The Dimensionality of Scene Appearance

Rahul Garg[1], Hao Du[1], Steven M. Seitz[1], and Noah Snavely[2]

[1]University of Washington
[2]Cornell University

Abstract

Low-rank approximation of image collections (e.g., via PCA) is a popular tool in many areas of computer vision. Yet, surprisingly little is known justifying the observation that images of an object or scene tend to be low dimensional, beyond the special case of Lambertian scenes. This paper considers the question of how many basis images are needed to span the space of images of a scene under real-world lighting and viewing conditions, allowing for general BRDFs. We establish new theoretical upper bounds on the number of basis images necessary to represent a wide variety of scenes under very general conditions, and perform empirical studies to justify the assumptions. We then demonstrate a number of novel applications of linear models for scene appearance for Internet photo collections. These applications include, image reconstruction, occluder-removal, and expanding field of view.

1. Introduction

Real world scenes vary in appearance as a function of viewpoint, lighting, weather and other effects. What is the *dimensionality* of this appearance space? More specifically, suppose you stacked all photos taken of a particular scene as rows in a matrix – what is the rank of that matrix?[1]

It is well known that certain types of image collections tend to be low-rank in practice, and can be spanned via linear combination of a small number of basis views computed via tools like Principle Component Analysis (PCA) or Singular Value Decomposition (SVD). First exploited in the early work on eigenfaces [12, 27], these rank-reduction methods have become the basis for a broad range of successful applications in recognition [18, 16], tracking [9], background modeling [17], image-based rendering [28], BRDF modeling [10, 15], compression and other domains.

In spite of the wide-spread use of rank-reduction on images, however, there is little theoretical justification that appearance space should be low-rank in general. An exception is the case of Lambertian scenes, for which a number of elegant results exist. Shashua [23] proved that three images are sufficient to span the full range of images of a Lambertian scene rendered under distant lighting and a fixed viewpoint, neglecting shadows. Belhumeur and Kriegman [2] considered the case of attached shadows, observing that the valid images lie in a restricted range of 3D subspace which they called the *illumination cone*. Basri and Jacobs [1], and Ramamoorthi and Hanrahan [20] independently showed that illumination cone is well approximated with 9 basis images. Ramamoorthi more recently [19] improved this bound to 5 images, bringing the theory in line with empirical studies on the dimensionality of face images [4].

Very little is known, however, about the dimensionality of images of *real-world scenes*, composed of real shapes, BRDFs, and illumination conditions. Consider, for example, the images of tourist sites on Flickr [5], which exhibit vast changes in appearance. While it may seem difficult to prove anything about such collections, a key property of real-world scenes is that they are not random. In particular, man-made scenes tend to be dominated by a small number of surface orientations. And while BRDFs can be very complex, real BRDFs can be well-approximated by a low-rank linear basis [15]. Similar considerations apply for illumination; for example, studies have shown that the space of daylight spectra is roughly two- or three-dimensional [26]. Based on these observations, this paper introduces new theoretical upper bounds on the dimensionality of scene appearance (improving on previous results by Belhumeur and Kriegman [2]). While we make a few limiting assumptions (distant lighting, distant viewer, no cast shadows, interreflections or subsurface scattering), these results bring the theory to the point where it can capture much of the extreme variability in these Internet photo collections.

The highlights of this paper include a factorization framework for analyzing dimensionality questions, introduced in section 2. Using this framework, we prove new upper bounds on the number of basis images, allow-

[1]By dimensionality, we refer to linear dimensionality in this paper.

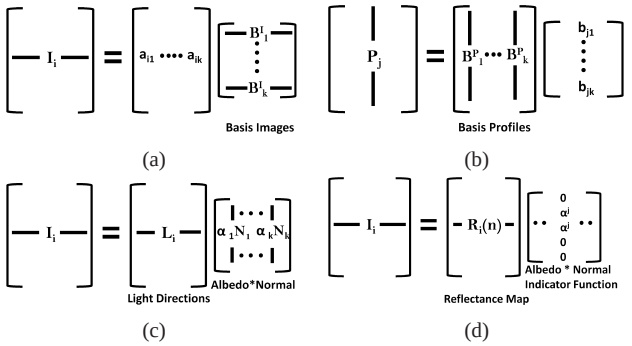


Figure 1: Four interpretations of factorization of image matrices. First, each image can be expressed as a linear combination of a set of basis images (a). Alternatively, the profile of each pixel can be expressed as a linear combination of a set of basis profiles (b). In the case of a Lambertian scene, the basis profiles and basis images assume special meaning (c). Finally, (d) shows the reflectance map interpretation.

ing for variable illumination direction and spectra, viewpoint, BRDFs, and convolution effects (e.g., blur). Importantly, all prior low-rank results for Lambertian scenes [23, 2, 1, 20, 19] do not apply under variations in light spectrum (even if the images are grayscale). We introduce new results that allow the light spectrum to vary in certain ways, greatly broadening the scope of application (e.g., to outdoors). Finally we demonstrate a number of interesting applications of low-rank linear models to problems in computational photography (Section 3.3).

2. Rank of the Image Matrix

We present theoretical results in this section. We first introduce a new framework to analyze the factorization of images (Section 2.1) which yields new insights and results in Section 2.2. Finally, we introduce wavelength (Section 2.3) bringing the theory closer to the real world images captured by cameras.

