

Rank Constrained Recognition under Unknown Illuminations

Shaohua Zhou and Rama Chellappa
Center for Automation Research (CfAR)
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
{shaohua, rama}@cfar.umd.edu

Abstract

Recognition under illumination variations is a challenging problem. The key is to successfully separate the illumination source from the observed appearance. Once separated, what remains is invariant to illuminant and appropriate for recognition. Most current efforts employ a Lambertian reflectance model with varying albedo field ignoring both attached and cast shadows, but restrict themselves by using object-specific samples, which undesirably deprives them of recognizing new objects not in the training samples. Using rank constraints on the albedo and the surface normal, we accomplish illumination separation in a more general setting, e.g., with class-specific samples via a factorization approach. In addition, we handle shadows (both attached and cast ones) by treating them as missing values, and resolve the ambiguities in the factorization method by enforcing integrability. As far as recognition is concerned, a bootstrap set which is just a collection of 2D image observations can be utilized to avoid the explicit requirement that 3D information be available. Our approaches produce good recognition results as shown in our experiments using the PIE database.

1 Introduction

Recognition under illumination variations is a challenging problem. The key is to successfully separate the illumination source from the observed appearance. Once separated, what remains is illuminant-invariant and appropriate for recognition. Most current efforts [15, 16, 23] employ a Lambertian reflectance model with varying albedo field ignoring both attached and cast shadows, which derives a subspace completely determined by three images at a fixed pose illuminated by three independent lighting sources. If an ambient component is added [22], this subspace becomes 4-D. If attached shadows are considered as in [1, 2], the subspace dimension grows to infinity (also see [3, 10]) but most of its energy is packed in a limited number of harmonic components, thereby leading to a good approximation.

However, all the above methods (except [16]) commonly restrict themselves by using object-specific samples, thereby unable to generalize to new objects not seen during the training phase. In [16], an ideal class is defined as a collection of 3D objects that possess the same shape but different albedo field. The generalization capability is authorized by assuming that the albedo field is in the rational span of those of a bootstrap set and introducing a quotient image. But, the same shape assumption is still somewhat restrictive.

In this paper, we remove these restrictions by imposing rank constraints on the albedo and surface normal¹. The rank constraint is a result from some parsimony existing in the data and it can arise from the physical or geometrical nature of the problem [18, 12], or from the statistical distribution of the data such as principal component analysis [19], or from combination of both [6, 4].

Our rank constraints enable us to (i) accomplish, in Sec. 2, illumination separation in a more general setting, e.g., with class-specific samples – leading to a singular value decomposition (SVD) approach; and (ii) generalize, in Sec. 4, to recognize new objects by obtaining their illuminant-invariant signatures. As shown in Sec. 5, these methods produce good recognition results on the PIE database [17]. In Sec. 6, we argue that our methods can be regarded in some sense as a generalized 'EigenFace' analysis [19] under illumination variations and discuss related face recognition approaches.

In addition, in Sec. 3 we handle shadows (both attached and cast ones) via the factorization approach with missing values [5], which maintains a minimal rank of 3 for illumination variations (otherwise it goes to a higher rank [1]), and resolve the ambiguities inherent in the factorization approach using the integrability constraint [8, 22].

¹This is not a 'constraint' in practice since we can make the rank as large as possible to span the target space by covering desired variations.

1.1 Notation

a is a scalar, \mathbf{a} is a column vector, $\mathbf{A}_{r \times c}$ is a matrix with r rows and c columns. \mathbf{A}^T denotes the transpose. $[\Rightarrow_i \mathbf{A}_i]$ and $[\Downarrow_i \mathbf{A}_i]$ are horizontal and vertical concatenations, respectively, and the matrix dimension is given by conformation. \mathbf{I} is the identity matrix; $\mathbf{1}$ is the vector or matrix of ones. \otimes denotes the Kronecker (tensor) product; \circ denotes the Hadamard (element-wise) product; \odot denotes the 'tiled' Hadamard (element-wise) product, e.g. $\mathbf{A}_{d \times 1} \odot \mathbf{B}_{d \times 3} = (\mathbf{A}_{d \times 1} \otimes \mathbf{1}_{1 \times 3}) \circ \mathbf{B}_{d \times 3}$; \mathbf{A}^\dagger denotes the pseudo-inverse. \doteq means by definition. Readers are encouraged to refer to [13] for more details on matrix operators.

2 Setting and rank constraints

We assume a Lambertian imaging model with varying albedo field. For image \mathbf{h} , a collection of d pixels, each pixel $h_i, i = 1, \dots, d$, is formulated as follows:

$$h_i = p_i \mathbf{n}_i^T \mathbf{s} = p_i q_i, \quad (1)$$

where p_i is the albedo at pixel i , \mathbf{n}_i (a 3×1 unit vector) is the surface normal at pixel i , \mathbf{s} (a 3×1 unit vector multiplied by its intensity) specifies a distant illuminant, and $q \doteq \mathbf{n}^T \mathbf{s}$. Stacking all the pixels into a column vector \mathbf{h} , we have

$$\begin{aligned} \mathbf{h}_{d \times 1} &= [\Downarrow_i h_i] = [\Downarrow_i p_i \mathbf{n}_i^T] \mathbf{s} = (\mathbf{p}_{d \times 1} \odot \mathbf{N}_{3 \times d}^T) \mathbf{s}_{3 \times 1} \\ &= \mathbf{T}_{d \times 3} \mathbf{s}_{3 \times 1}, \end{aligned} \quad (2)$$

where $\mathbf{p} \doteq [\Downarrow_i p_i]$ is the vector of albedos, $\mathbf{N} \doteq [\Rightarrow_i \mathbf{n}_i]$ is the surface normal matrix, and $\mathbf{T} \doteq \mathbf{p} \odot \mathbf{N}^T$ contains all albedo and shape information. We call the \mathbf{T} matrix as *object-specific albedo-shape* matrix,

