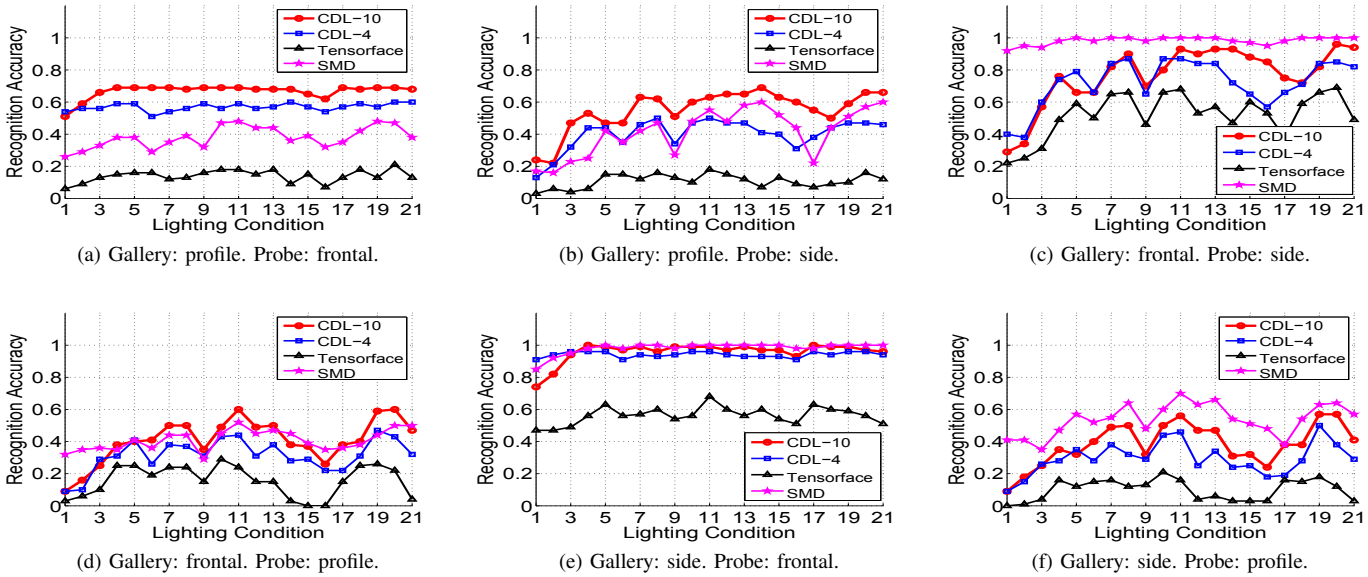
As discussed, in Table I, we obtain one unique domain-invariant sparse representation for each subject during the domain base dictionary learning. During testing, we extract the subject codes from each testing face image and classify it based on the best match in unique sparse representation of each subject learned from the training data. We now adopt a different experimental protocol on the extended YaleB dataset as discussed in [45]. We randomly select 5 images per subject from the training data as the reference and use the remaining images as the query images. The same base dictionary $\mathbf{D}_{32}$ is adopted. We obtain recognition accuracy 99.80%, which is comparable to 97.80% reported in [45].

E. Mean Code and Error Analysis

As discussed in Sec. II-B, the Tensorface method shares a similar multilinear framework to the proposed CDL method. However, we showed through the above experiments that the proposed method based on sparse decomposition significantly outperforms the N-mode SVD decomposition for face recognition across pose and illumination. In this section, we analyze in more detail the behaviors of the proposed CDL and Tensorfaces, by comparing subject and domain codes extracted from a face image using these two methods.

For the experiments in this section, we adopt the base dictionary $\mathbf{D}_{10}$ for CDL, and the same training data and sparsity values of $\mathbf{D}_{10}$ for Tensorfaces to learn the core tensor and the associated mode matrices. The same testing data is used for both methods, i.e., 68 subjects in the PIE dataset under 21 illumination conditions in the $c27$ (frontal), $c05$ (side) and $c22$ (profile) poses, which are three unseen poses not present in the training data.

