Department of Electrical
and
Computer Systems Engineering

Technical Report
MECSE-6-2003

An Analysis of Linear Subspace Approaches for Computer Vision and Pattern Recognition

P. Chen and D. Suter

MONASH
UNIVERSITY

An Analysis of Linear Subspace Approaches for Computer Vision and Pattern Recognition

Pei Chen and David Suter
Dept ECSE, P. O. Box 35, Monash University, Australia, 3800
{pei.chen, d.suter}@eng.monash.edu.au

***Abstract**: Linear subspace analysis (LSA) has become rather ubiquitous in the solution of a wide range of problems arising in pattern recognition and computer vision. The essence of these approaches is that certain structures are intrinsically (or approximately) low dimensional: for example, the factorization approach to the problem of structure from motion (SFM) and principal component analysis (PCA)-based approach to face recognition. In LSA, the singular value decomposition (SVD) is usually the basic mathematical tool. However, researchers have rather blindly used a SVD, without knowing the essential characteristics of its performance in the noise-corrupted environment. With the help of matrix perturbation theory, we present such an analysis here. First, the "denoising capacity" of the SVD is analysed. Second, we study the "learning capacity" of the LSA-based recognition system in a noise-corrupted environment. These results should help one to design more optimal systems in computer vision, particularly in tasks, such as SFM and face recognition.*

A direct application is that we clarify some issues regarding an optimal learning strategy for face recognition. Our analysis agrees with certain observed phenomenon, and these observations, together with our simulations, verify the correctness of our theory.

***Index terms**: Singular value decomposition, Linear subspaces, Principal component analysis, Structure from motion, Face recognition, Matrix perturbation, First-order perturbation, Multiple eigenvalue/singular value.*

1. Introduction

Linear subspace analysis has found application in many problems in computer vision and pattern recognition, where the high-dimensional representations of certain structures are intrinsically (or approximately) low dimensional. In this paper we focus on two very prominent problems: Structure from Motion (SFM), and PCA-based face recognition, although a whole host of other computer vision and pattern recognition tasks fall within the framework of our analysis.

1.1 Applications of Linear Subspace Analysis

In the SFM context, and in related multi-view analysis tasks, one extracts from the image sequence the coordinates of various tracked points (or other geometric features such a lines). These coordinates may be assembled into a measurement matrix, which is essentially low dimensional despite the matrix (itself) usually being physically huge. For example, under the affine models, the measurements are generally restricted to a rank-*4* subspace [13,14,16,17,23,32]. (Although the registered measurement matrix can be of rank *3* [16,17,23,32].) Another example is that the homographies of multiple planes between two views reside in a rank-*4* subspace [36,37]. Similarly, the homographies between two planes over multiple (*>2*) views lie in a rank-*4* subspace [29]. Moreover, the rank-*4* constraint also holds for the case of multiple-planes-over-multiple-views [36,37]. Exploitation of this low rank constraint is essential to solving for the quantities of interest (e.g., the 3-D structure of the scene being imaged).

Another particularly active area of computer vision research, also employing subspace analysis, is that of PCA-based face recognition* [6,10,33]. A human face, in typical applications, must be recognised despite illumination changes between the target image (to be recognised) and the database of candidate images. It has been observed that: "the variations between the images of the same face due to illumination and viewing direction are almost larger than image variations due to change in face identity" [19]. The issue of large illumination effects makes the problem of face recognition challenging [3,4,7,8,15,28]. In order to tackle this issue, PCA has been utilized to model the lighting variation in images; because it has been proved, experimentally [20-22,6,10,35] and theoretically [1,2,25,26], that the possible images of the same Lambertian object, under different lighting conditions, approximately concentrate in a low-dimensional subspace, although the dimension of the image set for an object is "equal to the number of distinct surface normals" [4]. Experimental observations [6,10,35] have also helped firmly establish that the images of the same face, produced under different lighting conditions, also approximately lie in a low-dimensional subspace. Similar approaches can be used in general object recognition and pose determination systems. A particularly influential example of such was the SLAM system [20-22], which captured the variations due to pose and illumination by a 20-dimensional (or less) subspace. Recently, it was proved, by