In the case of photometric stereo, we have n images of the *same* object, say $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$, observed at a fixed pose illuminated by different n lighting sources. Simple algebraic manipulation gives:

$$\mathbf{H}_{d \times n} = [\Rightarrow_i \mathbf{h}_i] = \mathbf{T} [\Rightarrow_i \mathbf{s}_i] = \mathbf{T}_{d \times 3} \mathbf{S}_{3 \times n}, \quad (3)$$

where $\mathbf{S} \doteq [\Rightarrow_i \mathbf{s}_i]$. Hence photometric stereo is rank-3 constrained. In other words, given at least 3 exemplar images for one object under 3 different independent illuminations, we can determine the identity of a new probe image by checking if it lies in the linear span of the 3 exemplar images. This requires us to store at least 3 images for one object in the gallery set, which is very prohibitive for a gallery set, though for a training set we can usually assume this knowledge. Note that in this recognition setting, there is no requirement of the training set. In other words, the training set is equivalent to the gallery set.

However, our interest lies in recognizing objects not seen in the training stage, i.e., there is no overlap between the

gallery set and the training set in terms of identity. In addition, we pose minimum requirement on the gallery objects by storing only one exemplar image for each object in the gallery set and assume no prior knowledge about the lighting conditions for both the gallery and probe sets. The only assumption is that all images in the training, gallery, and probe sets belong to the same class (e.g. the face class in our example), which naturally invites us the rank constraints.

2.1 The first rank constraint

We impose the rank constraint on the \mathbf{T} matrix by assuming that any \mathbf{T} matrix is a linear combination of some basis matrices $\{\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_m\}$, i.e., there exist coefficients c_i 's such that

$$\begin{aligned} \mathbf{T}_{d \times 3} &= \sum_{i=1}^m c_i \mathbf{T}_i = [\Rightarrow_i \mathbf{T}_i] (\mathbf{c} \otimes \mathbf{I}_3) \\ &= \mathbf{W}_{d \times 3m} (\mathbf{c}_{m \times 1} \otimes \mathbf{I}_3)_{3m \times 3}, \end{aligned} \quad (4)$$

where $\mathbf{c} \doteq [\Downarrow_i c_i]$ and $\mathbf{W} \doteq [\Rightarrow_i \mathbf{T}_i]$. We call the \mathbf{W} matrix as *class-specific albedo-shape* matrix. This assumption generalizes many approaches in the literature and is quite easy to be satisfied. For example, if $m = 1$, this reduces to photometric stereo case; if the surface normal is fixed and the albedo field lies in a rank- m linear subspace, we have Eq. (4) satisfied too. See Sec. 6.1 for justification of this assumptions.

Substitution of Eq. (4) into Eq. (2) yields

$$\begin{aligned} \mathbf{h}_{d \times 1} &= \mathbf{T} \mathbf{s} = \mathbf{W} (\mathbf{c} \otimes \mathbf{I}_3) \mathbf{s} = \mathbf{W} (\mathbf{c} \otimes \mathbf{s}) \\ &= \mathbf{W}_{d \times 3m} \mathbf{k}_{3m \times 1}, \end{aligned} \quad (5)$$

where $\mathbf{k} \doteq \mathbf{c} \otimes \mathbf{s}$.

With the availability of n images $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$ for *different* objects, observed at a fixed pose illuminated by different n lighting sources, we have

$$\begin{aligned} \mathbf{H}_{d \times n} &= [\Rightarrow_i \mathbf{h}_i] = \mathbf{W} [\Rightarrow_i (\mathbf{c}_i \otimes \mathbf{s}_i)] = \mathbf{W} [\Rightarrow_i \mathbf{k}_i] \\ &= \mathbf{W}_{d \times 3m} \mathbf{K}_{3m \times n}, \end{aligned} \quad (6)$$

where $\mathbf{K} \doteq [\Rightarrow_i \mathbf{k}_i]$. It is a rank- $3m$ problem.

Our immediate goal is to estimate \mathbf{W} and \mathbf{K} from the observation matrix \mathbf{H} . The first step is to invoke an SVD factorization, say $\mathbf{H} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T$, and retain the top $3m$ components as $\mathbf{H} = \mathbf{U}_{3m} \mathbf{\Lambda}_{3m} \mathbf{V}_{3m}^T$. Thus, we can recover \mathbf{W} and \mathbf{K} up to an $3m \times 3m$ matrix. i.e., there exists matrices $\mathbf{P}_{3m \times 3m}$ and $\mathbf{Q}_{3m \times 3m}$ such that $\mathbf{W} = \mathbf{U}_{3m} \mathbf{Q}$, $\mathbf{K} = \mathbf{R} \mathbf{V}_{3m}^T$ and $\mathbf{Q} \mathbf{R} = \mathbf{\Lambda}_{3m}$. However, caution should be exercised in SVD because the imaging mechanism violates Eq. (1) very often especially due to shadows. The next step is to handle shadows and resolve ambiguities as shown in Sec. 3.