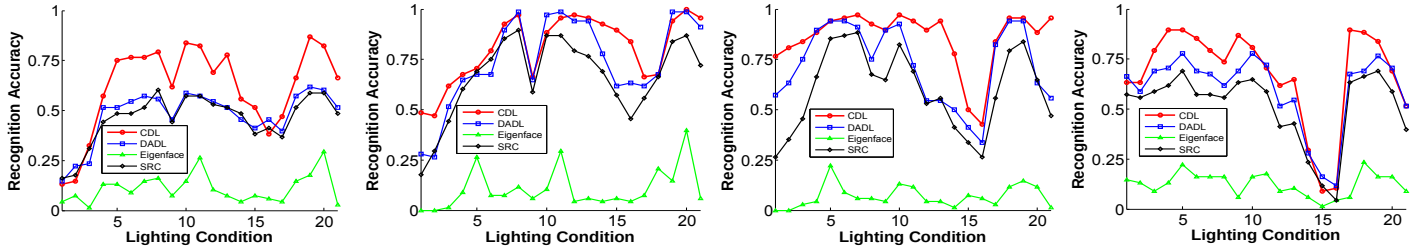
Fig. 14 and Fig. 15 shows the mean subject codes of subject $s1$ and $s2$ over 21 illumination conditions in each of the three testing poses, and the associated standard errors. In each of the two figures, we compare the first row, the subject codes from CDL, with the second row, the subject codes from Tensorfaces. We can easily notice the following: first, the subject codes extracted using CDL are more sparse; second, CDL subject codes are more consistent across pose; third, CDL subject codes are more consistent across illumination, which is indicated by the smaller standard errors. By comparing Fig. 14 with Fig. 15, we also observe that the CDL subject codes are more discriminative. Table III further shows the square root of the pooled variances of subject codes for all 68 subjects over 21 illumination conditions in each of the three testing poses. The significantly smaller variance values obtained using CDL indicate the more consistent sparse representation of subjects decomposed from face images. Therefore, face recognition using CDL subject codes significantly outperforms recognition using Tensorfaces subject codes.



(a) Gallery: profile. Probe: frontal. (b) Gallery: profile. Probe: side. (c) Gallery: frontal. Probe: side.

(d) Gallery: frontal. Probe: profile. (e) Gallery: side. Probe: frontal. (f) Gallery: side. Probe: profile.

Fig. 11: Face recognition under combined pose and illumination variations for the CMU PIE dataset. Given three testing poses, Frontal ($c27$), Side ($c05$), Profile ($c22$), we show the percentage of correct recognition for each disjoint pair of Gallery-Probe poses. See Fig. 5 for poses and lighting conditions. Methods compared here include Tensorface [7], [12], SMD [8] and our compositional dictionary learning (CDL) method . CDL-4 uses the dictionary $\mathbf{D}_4$ and CDL-10 uses $\mathbf{D}_{10}$. To the best of our knowledge, SMD reports the best recognition performance in such experimental setup. 4 out of 6 Gallery-Probe pose pairs, i.e., (a), (b), (d) and (e), our results are comparable to SMD.



(a) Pose c02 (b) Pose c05 (c) Pose c29 (d) Pose c14

Fig. 13: Face recognition accuracy on the CMU PIE dataset using the experimental protocol in [18]. The domain base dictionary is learned from five training poses $c11$, $c22$ $c27$, $c34$, and $c37$. The classification accuracy is reported on 68 subjects 5712 face images in four testing poses $c02$, $c05$, $c14$, and $c29$ over 21 different lighting conditions. The proposed method is denoted as CDL in color red. The proposed method significantly outperforms state-of-the-art methods DADL [18] and SRC [36] for face recognition across domains.

TABLE II: Face recognition rate (%) on the PIE face dataset (pose $c27$) under varying illumination.