* Here, we have to clarify the difference between the common PCA [6,10,33] and linear subspace analysis [1-3]. In face recognition and related applications, several terminologies, like PCA [33], eigenface [33] and eigenimage [6,10], have been used for such dimensionality reduction techniques. PCA [6,10,33] works on the correlation matrix, where the mean of the images was first subtracted. While, in linear subspace analysis, we work directly on the original data [1-3], without subtracting their mean. Recently, some theoretical analysis [1,2,25,26] and experimental result [3] prove that better performance can be obtained directly by using the linear subspace analysis, without subtracting the mean. In section 5, we analyze the performance of the linear subspace analysis, without subtracting the mean, as in [26].

using spherical harmonics, that "all Lambertian reflectance functions obtained with arbitrary distant sources lie in close to a 9D linear subspace": Basri and Jacobs [1,2] and Ramamoorthi and Hanrahan [25,26].

1.2 Noise Effects

Despite such a plethora of applications where one expects, in principle, the measurements to be of low rank; it is widely understood that noise is inevitably introduced in the data. In the presence of noise, the matrix in question quickly becomes full rank. Thus, the matrix has to be fitted to its closest low-rank approximation. The SVD gives the best solution to this problem [9], measured by the Frobenius norm and 2-norm. The result is guaranteed to be optimal [24] if the noise is i.i.d. Gaussian. Not surprisingly, therefore, the SVD has become a widely used tool. For example, the factorization method [23,32] achieves a *Maximum Likelihood* affine reconstruction from multiple (>2) views, as pointed out in [11,27].

From a related point of view, the low-rank approximation can be regarded as a "denoising" tool, where we refer to the measure of the sum of squared difference (SSD)* between the noise-corrupted matrix (or the "denoised" matrix) and the noise-free matrix. Compared with a noisy matrix that is always of full rank, its low-rank approximation matrix, obtained by SVD, is always closer to the noise-free matrix, i.e. the underlying ground truth. For example, the multiview subspace constraint was utilized to improve the accuracy of recovered homographies, especially for those that have small regions [36,37].

---

* In image denoising, we usually use the terminology of mean square error (MSE).

Thus, linear subspace approximation is sometimes a model simplification and sometimes a denoising process (and often both, simultaneously).

1.3 Performance Questions

1.3.1 Denoising capacity of SVD

Although SVD is widely employed to fit a large matrix to its low-dimensional subspace, little work has been done to analyze the performance of SVD in such noise-corrupted cases. It is well known [9] that one can, by SVD, obtain the best solution to the low-rank approximation, measured by 2-norm or Frobenius-norm. However, the optimality is against the noise-corrupted matrix: the rank-*r* approximation matrix, obtained by SVD, is the closest rank-*r* matrix to the noise-corrupted matrix. However, we don't know its capacity of separating the signal from the noise. Supposing the noise level is small enough, how much signal is retained by keeping the largest *r* components? Or, how much noise has been reduced by discarding the other components? In this sense, we are *blindly* using SVD, *without knowing* its *denoising* capacity: how close is the low-rank approximation matrix to the noise-free matrix, or how close is the SVD-based subspace to the ground-truth subspace. The lack of such performance analysis impedes the careful design of optimal systems. A natural issue arising is to characterise the achieved accuracy with the growth in data (in the SFM context, this can be either through a growth in the number of frames analysed, or by a growth in the number of features tracked). In the factorization approach to SFM, it is widely accepted that more frames produce more accurate result than a few ("few" typically being little more than 3) frames. It was even claimed [31] that the 3D scene could be reconstructed to arbitrary accuracy given enough frames.

However, *what is the gain of adding the data from one extra frame to a very large measurement matrix? What happens as the number of the frames approaches infinity? Can the 3D scene be truly reconstructed with arbitrary accuracy? Can such arbitrary accuracy only be achieved by the increase of the frames (while the number features don't increase)? Is an increase in the number of frames the most efficient way to obtain an increase in accuracy?*

In the example of SFM, as suggested above we can also possibly augment the number of feature points, or we can augment the number frames, or we can do both: i.e., both the row and the column of the matrix can grow towards the infinite in size. However, in a related problem, the matrix consisting of the homographies over two views, is restricted to a class of $m\times 9$ matrices [36,37]. Such a matrix can only "grow" in one dimension, not both. We introduce some terminology to describe this difference: We call the matrix potentially-double-infinite if it has infinite rows and columns, and potentially-single-infinite for those who has constant rows (columns) and infinite columns (rows). This raises another question: *What is the difference between these two types of matrices in terms of the precision that can be achieved?*

In summary, the first aims of this paper are to analyze the *denoising* capacity of SVD, i.e., to identify the error that still resides in the low-rank approximation matrix and how this error relates to the growth of additional data.