2.2 The second rank constraint

By noting that $\mathbf{T} = \mathbf{p} \odot \mathbf{N}^T$, we introduce the second rank constraint which assumes that (i) any \mathbf{p} vector is a linear combination of some basis vectors $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m\}$ with $m < d$ and (ii) any \mathbf{N} matrix is a linear combination of some basis matrices $\{\mathbf{N}_1, \mathbf{N}_2, \dots, \mathbf{N}_l\}$ with $l < d$. This is a common constraint used in the face recognition literature. For example, in [14, 21, 7], they all assume that shape and texture have separate bases.

Hence, there exist two vectors $\mathbf{c}_{m \times 1} \doteq [\downarrow_i c_i]$ and $\mathbf{d}_{l \times 1} \doteq [\downarrow_i d_i]$ such that

$$\mathbf{p} = [\Rightarrow_i \mathbf{p}_i] \mathbf{c}; \quad \mathbf{N}^T = [\Rightarrow_i \mathbf{N}_i^T] (\mathbf{d} \otimes \mathbf{I}_3), \quad (7)$$

and similarly the image \mathbf{h} can be expressed as

$$\begin{aligned} \mathbf{h}_{d \times 1} &= [\Rightarrow_{ij} (\mathbf{p}_i \odot \mathbf{N}_j^T)] (\mathbf{c} \otimes \mathbf{d} \otimes \mathbf{s}) \\ &= \mathbf{Y}_{d \times 3ml} (\mathbf{c}_{m \times 1} \otimes \mathbf{d}_{l \times 1} \otimes \mathbf{s}_{3 \times 1}), \end{aligned} \quad (8)$$

where $\mathbf{Y} \doteq [\Rightarrow_{ij} (\mathbf{p}_i \odot \mathbf{N}_j^T)]$, and

$$\begin{aligned} \mathbf{H}_{d \times n} &= [\Rightarrow_i \mathbf{h}_i] \\ &= \mathbf{Y}_{d \times 3ml} [\Rightarrow_i (\mathbf{c}_i \otimes \mathbf{d}_i \otimes \mathbf{s}_i)]_{3ml \times n} \end{aligned} \quad (9)$$

Note that the second constraint can be thought as a special type of the first constraint if we treat \mathbf{Y} and $\mathbf{c} \otimes \mathbf{d}$ in Eq. (8) as \mathbf{W} and \mathbf{c} in Eq. (5), respectively. But, the bilinearity now becomes trilinearity and both \mathbf{c} and \mathbf{d} characterize the identity of \mathbf{h} . This also brings difficulties when one attempts to estimate the bases for \mathbf{p} and \mathbf{N} from the given observations.

3 Handling shadows and resolving ambiguities

3.1 Handling shadows

Practical imaging mechanism must take account of attached and cast shadows as well as sensor noise. We have

$$\tilde{h}_i = \max(h_i + v_i, 0) b(i \in \mathcal{C}^c) + \max(v_i, 0) b(i \in \mathcal{C}), \quad (10)$$

where \mathcal{C} is the cast shadow region with \mathcal{C}^c its complement, v_i is sensor noise at pixel i , and $b(\cdot)$ is a boolean indication function. The existence of shadows in principle increases the rank to infinity. Fortunately, we have a strong hint to distinguish pixels in shadows: their intensities are close to zero. In practice, we set those pixels whose intensities are less than a certain threshold as missing values.

Brand [5] developed an incremental algorithm to apply SVD to data with uncertainty and missing values. Suppose the data is arranged in columns. The idea is to iteratively update or increase the column bases by exhausting the data

column by column. When confronted by missing values, he predicts them using known data and column bases via a normal equation so that the rank growth is minimized. Rank constraint can also be forced by keeping the desired number of bases at each iteration. We have adopted this algorithm to perform SVD in our work. After SVD, we can also predict the missing values in any given image in a similar manner.

3.2 Resolving ambiguities

One common constraint used in shape from shading research is the integrability of the surface [8, 22, 10]. Suppose that the surface function is $z = z(\mathbf{x})$ with $\mathbf{x} \doteq (x, y)$, we must have $\frac{\partial}{\partial x} \frac{\partial z}{\partial y} = \frac{\partial}{\partial y} \frac{\partial z}{\partial x}$. If instead we are given the product of albedo and surface normal as in Eq. (1), say $\mathbf{t}(\mathbf{x}) \doteq (\alpha(\mathbf{x}) \doteq pn_x, \beta(\mathbf{x}) \doteq pn_y, \gamma(\mathbf{x}) \doteq pn_z)$, we have

$$\frac{\partial}{\partial x} \frac{\beta(\mathbf{x})}{\gamma(\mathbf{x})} = \frac{\partial}{\partial y} \frac{\alpha(\mathbf{x})}{\gamma(\mathbf{x})}, \quad (11)$$

i.e.,

$$\gamma(\mathbf{x}) \frac{\partial \beta(\mathbf{x})}{\partial x} - \beta(\mathbf{x}) \frac{\partial \gamma(\mathbf{x})}{\partial x} = \gamma(\mathbf{x}) \frac{\partial \alpha(\mathbf{x})}{\partial y} - \alpha(\mathbf{x}) \frac{\partial \gamma(\mathbf{x})}{\partial y}. \quad (12)$$

In the $m = 1$ case of photometric stereo, A. Yuille et al. [22] enforced the integrability constraint to successfully recover the shape. In principle, we can generalize their analysis for cases with $m > 1$. However, our preliminary experiments show that, especially when m is large, the algorithm is not robust enough, i.e., gets trapped in local minima. We are now working on robust algorithm for extracting the \mathbf{W} matrix from the given training set \mathbf{H} . Fortunately, in the next section, we show that the explicit knowledge of the \mathbf{W} matrix is not required for recognition and it can be completely replaced by practical imagery.