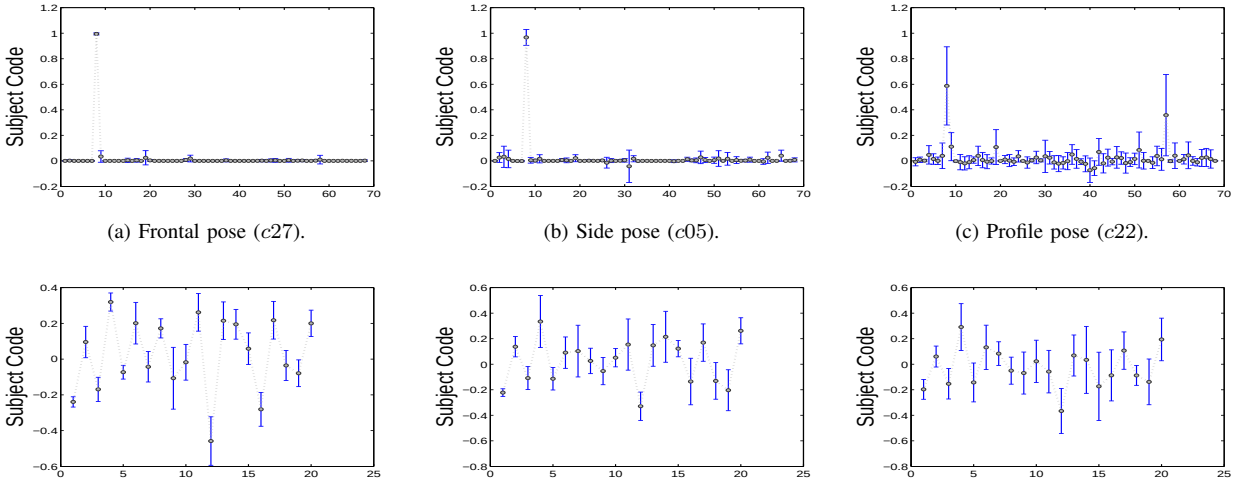
CDL	Gradientfaces [41]	LTV [42]	SQI [43]	MSR [44]
99.93	99.83	86.85	77.94	62.07

F. Pose and Illumination Estimation

In Section V-D, we report the results of experiments over subject codes using base dictionaries $\mathbf{D}_{10}$ and $\mathbf{D}_4$. While generating subject codes, we simultaneously obtain pose codes and illumination codes. Such pose and illumination codes can be used for pose and illumination estimation. In Fig. 16, we show the pose and illumination estimation performance on the PIE dataset using the pose and illumination sparse codes through both CDL and Tensorfaces. The proposed CDL method exhibits significantly better domain estimation accuracy than the Tensorfaces method. By examining Fig. 16, it can be noticed that the most confusing illumination pairs in CDL, e.g., $(f05, f18)$, $(f10, f19)$ and $(f11, f20)$ are very visually similar based on Fig. 5.

VI. Conclusion

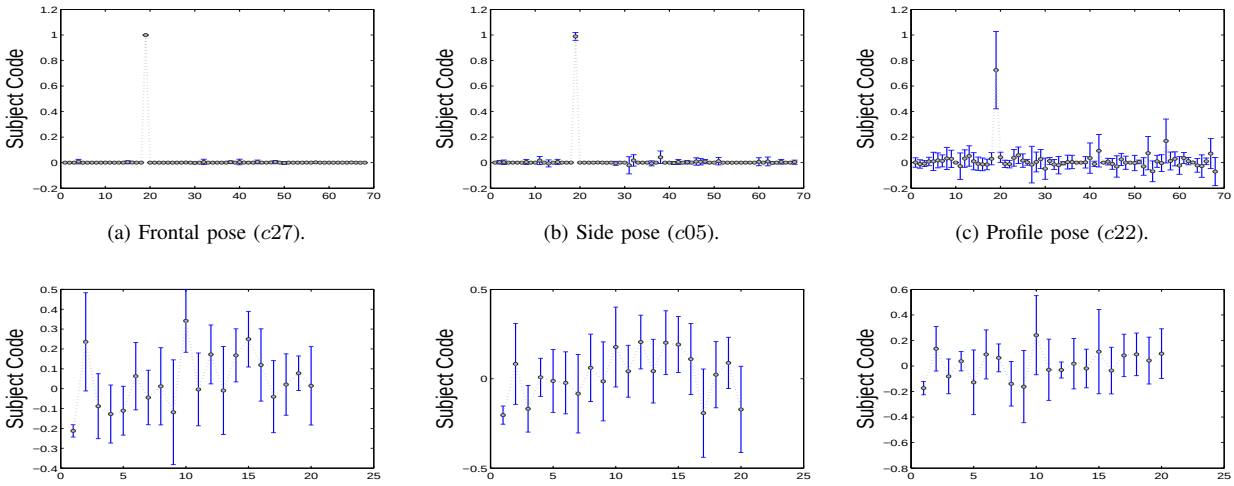
We presented an approach to learn domain adaptive dictionaries for face recognition across pose and illumination domain shifts. With a learned domain base dictionary, an unknown face image is decomposed into subject codes, pose codes and illumination codes. Subject codes are consistent across domains, and enable pose and illumination insensitive face recognition. Pose and illumination codes can be used to estimate the pose and lighting condition of the face. We plan to evaluate the proposed framework in representing 3D