1.3.2 Learning capacity of linear subspace analysis

Different questions, to those posed above, arise from the face recognition applications (including the object recognition and pose determination, and related applications). In the PCA-based face recognition approach, the eigenimage representation

relies on a compact approximation of the large image database (or "training" set), by spanning this set (approximately) with a few orthogonal basis images. Such an approach attempts to capture and characterise the essential object or face features, and their variations in appearance under lighting and pose changes. Although the “illumination cone” [4] (see also [38]) can be obtained by as little as three images, the result is usually not accurate enough. Firstly, there is inevitably some noise in the images, like quantization error. Secondly, it is difficult to satisfy the conditions in proposition 3 in that paper ([4]). Even if we can have three distinct light sources that can shed light on all the points of the surface, we can’t, in practice, exclude other light sources that cause attached or cast shadows on the subject. These considerations, plus general noise, have generally resulting in researchers trying to "learn" the eigenimages by a data reduction step applied to many "learning samples". Thus, many learning samples were needed to produce a good basis, for example, 66 images were used for one object [3]. *What is the relationship between the learning capacity and the size of the learning samples?* Note, the *learning* process will be explained in section 5, and a more detailed description of such learning processes can be found in [33,10,3].

Understanding the error, still residing in the basis images, will hopefully help to design the recognition system. Accurate basis images are desired because the recognition algorithm relies in projecting the test image, to be identified, on the basis images. Note that the test image itself contains noise. *Thus the noise in the LSA-based recognition system comes from two sources: one from the basis and the other from the test image. Do these two types of noise interfere with each other?*

The second aim of this paper is to present some theoretical analysis of the learning capacity of LSA-based recognition systems. Specifically, the error (measured by the sum of squared differences – SSD) of the LSA-based recognition system can be separated into two parts: one from the basis and the other from the test image, and we obtain some analytical results about their effects on the performance of the recognition system. We show that it is possible, theoretically, to design the optimal recognition system if we know the expectation of the test images.

In this paper, we answer the above questions by analyzing the performance of SVD in a noise-corrupted environment using the major tool of the matrix perturbation theory. In section 2, we first present our results. In section 3, some preliminary knowledge concerning the SVD and matrix perturbation theory is summarised. In sections 4 and 5, the justification of our results is developed, with the help of the matrix perturbation theory. In section 6, some simulation results are presented to testify to the correctness of our results and we explain some phenomena, observed by other researchers.

2. Major results

2.1 Notation

In the following, a matrix will be denoted by a bold capital letter, like **M**, and a bold lowercase letter represents a vector, e.g. **x**. $\mathbf{M}_i$ denotes the i^{th} column of **M**. A scalar entry in a vector or in a matrix will respectively be denoted by, for example, x_1 or $M_{1,2}$. $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, and $\mathbf{0}_{m,n}$ for a $m \times n$ zero-matrix. $\mathbf{e}_i$ is the i^{th} column of $\mathbf{I}_n$. $\mathbf{M}_{i:j,k:l}$, a notation from *Matlab*, denotes for the submatrix of **M**: the intersection of

the *i-to-j* rows and the *k-to-l* columns. A matrix **U**, $\mathbf{U} \in R^{m,n}$, is said to be orthonormal, *iff* $\mathbf{U}^T\mathbf{U} = \mathbf{I}_n$. The set of $m \times n$ orthonormal matrices is denoted by $O^{m,n}$. An orthonormal matrix will always be denoted by **U** or **V**. Two matrices, **M** and **N**, with same sizes, are said to be orthogonal to each other *iff* $\| \mathbf{M} \|_F = 1$, $\| \mathbf{N} \|_F = 1$, and $\sum M_{i,j} N_{i,j} = 0$. The Frobenius norm of a matrix **M** (or a vector) will be denoted as $\|\mathbf{M}\|_F$, where $\|\mathbf{M}\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$. $\mathbf{M}^r$ denotes the closest rank-*r* approximation of $\mathbf{M} \in R^{m,n}$, where $r \leq \min(m,n)$, as will be explained in section 3.1. The symbol "≈" means the first order perturbation, explained in appendix A. And, "≅" means the equality, in the sense of statistical expectation.