Throughout the paper, we assume that images are lit by distant light sources and observed from distant viewpoints. We ignore indirect illumination effects like transparent and translucent materials, interreflections, cast shadows and subsurface scattering. Our theory does account for attached shadows. Initially, we also make the assumption that images are taken from a fixed viewpoint, which we relax in Section 2.2.3.

2.1. Four Factorizations of the Image Matrix

Consider a set of m, n-pixel images of a scene, $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m$ taken under varying illumination conditions. Consider the $m \times n$ matrix $\mathbf{M}$ obtained by stacking $\mathbf{I}_1, \mathbf{I}_2, .., \mathbf{I}_m$ as rows of the matrix. Each row of $\mathbf{M}$ is an image, and each column describes the appearance of a single pixel, say x, under different illumination conditions, referred to as the *profile* of the pixel and denoted by $\mathbf{P}_x$, where $\mathbf{P}_x(i) = \mathbf{I}_i(x)$.

Consider a factorization of $\mathbf{M}$ into the product of two rank-k matrices:

$$\mathbf{M}_{m\times n} = \mathbf{C}_{m\times k}\mathbf{D}_{k\times n}. \tag{1}$$

Such a factorization may be obtained by PCA or SVD, for instance. We present four different interpretations of such a factorization, shown in Figure 1.

2.1.1 Basis Images

First, the rows of $\mathbf{D}$ can be interpreted as basis images, denoted by $\mathbf{B}^I$, and the rows of $\mathbf{C}$ can be interpreted as coefficients. This interpretation, shown in Figure 1(a), is commonly used. For instance, in work on eigenfaces [12], the eigenvectors obtained from PCA comprise the basis images (assuming mean subtracted data). Here each image $\mathbf{I}_i$ is a linear combination of basis images:

$$\mathbf{I}_i = \sum_{j=1}^{k} a_{ij}\mathbf{B}^I_j. \tag{2}$$

2.1.2 Basis Profiles

Another way to interpret this factorization is that each column (profile) $\mathbf{P}_j$ of $\mathbf{M}$ can be interpreted as a linear combination of columns of $\mathbf{C}$, with coefficients determined by columns of $\mathbf{D}$, as shown in Figure 1(b). In this interpretation, the columns of $\mathbf{C}$ form basis profiles, denoted by $\mathbf{B}^P$:

$$\mathbf{P}_j = \sum_{i=1}^{k} b_{ji}\mathbf{B}^P_i. \tag{3}$$

2.1.3 The Lambertian Case

For Lambertian scenes, neglecting *any* shadows, the rank of $\mathbf{M}$ is 3, and the basis profiles and the basis images assume a special meaning, shown in Figure 1(c). $\mathbf{D}$ is a $3 \times n$ matrix, where the j^{th} column of $\mathbf{D}$ encodes the normal times the albedo at the j^{th} pixel in the scene. $\mathbf{C}$ is a $m \times 3$ matrix where the i^{th} row encodes the lighting direction times the light intensity for the i^{th} image. Hence, the basis images represent scene properties (normals and albedos) and the basis profiles encode illumination properties. In particular, each basis profile contains the light intensity along a coordinate axis for each image.

2.1.4 The Reflectance Map Interpretation

The reflectance map [11], is defined for an image of a scene with a single BRDF as a function $R(\hat{\mathbf{n}})$ that maps scene normals to image intensity. $R(\hat{\mathbf{n}})$ can be encoded as an image

of a sphere with the same BRDF as the scene and taken from the same viewpoint under identical illumination conditions.

We denote the image of the sphere corresponding to $\mathbf{I}_i$ by $\mathbf{R}_i$, and write

$$\mathbf{I}_i{}^T = \mathbf{R}_i{}^T\mathbf{D} \tag{4}$$

where $\mathbf{D}$ is defined as:

$$\mathbf{D}(j,k) = \begin{cases} 0 & \text{if } \hat{\mathbf{n}}^k \neq \hat{\mathbf{m}}_j \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where $\hat{\mathbf{m}}_j$ is the normal at the j^{th} pixel of $\mathbf{R}_i$ and $\hat{\mathbf{n}}^k$ is the normal at the k^{th} pixel in the scene. The k^{th} column of $\mathbf{D}$ can be thought of as a *normal indicator function* $\mathbf{v}_k$.

It often happens that the BRDF is same across the scene save for a scaling factor (the albedo). This factorization can also incorporate per pixel albedos if we define the k^{th} column of $\mathbf{D}$ as $\alpha^k\mathbf{v}_k$ where α^k is the albedo of the k^{th} pixel.

Now, observing that $\mathbf{D}$ does not depend on i, one can write $\mathbf{M} = \mathbf{CD}$ where $\mathbf{C}$ contains the reflectance maps $\mathbf{R}_i$'s stacked as rows.

We can alternately define $R_i(\hat{\mathbf{n}})$ using the rendering equation which under our assumptions, can be written as

$$\mathbf{I}_i(x) = \int_\Omega \alpha^x\rho^x(\omega',\omega)L_{f_i}(\omega')(-\hat{\omega'}.\hat{\mathbf{n}}^x)_+\mathbf{d}\omega' \tag{6}$$

where the integral is over a hemisphere of inward directions ω', ω is the viewing direction for point x, ρ^x is the reflectance function at point x (evaluated at ω',ω), $L_{f_i}(\omega')$ is the light arriving from direction ω' for image $\mathbf{I}_i$, and $\hat{\mathbf{n}}^x$ is the normal at x. The + subscript on the dot product indicates that it is clamped below to 0 to account for attached shadows.