4 Separating illumination

We now consider recognition under illumination variations assuming the availability of the class-specific shape-albedo \mathbf{W} matrix, say learned from the training set. Given a gallery set of different images for different objects captured under different unknown illuminations, we are required to identify the object based on a probe image captured under unknown illumination.

The problem reduces to finding the coefficient \mathbf{c} under the first constraint or \mathbf{c} and \mathbf{d} under the second constraint (but we simply refer as $\mathbf{c} = [\mathbf{c}^T, \mathbf{d}^T]^T$ in the sequel) and illuminant vector \mathbf{s} for an arbitrary image \mathbf{h} . Sec. 4.1 presents such recovery algorithms, which also normalize the solutions to same range.

Once coefficient \mathbf{c} 's have been recovered for both gallery and probe sets, we apply a simple nearest-neighbor classifier which determines the class label of a probe p according

to its correlation-based similarity measure to those in the gallery. The similarity measure d is defined as

$$d = \frac{\mathbf{c}_p^T \mathbf{c}_g}{\sqrt{\mathbf{c}_p^T \mathbf{c}_p} \sqrt{\mathbf{c}_g^T \mathbf{c}_g}}, \quad (13)$$

where \mathbf{c}_p and \mathbf{c}_g are coefficients for the probe and the gallery images, respectively.

4.1 Recovering \mathbf{c} and \mathbf{s} from \mathbf{h}

Given the (5), the recovery task is equivalent to finding \mathbf{c} and \mathbf{s} , which minimizes the least square (LS) cost, i.e.,

$$\min_{\mathbf{c}, \mathbf{s}} \|\mathbf{h} - \mathbf{W}(\mathbf{c} \otimes \mathbf{s})\|^2. \quad (14)$$

Obviously, we can generalize \mathbf{s} to have an arbitrary dimension r , which is not necessarily 3.

Note that \mathbf{c} and \mathbf{s} can be recovered only up to a non-zero scalar, i.e., one can always multiply \mathbf{c} by a non-zero scalar and divide \mathbf{s} by the same scalar. Therefore, without loss of generality, we can simply impose an additional constraint: $\mathbf{1}^T \mathbf{c} = 1$, where $\mathbf{1}_{m \times 1}$ is a vector of 1's.

One way to solve this is indicated in [16]. It is a two-step algorithm. First, \mathbf{k} is approximated by $\mathbf{k} = \mathbf{W}^\dagger \mathbf{h}$. Then $\mathbf{k} = \mathbf{c} \otimes \mathbf{s}$ is used to solve for \mathbf{c} and \mathbf{s} , again using the LS approximation, i.e. finding \mathbf{c} and \mathbf{s} such that the cost $\|\mathbf{k} - \mathbf{c} \otimes \mathbf{s}\|^2$ is minimized. Again, we can impose $\mathbf{1}^T \mathbf{c} = 1$ on the above solutions.

As pointed out in [16], the above algorithm is not robust [16] since two approximations are involved. We now present a simple algorithm which is more robust by observing that Eq. (5) provides a series of sub-equations, which is linear in \mathbf{c} if \mathbf{s} is fixed and in \mathbf{s} if \mathbf{c} is fixed. The algorithm has two iterations. In the first iteration, we solve for the LS solution to \mathbf{s} , given \mathbf{c} .

$$\mathbf{s} = \mathbf{A}^\dagger \mathbf{h}, \quad (15)$$

where $\mathbf{A}_{r \times d} = [a_{i,j}]$ with

$$a_{i,j} = [w_{j,i}, w_{j,i+r}, \dots, w_{j,i+r(m-1)}] \mathbf{c}. \quad (16)$$

In the second iteration, we solve for the LS solution to \mathbf{c} , given \mathbf{s} . Here, we can also include the additional constraint $\mathbf{1}^T \mathbf{c} = 1$ since it is a linear equation too. So,

$$\mathbf{c} = \mathbf{B}^\dagger \begin{bmatrix} \mathbf{h} \\ 1 \end{bmatrix}, \quad (17)$$

where $\mathbf{B}_{(m+1) \times d} = [b_{i,j}]$ with

$$\begin{aligned} b_{i,j} &= [w_{j,(i-1)r+1}, w_{j,(i-1)r+2}, \dots, w_{j,ir}] \mathbf{s} \\ &= 1 \text{ if } i=m+1 \end{aligned} \quad (18)$$

We found that our algorithm is very stable in the sense that it always reaches the same solution regardless of initial conditions and invokes a smaller residual than the algorithm reported in [16]. Appendix shows how to learn \mathbf{c} , \mathbf{d} , and \mathbf{s} from \mathbf{h} using the second constraint. The presented algorithm is actually very general.

4.2 Bootstrap set

The use of the \mathbf{W} matrix is very restrictive in real application since it actually needs 3D information. We now show that, under the first constraint, the \mathbf{W} matrix can be replaced by a bootstrap set containing m exemplar objects captured at a fixed pose, each with three images illuminated by three independent but fixed lighting sources. Denote \mathbf{h}_{ij} as the image for the object i illuminated by the lighting source j .