(a) Frontal pose ($c27$). (b) Side pose ($c05$). (c) Profile pose ($c22$).

(d) Frontal pose ($c27$). (e) Side pose ($c05$). (f) Profile pose ($c22$).

Fig. 14: Mean subject code of subject $s1$ over 21 illumination conditions in each of the three testing poses, and standard error of the mean code. (a),(b),(c) are generated using CDL with the base dictionary $\mathbf{D_{10}}$. (d),(e),(f) are generated using Tensorfaces.



(a) Frontal pose ($c27$). (b) Side pose ($c05$). (c) Profile pose ($c22$).

(d) Frontal pose ($c27$). (e) Side pose ($c05$). (f) Profile pose ($c22$).

Fig. 15: Mean subject code of subject $s2$ over 21 illumination conditions in each of the three testing poses, and standard error of the mean code. (a),(b),(c) are generated using CDL with the base dictionary $\mathbf{D_{10}}$. (d),(e),(f) are generated using Tensorfaces.

TABLE III: The square root of the pooled variances of subject codes for 68 subjects over 21 illumination conditions in each of the three testing poses.

	Frontal pose (c27)	Side pose (c05)	Profile pose (c22)
CDL	0.0351	0.0590	0.0879
Tensorfaces	0.1479	0.1758	0.1814

faces. A face image captured by a RGB-D camera provides projected 2D images at various poses. Together with synthesized light sources, we can construct the proposed domain base dictionary; and the learned dictionary can then be used to decompose any given 2D face image for domain-invariant subject representation. We will also experiment the proposed method as a novel way to synthesize more training samples from unseen pose and illumination conditions.

ACKNOWLEDGMENTS

This research was partially supported by a MURI from the Office of Naval research under the Grant N00014-10-1-0934..

REFERENCES

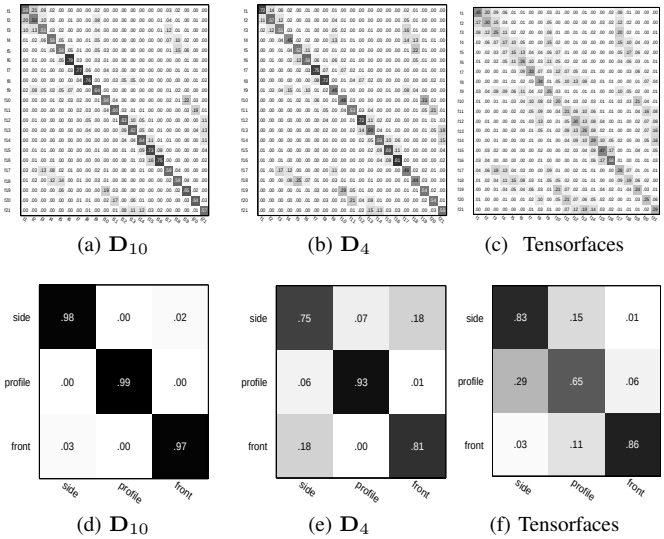
[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, pp. 151–175, 2010.