Given this, we can define $R_i(\hat{\mathbf{n}})$ as

$$R_i(\hat{\mathbf{n}}) = \int_\Omega \rho(\omega',\omega)L_{f_i}(\omega')(-\hat{\omega'}.\hat{\mathbf{n}})_+\mathbf{d}\omega' \tag{7}$$

where ρ^x has been replaced by ρ, as R_i represents a scene with a single BRDF.

2.2. Upper Bound on Rank of M

Belhumeur and Kriegman [2] proved that, given an arbitrary scene with a single material and k_n distinct normals, the space of images of the scene taken from a fixed, distant viewpoint with distant lighting and no cast shadows is *exactly* k_n-dimensional. This result justifies the use of linear models for real-world scenes. For instance, many man-made scenes consist of large planar regions (such as walls and ground), and therefore contain only a small number of distinct normals. Curved surfaces may also be approximated by piecewise planar surfaces.

We first show how an upper bound of k_n can be seen to hold true for a scene with a single BRDF using the reflectance map interpretation of the factorization and then extend it to a more general case.

From the reflectance map interpretation $\mathbf{M} = \mathbf{CD}$, it is easy to see that only k_n rows of $\mathbf{D}$ will be non-zero when the number of distinct normals in the scene is k_n. Hence, $rank(\mathbf{D}) \leq k_n$, which gives us an upper bound of k_n on $rank(\mathbf{M})$ as well.

Now consider the more general case where there are k_ρ materials and k_n normals in the scene. In this case, we first define reflectance maps corresponding to every BRDF for each image, i.e.,

$$\mathbf{I}_i{}^T = \sum_{l=1}^{k_\rho} \mathbf{R}_{il}{}^T\mathbf{D}_l \tag{8}$$

where $\mathbf{D}_l$ now encodes the distribution of normals corresponding to the l^{th} material, i.e.,

$$\mathbf{D}_l(j,k) = \begin{cases} 0 & \text{if } \hat{\mathbf{n}}^k \neq \hat{\mathbf{m}}_j \text{ or } \rho^k \neq \rho_l \\ \alpha^k & \text{otherwise} \end{cases} \tag{9}$$

Hence, $\mathbf{M}{=}\sum_{l=1}^{k_\rho}\mathbf{C}_l\mathbf{D}_l$, which implies that $rank(\mathbf{M}){\leq}k_\rho k_n$. More precisely $rank(\mathbf{M}){\leq}\sum_{l=1}^{k_\rho}N(l)$ where $N(l)$ is the number of orientations corresponding to material l. Hence we have proven the following:

Theorem 1 *Consider a scene with k_ρ different BRDFs and k_n distinct normals. Consider the images $\mathbf{I}_1,\mathbf{I}_2,\ldots,\mathbf{I}_m$ of the scene obtained from a fixed distant viewpoint under different distant illuminations $L_{f_1},L_{f_2},\ldots,L_{f_m}$. Assuming that there are no cast shadows, the rank of the matrix M obtained by stacking $\mathbf{I}_1,\mathbf{I}_2..\mathbf{I}_m$ as rows is at most $k_\rho k_n$.*

It is also instructive to write $\mathbf{M} = \sum_{l=1}^{k_\rho}\mathbf{C}_l\mathbf{D}_l$ in the form $\mathbf{M} = \mathbf{CD}$ so that basis images and basis profiles can be explicitly defined. This can be done by stacking $\mathbf{C}_l$ side by side, i.e., $\mathbf{C} = [\mathbf{C}_1|\mathbf{C}_2|..|\mathbf{C}_{k_\rho}]$ and stacking $\mathbf{D}_l$ one over another. Finally, we can remove all zero rows from $\mathbf{D}$ and corresponding columns from $\mathbf{C}$ leaving at most $k_\rho k_n$ rows in $\mathbf{D}$, which correspond to basis images, and basis profiles are the remaining columns of $\mathbf{C}$. The columns of $\mathbf{D}$ are of the form $\mathbf{d}_k = \alpha^k\mathbf{v}_k$ where $\mathbf{v}_k$ is a $0,1$ vector that can be thought of as a *normal-material indicator function*.

The result may be modified to accommodate anisotropic BRDFs as well. For anisotropic materials, one needs to parametrize by both the *orientation* and the normal. Hence, one can derive the same bound where k_n now refers to number of distinct orientations times normals in the scene.

In the following sections, we extend this result to a number of common scenarios.

2.2.1 Linear Families of BRDFs

While the world is composed of diverse materials, it has been argued [21, 15] that the space of BRDFs is low dimensional. We also verify this by conducting experiments on CUReT [3] database of BRDFs [6].

Thus, we now generalize to the case when ρ^x is contained in the linear span of $\{\rho_1, \rho_2, \ldots, \rho_{k_\rho}\}$, i.e., $\rho^x = \Sigma_{l=1}^{k_\rho} c_l(x)\rho_l$. In this case, $\mathbf{I}_i$ can be represented as a sum of matrix product, ${\mathbf{I}_i}^T = \sum_{l=1}^{k_\rho} {\mathbf{R}_{il}}^T \mathbf{D}_l$, where

$$\mathbf{D}_l(j,k) = \begin{cases} 0 & \text{if } \hat{\mathbf{n}}^k \neq \hat{\mathbf{m}}_j \\ \alpha^x c_l(x) & \text{otherwise} \end{cases} \tag{10}$$

Hence the upper bound of $k_\rho k_n$ still holds, i.e., rank is dimensionality of BRDF family times the number of normals.