Using Eq. (5), we can write \mathbf{h}_{ij} and the bootstrap set \mathbf{B} as

$$\mathbf{h}_{ij} = \mathbf{W}(\mathbf{c}_i \otimes \mathbf{s}_j); \quad i = 1, \dots, m; j = 1, 2, 3. \quad (19)$$

$$\begin{aligned} \mathbf{B}_{d \times 3m} &= [\Rightarrow_{ij} \mathbf{h}_{ij}] = \mathbf{W}[\Rightarrow_{ij} (\mathbf{c}_i \otimes \mathbf{s}_j)] \\ &= \mathbf{W}_{d \times 3m} (\mathbf{C}_{m \times m} \otimes \mathbf{S}_{3 \times 3}), \end{aligned} \quad (20)$$

where $\mathbf{C} \doteq [\Rightarrow_i \mathbf{c}_i]$ and $\mathbf{S} \doteq [\Rightarrow_j \mathbf{s}_j]$ define the (though not orthogonal) bases for the identity coefficients and the light sources, respectively. Therefore, finding \mathbf{c} and \mathbf{s} for image \mathbf{h} is equivalent to finding \mathbf{b} and \mathbf{t} , which relates \mathbf{c} with \mathbf{C} and \mathbf{s} with \mathbf{S} , defined below:

$$\mathbf{c} = \sum_{i=1}^m b_i \mathbf{c}_i = \mathbf{C} \mathbf{b}; \quad \mathbf{s} = \sum_{j=1}^3 t_j \mathbf{s}_j = \mathbf{S} \mathbf{t}. \quad (21)$$

$$\begin{aligned} \mathbf{h}_{d \times 1} &= \mathbf{W}(\mathbf{c} \otimes \mathbf{s}) = \mathbf{W}(\mathbf{C} \mathbf{b} \otimes \mathbf{S} \mathbf{t}) \\ &= \mathbf{W}(\mathbf{C} \otimes \mathbf{S})(\mathbf{b} \otimes \mathbf{t}) \\ &= \mathbf{B}_{d \times 3m} (\mathbf{b}_{m \times 1} \otimes \mathbf{t}_{3 \times 1}) \end{aligned} \quad (22)$$

where \mathbf{B} is the bootstrap set.

The use of the bootstrap set has an additional benefit. As indicated in Sec. 1, the rank for covering illumination variations in practice exceeds 3. Suppose that this rank is $r > 3$, we can use a bootstrap set of dimension d by rm , i.e. using images for m objects taken under r lighting conditions, to improve the recognition performance. A final note is that no bootstrap set can be developed for the second constraint using exemplar images.

5 Experiments

5.1 PIE database

We use the Pose and Illumination and Expression (PIE) database [17] to demonstrate the effectiveness of our ap-

proach². Fig. 1 shows the distribution of all 21 flashes. For illustrative purposes, we move their positions on a unit sphere as only the illuminant directions matter. Since the flashes are almost symmetrically distributed about the head position, we only use 12 of them. In total, we used $68 \times 12 = 816$ images in a fixed view as there are 68 subjects in the PIE database. Also, we only study gray images by taking the average of the red, green, and blue channels of their color versions.

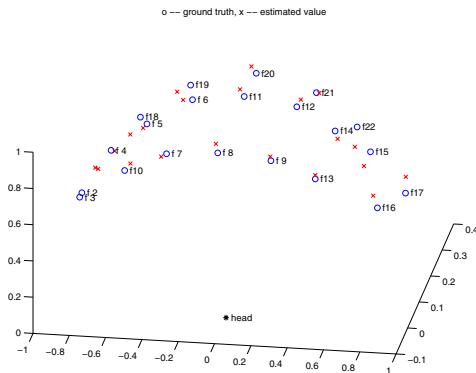


Figure 1: Flash distribution in PIE database. 'o' means the ground truth and 'x' the estimated values.

Registration is performed by aligning the eyes and mouth to desired positions. No flow computation [14] is carried on for further alignment. We only focus on inner face region by putting an ellipse-shaped mask to remove 'bad' pixels such as hair, etc. After the pre-processing step, the number of pixels we utilize is 4805, i.e. $d = 4805$.

A typical recognition scenario using the PIE database is as follows. The bootstrap set (or the training set) is from the Yale's illumination database [10] or Vetter's 3D face database [14]. The gallery set is taken as all images under one unknown illumination condition and the probe set as all images under another unknown illumination condition. Assuming the identity information of the gallery set, we infer the identity for the probe set.

5.2 Recognition performance

We first assume that all the images have been captured in frontal view, but we do not know the directions and intensities of the illuminants.

The bootstrap set is first taken as the Yale's illumination database [10]. There are only 10 subjects (i.e. $m = 10$) in this database and each subject has 64 images in frontal view illuminated by 64 different lights. We pick out images under 9 lights (i.e., $r = 9$) in order to cover up to second-order

²We use the 'illum' part of the PIE database not the 'light' part as this is closer to the Lambertian model.

harmonic components [1]. Fig. 2 shows some example images in the bootstrap set.

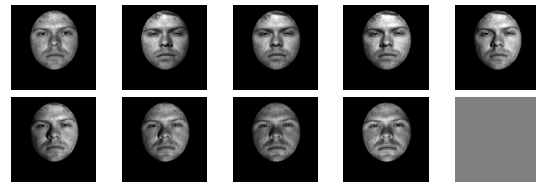


Figure 2: Yale database: one subject in 9 illuminants.