[2] H. Daumé, III and D. Marcu, "Domain adaptation for statistical classifiers," *J. Artif. Int. Res.*, vol. 26, pp. 101–126, May 2006.

[3] A. Farhadi and M. K. Tabrizi, "Learning to recognize activities from the wrong view point," in *Proc. European Conference on Computer Vision, Marseille, France*, Oct. 2008.

[4] R. Gopalan, R. Li, , and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. Intl. Conf. on Computer Vision, Barcelona, Spain*, Nov. 2011.

[5] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination," in *Handbook of Face Recognition*. Springer-Verlag, 2004.



(a) $\mathbf{D}_{10}$ (b) $\mathbf{D}_4$ (c) Tensorfaces

(d) $\mathbf{D}_{10}$ (e) $\mathbf{D}_4$ (f) Tensorfaces

Fig. 16: Illumination (a-c) and pose (d-f) estimation on the CMU PIE dataset using base dictionaries $\mathbf{D}_4$ and $\mathbf{D}_{10}$. Average accuracy: (a) 0.63, (b) 0.58, (c) 0.28, (d) 0.98, (e) 0.83, (f) 0.78. The proposed CDL method exhibits significantly better domain estimation accuracy than the Tensorfaces method.

[6] C. Castillo and D. Jacobs, "Wide-baseline stereo for face recognition with large pose variation," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Colorado Springs, CO*, June 2011.

[7] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. European Conf. on Computer Vision, Copenhagen, Denmark*, 2002.

[8] C. D. Castillo and D. W. Jacobs, "Using stereo matching for 2-d face recognition across pose," *IEEE Trans. on Patt. Analysis and Mach. Intell.*, vol. 31, pp. 2298–2304, 2009.

[9] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Providence, Rhode Island*, June 2012.

[10] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Columbus, Ohio*, June 2014.

[11] X. Xiong and F. D. la Torre, "Supervised descent method and its application to face alignment," in *IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Portland, Oregon*, June 2013.

[12] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *Proc. Intl. Conf. on Patt. Recn., Quebec, Canada*, Aug, 2002.

[13] S. W. Park and M. Savvides, "Individual kernel tensor-subspaces for robust face recognition: A computationally efficient tensor framework without requiring mode factorization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 5, pp. 1156–1166, 2007.

[14] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247–1283, 2000.

[15] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045 –1057, June 2010.

[16] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, pp. 1031–1044, June 2010.

[17] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *Proc. Intl. Conf. on Computer Vision, Barcelona, Spain*, Nov. 2011.

[18] Q. Qiu, V. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *Proc. European Conference on Computer Vision, Florence, Italy*, Oct. 2012.

[19] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, pp. 33–61, 1998.

[20] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *Proc. 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA*, pp. 40–44, 1993.

[21] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. on Information Theory*, vol. 50, pp. 2231 – 2242, 2004.

[22] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[23] M. Aharon, M. Elad, and A. Bruckstein, "k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, pp. 4311 –4322, Nov. 2006.

[24] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Maui, Hawaii*, June 1991.

[25] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, pp. 705–740, May 1995.

[26] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on Patt. anal. and Mach. Intell.*, vol. 19, pp. 711–720, July 1997.

[27] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. on Patt. anal. and Mach. Intell.*, vol. 25, pp. 218–233, February 2003.

[28] Z. Yue, W. Zhao, and R. Chellappa, "Pose-encoded spherical harmonics for face recognition and synthesis using a single image," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 65, January 2008.

[29] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for reflection," *ACM Transactions on Graphics*, vol. 23, pp. 1004–1042, Oct 2004.

[30] Y. Tanabe, T. Inui, and Y. Onodera, *Group Theory and Its Applications in Physics*. Springer, Berlin, Germany, 1990.

[31] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," 2008, Technical Report - CS Technion.

[32] B. Sturm and M. Christensen, "Comparison of orthogonal matching pursuit implementations," in *2012 Proceedings of the 20th European Signal Processing Conference*, Aug. 2012.