2.2.2 Low-dimensional BRDFs

Certain BRDFs tend to be *low-dimensional*. For example, three basis images suffice to span images of a Lambertian scene captured under different lighting conditions, in the absence of shadows. Formally, we call a BRDF K-dimensional if the rank of the matrix $\mathbf{C}$ obtained by stacking reflectance maps obtained under arbitrary sampling of illumination conditions is always at most K. In the presence of such materials, the upper bound may be reduced to $\sum_{i=1}^{k_\rho} K(i)$, where $K(i)$ is the rank of the i^{th} BRDF.

We again used the CUReT database for estimating the dimensionality of each material in the database and found that for 49 of the 61 material, the reconstruction error is less than 10% using 9 basis vectors [6].

2.2.3 Varying Viewpoint

Given images taken from different viewpoints, it is trivial to extend the upper bound to $k_v k_\rho k_n$ where k_v is the number of distinct viewpoints. However, the bound of $k_\rho k_n$ holds true if we know the pixel corresponding to a point x' in the scene in every image. This correspondence can be found, for instance, if the camera parameters of each image and the 3D geometry of the scene are known. Using this, we can rearrange the pixels in each image so that the x^{th} pixel in every image corresponds to the same scene point. We assume that every scene point is seen by every image (We relax this assumption in Section 3.1). We again consider the rank of the matrix $\mathbf{M}$ obtained by stacking these rearranged images. The argument for Theorem 1 still holds, with R_{il} now defined as:

$$R_{il}(\hat{\mathbf{n}}) = \int_\Omega \rho_l(\omega', \omega_\mathbf{i}) L_{f_i}(\omega')(-\hat{\omega'}.\hat{\mathbf{n}})_+ \mathbf{d}\omega'. \tag{11}$$

where $\omega_\mathbf{i}$ is the viewing direction for image i.

2.2.4 Filtered Images

Many real-world images are blurry due to camera shake, or have been otherwise filtered (e.g., software sharpening). We extend the above result to filtered images.

Consider the family of images obtained by convolving image $\mathbf{I}(x,y)$ by an arbitrary $K \times K$ kernel $\mathbf{F}$. The resulting image can be expressed as $\mathbf{I_{conv}}(x,y) = \sum_{i=1}^{K}\sum_{j=1}^{K} \mathbf{F}(i,j)\mathbf{I}(x-i, y-j)$. Since the space of each of the shifted images $\mathbf{I}(x-i, y-j)$ is at most rank $k_\rho k_n$, it follows that the space of all filtered images of the scene is at most rank $K^2 k_\rho k_n$.

An important special case is the family of radially symmetric filters (e.g., blur, sharpen). These filters can be spanned by a few *basis filters* (The basis filters are simply circles of varying radii.)

Suppose that the family of filters we are concerned with can be spanned by k_f basis filters. Consider convolving each of the $k_\rho k_n$ basis images with each of the k_f basis filters to yield $k_f k_\rho k_n$ images. Any filtered image can then be expressed as a linear combination of these filtered basis images. Hence, the bound reduces to $k_f k_\rho k_n$.

2.3. Light Spectra

Up until now, we assumed that all measurements are done at a particular wavelength of light, and that the spectrum of light is constant over all images. We now consider the case when the camera sensors and light spectra vary between images. Surprisingly, in general, the appearance space of a simple Lambertian scene with a single infinite plane can have unbounded dimension, even for grayscale images. Albedos, which were before treated as fixed scalars for every pixel, are now functions of wavelength, allowing the scene to have arbitrary appearance for different wavelengths. In the general case, using a linear response model,

$$\mathbf{I}_i(x) = \int s_i(\lambda)\mathbf{I}_i(x,\lambda)d\lambda \tag{12}$$

where $s_i(\lambda)$ is the spectral response of the sensor i and $\mathbf{I}_i(x,\lambda)$ is the intensity of light of wavelength λ arriving at the sensor. We begin by analyzing the general case, then discuss results for some common special cases.

2.3.1 The General Case

Consider the matrix $\mathbf{M}$ obtained by stacking images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m$ captured by arbitrary sensors. We claim that the rank of $\mathbf{M}$ is bounded by $k_\rho k_n k_\alpha$, where k_α is the number of distinct albedos in the scene.

This result can again be derived from the reflectance map interpretation. We define a reflectance map corresponding to every pair of albedo and BRDF in the scene, with the incidence of normals encoded in the $\mathbf{D}$ matrix. More precisely, ${\mathbf{I}_i}^T = \sum_{h=1}^{k_\alpha}\sum_{l=1}^{k_\rho} \mathbf{R}_{ihl}^T \mathbf{D}_{hl}$ where $\mathbf{R}_{ihl}$ is the image of a sphere with BRDF ρ_l and albedo α_h captured under identi-

cal illumination conditions by the same sensor, and

$$\mathbf{D}_{hl}(j,k) = \begin{cases} 0 & \text{if } \hat{\mathbf{n}}^k \neq \hat{\mathbf{m}}_j \text{ or } \alpha^x \neq \alpha_h \text{ or } \rho^x \neq \rho_l \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

Again, we can write $\mathbf{M} = \sum_{h=1}^{k_\alpha} \sum_{l=1}^{k_\rho} \mathbf{C}_{hl}\mathbf{D}_{hl}$ by stacking up the $\mathbf{R}_{ihl}$'s,. It follows that $rank(\mathbf{M}) \leq k_\rho k_n k_\alpha$.