Table 1 lists the recognition rate for the PIE database using Yale's database as the bootstrap set. Even with $m = 10$, we obtain quite good results. One observation is that when the flashes become separated, the recognition rate is lowered. Also, using images under frontal or near-frontal illuminants as galleries produces good results.

For comparison, we also implemented the 'EigenFace' approach [19] by training the eigenvector from the same bootstrap set. The recognition rates are also presented in Table 1. Obviously, the performance is much worse than that of our approach. This highlights the virtue of decoupling the illumination variations.

It is rather restrictive to have only 10 subjects in the bootstrap set. We now increase m from 10 to 100 by using Vetter's 3D face database [14]. In the original database, only the surface depth and texture map is provided, but we need surface normal and albedo information. The surface normal is easy to calculate from the given surface depth data. To estimate the albedo, we first use our algorithm to estimate illuminant directions using the Yale's database (assuming the directions are known in the Yale's database), then we use the Lambertian model to compute the albedo field. Therefore, in this case, we actually have \mathbf{p} , \mathbf{N} , \mathbf{T} , and \mathbf{W} available. However, we believe that using a bootstrap set of $m = 100$ from other sources can yield similar performances.

Table 2 tabulates the obtained recognition rates by imposing the first rank constraint. Significant improvements have been achieved due to the increase in m . This seems to suggest that a moderate sample size of 100 is enough to span the entire face space.

We then experiment with the second rank constraint. Note that here we need explicit knowledge of \mathbf{p} and \mathbf{N} , while we only need a bootstrap set if the first constraint is used. Table 2 also tabulates the obtained recognition rate. It seems that the use of the second constraint does not help too much. In fact, it is slightly worse due to possible over-parameterization. Given the difficulty of acquiring the knowledge of \mathbf{p} and \mathbf{N} and the slower computation, it seems beneficial to use the first rank constraint.

We now present our preliminary results on recognition across pose. Our approach in principle can also handle pose variation since the \mathbf{W} matrix contains all the needed 3D in-

formation, i.e., we can recover the 3D model from it. But as mentioned earlier, learning the W matrix is not robust. Here, we simply use Vetter's database to handle pose variations. Pose is roughly estimated from the geometric calibration information provided in the PIE database. We then warp the 3D model to the desired pose. The rest just follows using the first constraint approach. Table 3 lists the recognition results obtained. In general, using the side view does not hurt the recognition rate too much. Our rates are quite good, but compared to [14], there is some room to improve. We believe that this can be accomplished using improved pre-processing technique such as flow based correspondence and refinement of pose estimation.

5.3 Illuminant estimation and face synthesis

In the above process, we achieve two byproducts: illuminant estimation, and face synthesis. Fig. 1 also shows the estimated illuminant directions. It is quite accurate for estimation of directions of flashes near frontal pose. But when the flashes are very off-frontal, accuracy slightly goes down.

Face synthesis has also been done in Freeman's bilinear analysis [9]. Fig. 3 displays some synthesized faces. The synthesis results bases on the Yale database are not far from perfect, but satisfactory considering the small database size. Using Vetter's bootstrap set greatly improves the synthesis performance since it spans the face space more accurately. However, this is only the face space in the frontal view. For a non-frontal view where pose variation is dominant, the synthesis performance degrades, e.g., one cannot observe the left face contour in the synthesized image in the fifth column.

6 Discussions and summary

6.1 Are our rank constraints valid?

We justify the validity by claiming that the famous 'EigenFace' approach [19] is just a special case of our approach, but this is only for a fixed illumination source. Suppose that the illuminant vector is \tilde{S} , Eq. (5) becomes

$$h_{d \times 1} = W(c \otimes I_3)\tilde{S} = \tilde{W}_{d \times m} c_{m \times 1}, \quad (23)$$

where $\tilde{W} \doteq [\Rightarrow_i T_i \tilde{S}]$, and consequently Eq. (6) becomes

$$H_{d \times n} = [\Rightarrow_i h_i] = \tilde{W}[\Rightarrow_i c_i] = \tilde{W}_{d \times m} C_{m \times n}. \quad (24)$$

This elegantly reduces to the regular 'EigenFace' analysis. Therefore, our approach can be regarded as a generalized 'EigenFace' analysis under illumination variation.

The requirement of fixed lighting source also explains why the 'EigenFace' approach does not work well for recognizing faces under illumination variations. Our experiments also confirm this.



Figure 3: Row 1: original images. Row 2: synthesized images using Yale's bootstrap set. Row 3: synthesized images using Vetter's bootstrap set.

6.2 Related works

Our approach is closely related to those dealing with illuminations, such as photometric stereo [15], illumination cone [3] and others [22, 1, 2]. This has been briefly surveyed in Sections 1 and 2. We now only list some related works in face recognition and synthesis literature.

The bilinear approach [9] studied the illumination variations. Our approach essentially coincides with their study. It seems that they assumed the bilinearity without justification. But we principally justified why a bilinear analysis is appropriate to attack illumination variations by showing this is a rank- $3m$ problem, and focused on recognition.

The 'TensorFace' approach [20] carries on a multilinear analysis to take account different factors, such as illumination, expression, pose, identity. However, the multilinear assumption is rather weak, especially on view and expression because of corresponding problems, i.e., the corresponding pixels under different poses and expressions are completely misaligned. Also, in [20], recognition is still performed based on object-specific samples; so generalization to class-specific samples is not available. Finally, the illumination variation is considered explicitly, but not decoupled in the recognition process, which might comprise the recognition performance.