2.2 Major results

Here, we present the major results of this paper, by which we can answer the questions in the introduction. The justification of these results will be deferred until section 4 and section 5.

Result 1 (*Denoising capacity of SVD*): Suppose a matrix $\mathbf{A} \in \mathbf{R}^{m,n}$ lies in a low-dimensional, *r*, subspace. It is corrupted by i.i.d. Gaussian noise producing another matrix **B,** which is directly observed. Then, the error that still resides in the rank-*r* approximation matrix, $\mathbf{B}^r$, is

$$E \mid B_{i,j}^r - A_{i,j} \mid = \sigma \sqrt{\frac{r(m+n) - r^2}{mn}} \tag{1}$$

if the noise level σ, compared with the signal level, is small enough. Specially, as $m, n \to \infty$, the rank-r approximation of **B** approaches **A**, i.e. $\mathbf{B}^r \to \mathbf{A}$; and if $n \equiv k$ ($k \geq r$) and $m \to \infty$,

$$E \mid B_{i,j}^r - A_{i,j} \mid \to \sigma\sqrt{\frac{r}{k}} \tag{2}$$

Result 2 (*Learning capacity of LSA***)**: For a rank-r LSA-based recognition system, the "error measure" (the SSD) comes from two independent sources: the noise in the basis images and the noise in the test image. Specifically, the expectation of the SSD, over the learning samples, is:

$$(m-r)\sigma_t^2 + {(m-r)r\sigma_l^2}\Big/{n} \tag{3}$$

where m is the dimension of the object, n is the number of learning samples, σ_t and σ_l are the noise levels, for the test image and the learning samples respectively (Supposing both σ_t and σ_l are small enough, compared with the signal level σ_s). Moreover, for a *random* test image set, (3) is *optimal* among the size-n learning sets; and the size-n learning set is optimal *iff* it has r equal singular values.

Result 1 and result 2 will be motivated in section 4 and section 5 respectively.

3. Preliminary knowledge: SVD and perturbation theory

3.1 Singular value decomposition

The principle behind the SVD [9] states that any matrix, $\mathbf{M} \in R^{m,n}$, can be decomposed into

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}V^T \tag{4}$$

where $\mathbf{U} \in O^{m,m}$, $\mathbf{V} \in O^{n,n}$ and $\mathbf{\Sigma} = diag(\sigma_1, \sigma_2, \cdots, \sigma_p) \in R^{m,n}$, with $p = \min(m,n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Without loss of generality, suppose $m \geq n$. $\{\sigma_i^2 \mid i = 1,2,\cdots,n\}$ are the eigenvalues of $\mathbf{M}^T\mathbf{M}$, or the first *n* largest eigenvalues of $\mathbf{M}\mathbf{M}^T$. The first *n* left singular vectors of $\mathbf{M}$ are $\{\mathbf{U}_i \mid i = 1,2,\cdots,n\}$, where $\mathbf{U}_i$ is the eigenvector, corresponding to the eigenvalue of λ_i^2, of $\mathbf{M}\mathbf{M}^T$. Similarly, the right singular vectors of $\mathbf{M}$ are $\{\mathbf{V}_i \mid i = 1,2,\cdots,n\}$, where $\mathbf{V}_i$ is the eigenvector, corresponding to the eigenvalue of λ_i^2, of $\mathbf{M}^T\mathbf{M}$. Another important fact [9], is that one can easily construct $\mathbf{M}^k$, the closest rank *k* approximation of $\mathbf{M}$, measured by 2-norm or Frobenius-norm, by:

$$\mathbf{M}^k = \sum_{i=1}^{k} \sigma_i \mathbf{U}_i \mathbf{V}_i^T \tag{5}$$

Specifically,

$$\left\| \mathbf{M} - \mathbf{M}^k \right\|_2 = \sigma_{k+1} \tag{6}$$

$$\left\| \mathbf{M} - \mathbf{M}^k \right\|_F = \sqrt{\sum_{j=k+1}^{n} \sigma_j^2} \tag{7}$$