More generally, the albedos in a scene (as a function of wavelength) may be spanned by k_α basis albedos. It can be shown in a fashion similar to Section 2.2.1 that the bound of $k_\rho k_n k_\alpha$ extends to this case as well.

2.3.2 Light Sources with Constant Spectra

Belhumeur and Kriegman [2] showed that images of a Lambertian scene lit by light sources of identical spectra can be spanned by three basis images in the absence of shadows. We do a similar analysis in a more general setting.

Assume that (a) BRDFs do not depend on λ, (b) all images are lit by light sources with a constant spectrum across images (but with varying intensity and direction), and (c) all images are captured by identical sensors with spectral response $s(\lambda)$. Under these assumptions, the bound of $k_\rho k_n$ can be seen to hold true.

Under assumption (b), we can write $L_{f_i}(\omega', \lambda)$ as $K(\lambda)L'_{f_i}(\omega')$ and hence,

$$\mathbf{I}_i(x,\lambda) = K(\lambda) \int_\Omega \alpha^x(\lambda)\rho^x(\omega',\omega)L'_{f_i}(\omega')(-\hat{\omega'}.\hat{\mathbf{n}}^x)_+ \mathbf{d}\omega' \tag{14}$$

We can write $\mathbf{I}_i(x,\lambda) = K(\lambda)\sum_{j,k} a_{jk}(i)\mathbf{B}^I_{jk}(x,\lambda)$ by invoking the basis image representation for the expression in the integral (Theorem 1), where the number of basis images $\mathbf{B}^I_{jk}$'s is at most $k_\rho k_n$. Note that the coefficients do not depend on λ as wavelength dependent albedos are encoded in the basis images. Substituting into Eq. 12, we get

$$\mathbf{I}_i(x) = \sum_{j,k} a_{jk}(i) \int s(\lambda)K(\lambda)\mathbf{B}^I_{jk}(x,\lambda)d\lambda \tag{15}$$

which implies that $\mathbf{I}_i(x) = \sum_{jk} a_{jk}(i)\mathbf{B'}^I_{jk}(x)$ where the new basis images are obtained by integrating over λ, i.e., $\mathbf{B'}^I_{jk}(x) = \int s(\lambda)K(\lambda)\mathbf{B}^I_{jk}(x,\lambda)d\lambda$. Hence, these images can also be spanned by at most $k_\rho k_n$ basis images.

At first, these assumptions might appear too restrictive. We tested assumption (a) using the CUReT database and found strong support for it [6]. If albedos and camera spectral responses are unconstrained, the scene may have an unbounded rank. However, if the camera responses are similar, assumption (c) is a reasonable approximation. Other assumptions may be relaxed by extending the result. For instance, consider the case where a scene is lit by k_L light sources, each with its own spectrum that stays constant across all images. This can model outdoor illumination, which is often approximated as a combination of sunlight and skylight, each with its own spectrum [26]. Here, the bound can be seen to be $k_\rho k_n k_L$ by writing the illumination in the i^{th} image in the form $\sum_{l=1}^{k_l} K_l(\lambda)L_{lf_i}(\omega', \lambda)$.

Similarly, consider the case when $K(\lambda)$ varies from image to image but lies in a linear subspace of dimension k_s. For illumination in outdoor scenes, the spectra is well approximated by a two or three-dimensional subspace [26]. The bound can be shown to be $k_\rho k_n k_s$ in this case, by writing $K_i(\lambda) = \sum_{l=1}^{k_s} c_l(i)K_l(\lambda)$.

2.3.3 RGB Images

Images captured by conventional cameras contain three color channels. Consider RGB images $\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_m$, where we concatenate the channels together: $\mathbf{I}_i = [\mathbf{I}^1_i|\mathbf{I}^2_i|\mathbf{I}^3_i]$. Assume that each channel is captured by a separate sensor that is identical across all images,

$$\mathbf{I}^c_i(x) = \int s_c(\lambda)\mathbf{I}_i(x,\lambda)d\lambda \tag{16}$$

Consider the matrix $\mathbf{M}^c$ obtained by stacking channel c of all images. Under the assumptions of Section 2.3.2, we know that the rank of this matrix is bounded by $k_\rho k_n$ and it can be written as $\mathbf{M}^c = \mathbf{C}\mathbf{D}^c$ (The coefficients are embedded in the matrix $\mathbf{C}$ while the basis images $\mathbf{B'}^I_{jk}$ are stacked up in $\mathbf{D}$). Because $\mathbf{C}$ does not depend on c (From Eq. 15, we can see that $s_c(\lambda)$ is encoded in the basis images, i.e., $\mathbf{D}^c$), the rank of the matrix $\mathbf{M}$ obtained by concatenating the channels and stacking them is also bounded by $k_\rho k_n$ (We can write $[\mathbf{M}^1|\mathbf{M}^2|\mathbf{M}^3] = \mathbf{C}[\mathbf{D}^1|\mathbf{D}^2|\mathbf{D}^3]$).