The 3D morphable model approach [14] developed by Vetter *et al.* deals with both pose and illumination variations in an elegant manner. Our approaches look similar to this. But there are significant differences. The first rank constraint has no equivalence in [14]. Especially, using the first constraint frees us from the need to know 3D models; instead we use the bootstrap set, which is a collection of 2D images. Second, even the second rank constraint is different from [14], since we use albedo and surface normal information, while Vetter *et al.* use depth and texture map. Finally, in the experiments, we assume known pose but unknown illumination, but Vetter *et al.* assume unknown pose but known illumination. Also, compared to [14], our face alignment is rather crude and we believe that our recognition will

be improved by using more fine alignment via flow computation and pose refinement.

6.3 Summary of our approach

The above comparisons highlight the following aspects of our approach. It naturally combines the rank-constraint for identity with illumination modeling. This combination enables us to separate the illumination variations from the observation in a principled manner and generalizes us to recognize new objects not seen in the training phase. Our experimental study shows that recognition across illumination is very good, and recognition across pose is also promising, but we still need to investigate (i) extreme lighting condition: This cause more shadows. Ray tracing is one solution [10]; (ii) profile view: The issue here is how to relax the need for the 3D model. 'Eigen Light-Field' is a promising method [11]; and (iii) robust recovery of the \mathbf{W} matrix.

References

- [1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *Proc. of ICCV*, 2001.
- [2] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. *Proc. of CVPR*, 2001.
- [3] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *IJCV*, 28(3):245–260, 1998.
- [4] M. Brand. Morphable 3d models from video. *Proc. CVPR*, 2001.
- [5] M. Brand. Incremental singular value decomposition of uncertain data with missing values. *Proc. ECCV*, 2002.
- [6] C. Bregler, A. Hertzmann, and H. Biermann. Recovering nonrigid 3D shape from image streams. *Proc. CVPR*, 2000.
- [7] T. Cootes, G. Edwards, and C. Taylor. Active appearance model. *Proc. ECCV*, 1949.
- [8] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading problem. *IEEE Trans. PAMI*, PAMI-10(7):439–451, 1987.
- [9] W. T. Freeman and J. B. Tenenbaum. Learning bilinear models for two-factor problems in vision. *Prof. CVPR*, 1937.
- [10] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. PAMI*, 23(0):643–660, 2001.
- [11] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. *Prof. Face and Gesture Recognition*, 2002.
- [12] M. Irani. Multi-frame optical flow estimation using subspace constraints. *Prcc. of ICCV*, pages 626–633, 1999.
- [13] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley, 1999.
- [14] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. *Proc. ECCV*, 2002.
- [15] A. Shashua. On photometric issues in 3d visual recognition from a single 2d image. *IJCV*, 21(1):99–122, 1997.
- [16] A. Shashua and T. R. Raviv. The quotient image: Class based re-rendering and recognition with varying illuminations. *IEEE Trans. PAMI*, 23(2):129–139, 2001.
- [17] T. Sim, S. Baker, and M. Bast. The CMU pose, illumination, and expression (PIE) database. *Prof. Face and Gesture Recognition*, 2002.
- [18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Computer Vision*, 9(2):137–154, 1992.
- [19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neutoscience*, 3:72–86, 1991.
- [20] M. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. *Proc. ICPR*, 2002.
- [21] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. PAMI*, 11(7):733–742, 1997.
- [22] A. L. Yuille, D. Snow, E. R., and P. N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *IJCV*, 35(3):203–222, 1999.
- [23] W. Zhao and R. Chellappa. Symmetric shape from shading using self-ratio image. *IJCV*, 45(10):55–752, 2001.

Appendix: recovering $\{\mathbf{c}^1, \dots, \mathbf{c}^n\}$ from \mathbf{h}

The algorithm presented in Sec. 4.1 can be generalized to recover $\{\mathbf{c}^1, \dots, \mathbf{c}^n\}$ from \mathbf{h} if the following multilinear form is satisfied:

$$\mathbf{h}_{d \times 1} = \mathbf{W}_{d \times \prod_{i=1}^n m_i} (\mathbf{c}_{m_1 \times 1}^1 \otimes \dots \otimes \mathbf{c}_{m_n \times 1}^n), \quad (25)$$

where $\mathbf{W} \doteq [\Rightarrow_{j_1, \dots, j_n} \mathbf{w}_{j_1, \dots, j_n}]$. Again, we impose the addition constraints: $\mathbf{1}^T \mathbf{c}^i = 1$; $i = 1, \dots, n-1$.

In the iteration for computing \mathbf{c}^i given all other \mathbf{c}^j 's ($j \neq i$) fixed, we have,

$$\mathbf{h} = \mathbf{A}^i \mathbf{c}^i, \quad (26)$$

where $\mathbf{A}^i \doteq [\Rightarrow_{j_i=1}^{m_i} \mathbf{a}_{j_i}^i]$ and

$$\mathbf{a}_{j_i}^i = \sum_{j_1=1}^{m_1} \dots \sum_{j_n=1}^{m_n} c_{j_1}^1 \dots c_{j_{i-1}}^{i-1} c_{j_{i+1}}^{i+1} \dots c_{j_n}^n \mathbf{w}_{j_1, \dots, j_n}. \quad (27)$$