[33] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 25, pp. 1615 –1618, Dec. 2003.

[34] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 23, pp. 643–660, June 2001.

[35] R. Ruiters and R. Klein, "Btf compression via sparse tensor decomposition," *Computer Graphics Forum*, vol. 28, no. 4, pp. 1181–1188, 2009.

[36] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, vol. 31, pp. 210–227, Feb. 2009.

[37] H. Zhang, J. Yang, Y. Zhang, and T. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *Proc. Intl. Conf. on Computer Vision, Barcelona, Spain*, Nov. 2011.

[38] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proc. IEEE Computer Society Cnf. on Computer Vision and Patt. Recn., Colorado springs, CO*, June 2011.

[39] Q. Zhang and B. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn., San Francisco. CA*, June 2010.

[40] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn., San Francisco, CA*, June 2010, pp. 3360–3367.

[41] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. on Image Processing*, vol. 18, no. 11, pp. 2599–2606, 2009.

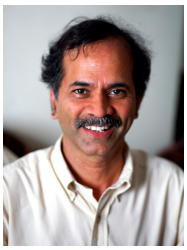
[42] T. Chen, W. Yin, X. Zhou, D. Comaniciu, and T. S. Huang, "Total variation models for variable lighting face recognition," *IEEE Trans. on Patt. anal. and Mach. Intell.*, vol. 28, no. 9, pp. 1519 – 1524, 2006.

[43] H. Wang, S. Z. Li, and Y. Wang, "Generalized quotient image," in *Proc. IEEE Computer Society Conf. on Computer Vision and Patt. Recn., Washington, DC*, June 2004.

[44] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multi-scale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[45] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Trans. on Patt. anal. and Mach. Intell.*, vol. 34, no. 12, pp. 2454–2466, 2012.

Qiang Qiu received his Bachelor's degree with first class honors in Computer Science in 2001, and his Master's degree in Computer Science in 2002, from National University of Singapore. He received his Ph.D. degree in Computer Science in 2012 from University of Maryland, College Park. During 2002-2007, he was a Senior Research Engineer at Institute for Infocomm Research, Singapore. He is currently a Postdoctoral Associate at the Department of Electrical and Computer Engineering, Duke University. His research interests include computer vision and machine learning, specifically on face recognition, human activity recognition, image classification, and sparse representation.



Rama Chellappa received the B.E. (Hons.) degree in Electronics and Communication Engineering from the University of Madras, India and the M.E. (with Distinction) degree from the Indian Institute of Science, Bangalore, India. He received the M.S.E.E. and Ph.D. Degrees in Electrical Engineering from Purdue University, West Lafayette, IN. During 1981-1991, he was a faculty member in the department of EE-Systems at University of Southern California (USC). Since 1991, he has been a Professor of Electrical and Computer Engineering (ECE) and an affiliate Professor of Computer Science at University of Maryland (UMD), College Park. He is also affiliated with the Center for Automation Research and the Institute for Advanced Computer Studies (Permanent Member) and is serving as the Chair of the ECE department. In 2005, he was named a Minta Martin Professor of Engineering. His current research interests span many areas in image processing, computer vision and pattern recognition. Prof. Chellappa has received several awards including an NSF Presidential Young Investigator Award, four IBM Faculty Development Awards, two paper awards and the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). He is a recipient of the Society, Technical Achievement and Meritorious Service Awards from the IEEE Signal Processing Society. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. He is a recipient of Excellence in teaching award from the School of Engineering at USC. At UMD, he received college and university level recognitions for research, teaching, innovation and mentoring undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. Prof. Chellappa served as the Editor-in-Chief of IEEE Transactions on Pattern Analysis and Machine Intelligence and as the General and Technical Program Chair/Co-Chair for several IEEE international and national conferences and workshops. He is a Golden Core Member of the IEEE Computer Society, served as a Distinguished Lecturer of the IEEE Signal Processing Society and as the President of IEEE Biometrics Council. He is a Fellow of IEEE, IAPR, OSA, AAAS, ACM and AAAI and holds four patents.