In fact, we can go further and show that profiles corresponding to a particular pixel are identical across channels save for a scaling factor, i.e., there exists $k_c(x)$ for each channel such that $\mathbf{P}^c_x/k_c(x)$ is same for all c. This can be seen by substituting for $\mathbf{I}_i(x,\lambda)$ from Eq. 14 in Eq. 16 and writing:

$$\mathbf{P}^c_x(i) = k_c(x) \int_\Omega \rho^x(\omega',\omega)L'_{f_i}(\omega')(-\hat{\omega'}.\hat{\mathbf{n}}^x)_+ \mathbf{d}\omega' \tag{17}$$

where

$$k_c(x) = \int s_c(\lambda)K(\lambda)\alpha^x(\lambda)d\lambda \tag{18}$$

2.4. Summary

We started by proving an upper bound of $k_\rho k_n$ in Theorem 1 and then showed that the same bound holds for images taken from different viewpoints and for linear families of BRDFs. In Section 2.2.2, we showed that certain BRDFs allow the bound to be lowered. In Section 2.2.4, it was shown how blurry (filtered) images can be handled

by raising the bound. Finally, we introduced wavelength in Section 2.3. While in the most general case, the theoretical bound can shown to be $k_\rho k_n k_\alpha$, the bound of $k_\rho k_n$ holds under certain assumptions.

3. Results on Internet Photo Collections

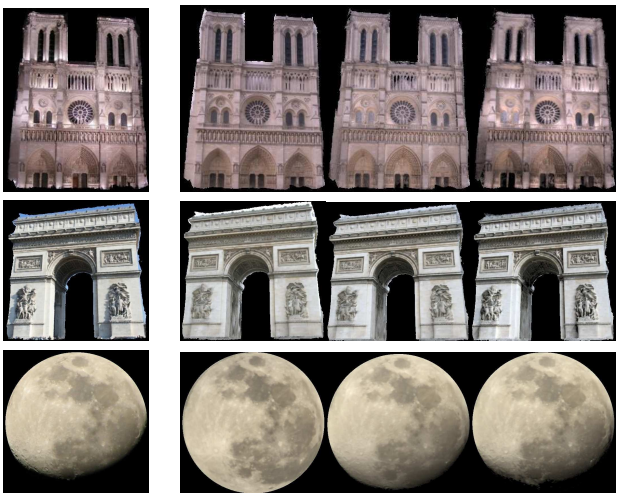
The theoretical results in Section 2 show that linear models can model a broad range of images of a scene. Much of the previous application of linear models has been to images captured in the lab under controlled conditions. Here, we apply it to a more challenging case, i.e., photos of popular locations downloaded from photo sharing websites [5]. The difficulties here stem from the wide variation in the scene appearance. Moreover, the images are captured using many different cameras and viewpoints.

3.1. Basis Computation

Because these photos are taken from different viewpoints, we first find pixel correspondences. We use the Structure from Motion (SfM) system of Snavely et al. [24] to recover the camera parameters. The 3D reconstruction uses the multi view stereo method of Goesele et al. [8]. The 3D models are *simplified* using qslim [7] to a mesh with $\sim 300,000$ faces. We use a simple representation where we associate a color corresponding to each mesh vertex. Images in this representation (which can be thought of as a *texture map*), can be treated in a fashion similar to images taken from a fixed viewpoint with mesh vertices assuming the role of pixels. However, a single image covers only a part of the scene, i.e., there is *missing data* in each texture map. To compute basis vectors with missing data, we use the EM based method of Srebro and Jaakkola [25] to compute SVD. However, the algorithm was found to be sensitive to initialization when the amount of missing data is large. We use the method of Roweis [22] which fills the missing data using EM based sensible PCA, to initialize.

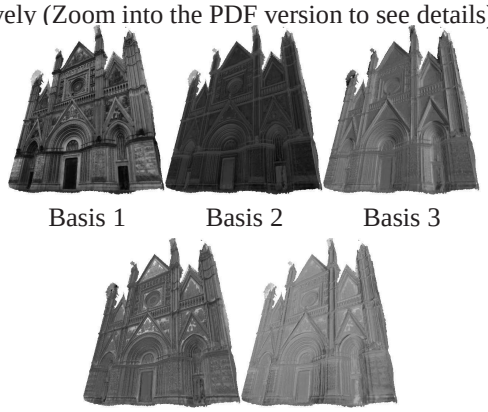
Internet photo collections are often dominated by people and other occluders who block the background scene. As our focus is modeling the scene and not the people, we start by manually removing images with significant occluders from which to compute a *clean* basis. We will show later how to handle occlusions in other images using this basis.

We cannot directly apply the ideas in Section 2.3.3 to these color images as the assumption of identical spectra and identical sensors does not hold for these collections. The selected clean set still has some outliers (e.g. cast shadows) and processing the three channels independently produces *rainbow* artifacts (examples in [6])) due to inconsistent fits between color channels. Instead, we make some simplifying assumptions that allow us to reconstruct the other channels given the reconstruction of one. Hence, we choose to process only the green channel of these images.



Rank 1 Rank 3 Rank 10

Figure 2: Left: example image from the dataset. Right: reconstruction obtained using $1, 3$ and 10 basis images respectively (Zoom into the PDF version to see details).