If $\mathbf{1}^T \mathbf{c}_i = 1$ is imposed for $i = 1, \dots, n-1$, the LS solution to \mathbf{c}^i is

$$\begin{aligned} \mathbf{c}^i &= \left[\begin{array}{c} \mathbf{A}^i \\ \mathbf{1}^T \end{array} \right]^\dagger \left[\begin{array}{c} \mathbf{h} \\ 1 \end{array} \right]; \quad i = 1, \dots, n-1 \\ &= [\mathbf{A}^n]^\dagger \mathbf{h}; \quad i = n. \end{aligned} \quad (28)$$

P\G	f08	f09	f11	f12	f13	f14	f15	f16	f17	f20	f21	f22	Avg.
f08	-	24/96	91/96	22/87	6/66	3/60	3/46	3/29	1/22	68/85	10/78	3/53	21/65
f09	38/94	-	40/96	94/96	40/90	34/87	4/56	7/40	4/24	31/84	82/96	10/68	35/75
f12	97/94	26/91	-	34/97	7/72	7/72	4/38	3/28	1/16	100/100	18/94	4/51	28/69
f13	28/88	99/94	40/97	-	43/88	53/93	12/57	9/41	4/28	38/94	100/100	19/76	40/78
f14	12/56	49/87	9/59	44/85	-	100/100	60/90	22/71	18/50	7/54	54/87	96/100	43/76
f15	10/51	49/85	10/63	56/93	100/100	-	72/90	19/66	16/49	9/59	74/91	100/99	47/77
f16	3/33	13/40	6/37	12/49	57/85	66/88	-	94/93	63/78	6/32	22/49	100/97	40/62
f16	3/19	9/26	4/26	9/32	24/59	24/44	79/84	-	99/93	3/26	13/31	40/63	28/46
f17	3/14	6/28	14/19	6/26	10/50	12/41	37/68	100/94	-	1/19	6/26	18/44	18/39
f20	84/90	31/85	100/99	31/97	7/65	9/69	4/38	3/26	1/21	-	25/93	5/53	27/67
f21	24/79	88/94	31/93	100/100	66/88	82/94	13/62	9/49	7/28	22/91	-	34/76	43/78
f22	9/43	24/65	6/46	32/75	96/99	99/99	100/97	20/76	32/59	8/43	47/74	-	46/70
Avg.	28/60	38/72	31/66	40/76	44/78	35/77	29/66	23/56	27/42	41/63	40/74	40/71	35/67

Table 1: Recognition rate obtained by 'EigenFace' approach and our approach using the 1st constraint and the Yale's database as the bootstrap set. In each cell, the left number is for 'EigenFace' and the right one for our approach. 'P' means probe, 'G' means gallery, 'fnn' means flash no. *nn*.

P\G	f08	f09	f11	f12	f13	f14	f15	f16	f17	f20	f21	f22	Avg.
f08	-	100/100	99/99	99/99	97/97	97/93	79/82	72/59	43/35	99/99	97/97	93/88	88/86
f09	100/100	-	99/99	99/99	99/99	99/99	97/91	91/84	60/53	97/99	97/99	97/96	94/92
f12	99/99	99/99	-	100/100	100/100	100/100	90/91	76/71	65/44	100/100	100/100	99/94	93/90
f13	99/99	99/99	100/100	-	100/100	100/100	100/99	93/90	76/72	100/100	100/100	100/99	97/96
f14	99/99	99/99	100/100	100/100	-	100/100	100/99	100/99	88/79	99/99	100/100	100/99	99/97
f15	99/99	99/99	100/100	100/100	100/100	-	100/99	100/97	96/87	99/99	100/100	100/99	99/98
f16	84/93	94/96	93/93	100/97	100/99	100/99	-	100/100	100/99	88/96	100/99	100/100	96/97
f16	69/75	87/90	78/69	90/93	100/97	100/99	100/100	-	100/99	69/69	90/94	100/100	89/89
f17	44/47	60/68	51/51	71/78	84/84	91/90	99/100	100/100	-	56/57	75/82	94/94	75/77
f20	97/99	97/99	100/100	100/100	100/99	100/100	90/91	74/76	68/51	-	100/100	99/94	93/92
f21	97/99	97/99	100/100	100/100	100/100	100/100	100/99	97/94	82/78	100/100	-	100/99	98/97
f22	90/97	97/96	96/96	100/99	100/99	100/99	100/100	100/100	99/90	97/96	100/99	-	98/97
Avg.	89/91	93/94	92/91	96/97	98/97	99/98	96/95	91/88	80/71	91/92	96/97	98/96	93/92

Table 2: Recognition rate obtained by our approach with the first and second constraints and Vetter's database. In each cell, the left number is when the 1st constraint is used and the right when the second constraint. 'P' means probe, 'G' means gallery, 'fnn' means flash no. *nn*.

	f08	f09	f11	f12	f13	f14	f15	f16	f17	f20	f21	f22	Avg.
Gly: Front f12, Prb: Front	99	99	100	-	100	100	97	93	78	100	100	99	96
Gly: Front f12, Prb: Side	85	89	88	94	96	96	88	81	68	86	95	91	88
Gly: Side f12, Prb: Front	92	91	99	97	85	87	72	53	33	94	96	87	82
Gly: Side f12, Prb: Side	100	100	100	-	100	100	99	85	63	100	100	100	95
Gly: Side f08, Prb: Side	-	100	100	100	99	97	72	59	35	100	100	90	86
Gly: Side f17, Prb: Side	26	41	37	57	76	84	100	100	-	43	65	91	66
Gly: Side f22, Prb: Side	75	97	88	99	100	100	100	100	100	91	100	-	95

Table 3: Recognition rate across pose. Front view is from camera 27, and side view from camera 05.