Basis 1 Basis 2 Basis 3

Basis 4 Basis 5

Figure 4: First 5 basis images for Orvieto. Basis 1 resembles the mean. Bases 2 and 3 model shading, and Bases 4 and 5 specularities.

Please refer to the technical report for details [6].

3.2. Evaluation

We present results on 6 datasets: Notre Dame Cathedral (212 images), Statue of Liberty (318 images), Orvieto Cathedral (228 images), Arc De Triomphe (268 images), Half Dome, Yosemite (95 images) and the Moon (259 images). The Moon presents an interesting case due to its retro-reflective nature. We are able to register the Moon images using SfM (There exists sufficient parallax for SfM to work [14]) and then fit a sphere to the 3D points obtained. The reconstruction is shown in [6].

All images were gamma corrected assuming $\gamma = 2.2$. We used the green channel of the images to find a basis. We observed that the reconstructions visually look reasonably good even with three or four basis vectors. With ten basis vectors, some of the finer details like specularities, self

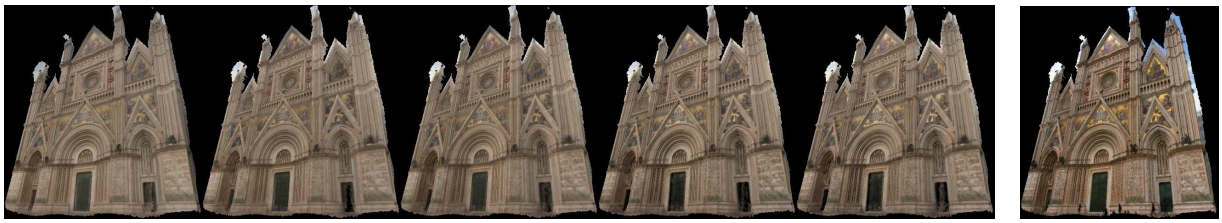


Figure 3: Reconstruction of an image of Orvieto Cathedral using $1, 2, 3, 4$ and 5 basis vectors. The image on the right is the original image.

shadowing, etc. are also modeled well (We use a basis of size ten to generate results in Section 3.3). There is little improvement in the reconstructions visually thereafter, but the numerical error stays at 12% even for 30 basis vectors. This error can be explained by the fact that even the *clean* set of images have a lot of noise. E.g., Half Dome's view is almost always partially occluded by trees.

Figure 2 shows an example image from these datasets and the corresponding reconstruction for $1, 3$ and 10 basis vectors. The top row (Notre Dame), shows that it becomes possible to model the appearance of night scenes using a larger basis. However, observe that such scenes have light sources close to the scene which violates our assumption of distant lighting. The configuration of lights is *fixed* across all night images and hence can be modeled by a single additional basis. The second row shows the reconstruction of an image of Arc De Triomphe demonstrating that it is possible to approximate cast shadowing using a larger basis. For the Moon, the appearance is modeled well using the first basis, while subsequent basis explain the shadows and the *texture at the terminator* [13]. We found that the model does not work well for the Half Dome dataset, as there are drastic appearance changes (such as seasonal snow).

An image of Orvieto Cathedral, whose facade is highly specular, is analyzed separately in Figure 3. Figure 4 shows the first 5 basis images. While the first basis simply looks like the mean image, the second and the third model the shading. The fourth and fifth bases seem to model view dependent effects (highlights). Again, note that specular highlight only on a part of the facade implies that the viewer is close to the scene which is a violation of our assumption of distant viewer. But as was the case in night scenes, a particular configuration of viewpoint and the lighting direction can be modeled by a single additional basis image.

3.3. Applications

We now show a few novel and interesting applications of linear scene appearance modeling.

3.3.1 View Expansion

As was mentioned in section 3.1, a single image might cover only part of the scene. However, since we use a method that can interpolate missing data, the derived basis images (and hence the reconstructions) cover the entire scene allowing us to hallucinate how the parts of the scene, not visible in the original image, would have appeared under similar illumination conditions (Figure 5).

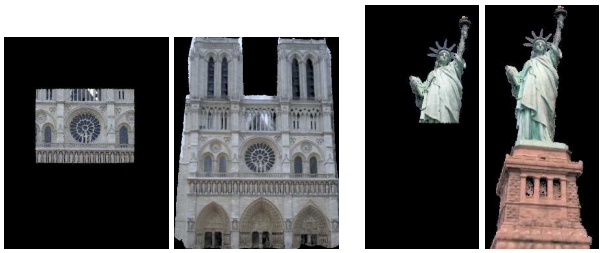


Figure 5: View Expansion: The left image in each image pair shows the original image with limited viewing area. The right image shows the reconstructed image.

3.3.2 Occluder Removal

Given the basis, we can project new images onto the computed basis. We choose a projection approach that is robust to outliers in the image. This allows us to handle occluders; for instance the bird in Figure 6(a). More precisely, in order to project a new image onto k basis images, we use a RANSAC approach where k pixels are sampled randomly and k coefficients are computed. The number of pixels that lie within a threshold of the original pixel values in the reconstruction obtained using these k coefficients are counted as *inliers*. Finally, the sample with the largest number of inliers is chosen and the estimate of coefficients is refined using all the inliers. Again, we can first reconstruct the green channel, and then reconstruct red and blue channels from it (explained in [6]). Some results are shown in Figure 6. See [6] for larger versions of these images.

4. Conclusion

This paper proved that scene appearance is low-rank under a variety of realistic conditions. These results are motivated by models of shape (particularly for man-made scenes), BRDFs, blur, and light spectra that approximate real-world scenes. We demonstrated the application of low-dimensional models to several large photo collections from

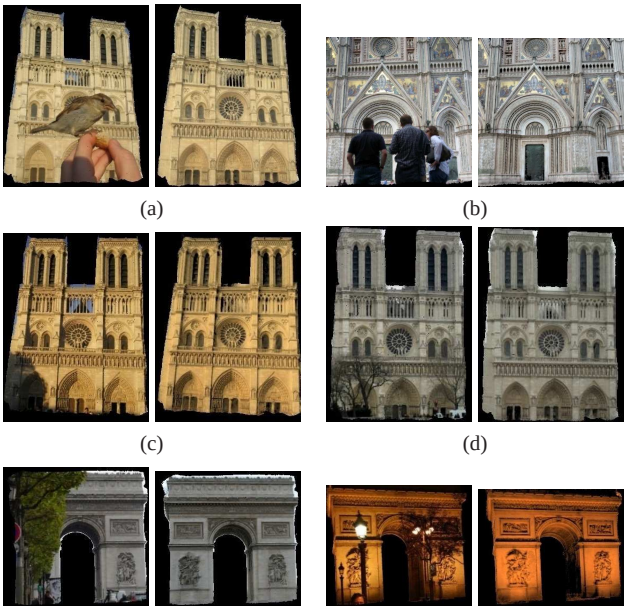


Figure 6: Occluder removal, where the occluder is removed and the scene behind is rendered under the same illumination conditions by robustly solving for basis image coefficients. (c) shows an example where the cast shadows are removed while the self shadows (which occur in large number of images and are modeled by the basis) are preserved.

the Internet, and showed compelling results for image reconstruction, view expansion, and occluder removal.

Acknowledgements: We thank Ryan Kaminsky for invaluable help with this project. This work was supported in part by National Science Foundation grant IIS-0811878, the Office of Naval Research, the University of Washington Animation Research Labs, and Microsoft. We are thankful to Flickr users whose photos we used.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[2] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions. *IJCV*, 28:270–277, 1998.

[3] K. Dana, B. Van-Ginneken, S. Nayar, and J. Koenderink. Reflectance and Texture of Real World Surfaces. *ACM Trans. on Graphics*, 18(1):1–34, 1999.

[4] R. Epstein, P. Hallinan, and A. Yuille. 5+/-2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. *IEEE Workshop on Physics-Based Modeling in Computer Vision*, 1995.

[5] http://www.flickr.com.

[6] R. Garg, H. Du, S. M. Seitz, and N. Snavely. The dimensionality of scene appearance. Technical Report UW-CSE-09-07-02, University of Washington, Dept. of Computer Science and Engineering, 2009.

[7] M. Garland and P. Heckbert. Surface simplification using quadric error metrics. *Proc. SIGGRAPH*, pages 209–216, 1997.

[8] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. Seitz. Multi-view stereo for community photo collections. *Proc. ICCV*, pages 1–8, 2007.

[9] G. Hager and K. Toyama. X vision: Combining image warping and geometric constraints for fast visual tracking. *Proc. ECCV*, pages 507–517, 1996.

[10] A. Hertzmann and S. Seitz. Shape and materials by example: a photometric stereo approach. *Proc. CVPR*, pages 533–540, 2003.

[11] B. K. Horn. *Robot Vision*. McGraw-Hill Higher Education, 1986.

[12] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *PAMI*, 12(1):103–108, 1990.

[13] J. J. Koenderink and S. C. Pont. Texture at the terminator. *3D Data Processing Visualization and Transmission, Int. Symp. on*, pages 406–415, 2002.

[14] http://en.wikipedia.org/wiki/libration.

[15] W. Matusik, H. Pfister, M. Br, and L. Mcmillan. A data-driven reflectance model. *ACM Trans. on Graphics*, 22:759–769, 2003.

[16] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearance. *IJCV*, 14(1):5–24, 1995.

[17] N. Oliver, B. Rosario, and A. Pentland. A Bayesian computer vision system for modeling human interactions. *PAMI*, 22(8):831–843, 2000.

[18] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. *Proc. CVPR*, pages 84–91, 1994.

[19] R. Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a lambertian object. *PAMI*, 24(10):1322–1333, 2002.

[20] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. *Proc. SIGGRAPH*, pages 117–128, 2001.

[21] R. Ramamoorthi and P. Hanrahan. Frequency space environment map rendering. *Proc. SIGGRAPH*, pages 517–526, 2002.

[22] S. Roweis. EM algorithms for PCA and SPCA. *Proc. NIPS*, 10:626–632, 1998.

[23] A. Shashua. Geometry and photometry in 3D visual recognition. Technical report, MIT AI Lab, 1992.

[24] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *Proc. SIGGRAPH*, pages 835–846, 2006.

[25] N. Srebro and T. Jaakkola. Weighted low-rank approximations. *Proc. NIPS*, pages 720–727, 2003.

[26] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? *Proc. CVPR*, pages 1–8, 2008.

[27] M. Turk and A. Pentland. Face recognition using eigenfaces. *Proc. CVPR*, pages 586–591, 1991.

[28] L. Wang, S. Kang, R. Szeliski, and H. Shum. Optimal texture map reconstruction from multiple views. *Proc. CVPR*, pages 347–354, 2001.