3.2 Perturbation theory

Only the perturbation theory concerning singular values/vectors is needed in this paper. However, we also include the perturbation theory concerning the eigenvalues/eigenvectors as a useful way to arrive at our results. With our objective, though, we need only consider *symmetric* matrices where the eigenvalues/eigenvectors are

concerned. To our best knowledge*, the perturbation expansion of the eigenvectors/singular-vectors is available only for those that correspond to a *simple* eigenvalue or singular value [30,34]. In this section, we review such theory, and present, in the next section, our new results for those that correspond to a *multiple* eigenvalue or singular value. In order to have a complete description of the perturbation theory, we give all the proofs, including those available in the textbooks [30,34], plus those within our new results. Detailed proofs are arranged in appendix A.

Theory 1 **[34]***: Consider a symmetric matrix,* $\mathbf{M} \in R^{m,m}$. *Suppose* $\mathbf{M}$ *has m distinct eigenvalues,* $\{\lambda_i \mid i = 1,2,\cdots,m\}$ *and the corresponding eigenvectors are* $\{\mathbf{x}_i \mid i = 1,2,\cdots,m\}$. *If* $\mathbf{M}$ *is perturbed by a matrix* $\mathbf{N}$*, the eigenvalues and the eigenvectors of* $\mathbf{M}+\mathbf{N}$ *are* $\{\lambda_i' \mid i = 1,2,\cdots,m\}$ *and* $\{\mathbf{x}_i' \mid i = 1,2,\cdots,m\}$ *respectively. Supposing every entry in* $\mathbf{N}$ *is small enough, the first-order perturbations of eigenvalues and eigenvectors are:*

$$\lambda_i' = \lambda_i + \beta_{i,i} \tag{8}$$

$$\mathbf{x}_i' = \mathbf{x}_i + \sum_{j \neq i} \frac{\beta_{j,i}}{\lambda_i - \lambda_j} \mathbf{x}_j \tag{9}$$

where $\beta_{i,j} = \mathbf{x}_i^T \mathbf{N} \mathbf{x}_j$.

Theorem 2 **[30]*:*** *Suppose* $\mathbf{A}$ (*not necessarily symmetric*) *is corrupted with* $\mathbf{N}$ *and we observe* $\mathbf{B}$*:* $\mathbf{B} = \mathbf{A} + \mathbf{N}$. *According to SVD, we have* $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, *where* $\mathbf{U} \in O^{m \times m}$, $\mathbf{\Sigma} = diag\{\kappa_1, \kappa_2, \cdots, \kappa_m\}$, $\mathbf{V} \in O^{m \times m}$. *Define* $\mathbf{C} = \mathbf{U}^{\mathrm{T}} \mathbf{N} \mathbf{V}$. *Suppose* κ_i *is a simple non-*

* Here, we'd like to express our appreciation to Prof. G. W. Stewart [32], who, by private correspondence,

zero singular value of **A**. *Then, the first order perturbations of the singular values* λ_i, *the right singular vector* $\mathbf{x}_i$, *and the left singular vector* $\mathbf{y}_i$, *of* **B** *are respectively*

$$\lambda_i = \kappa_i + C_{i,i} \tag{10}$$

$$\mathbf{x}_i = \mathbf{V}_i + \sum_{j \neq i} \frac{\lambda_j C_{j,i} + \lambda_i C_{i,j}}{\lambda_i^2 - \lambda_j^2} \mathbf{V}_j \tag{11}$$

$$\mathbf{y}_i = \mathbf{U}_i + \sum_{j \neq i} \frac{\lambda_i C_{j,i} + \lambda_j C_{i,j}}{\lambda_i^2 - \lambda_j^2} \mathbf{U}_j \tag{12}$$

The perturbation theory above, concerning the singular values/vectors, holds only for positive (and significantly large) singular values [30] (Note: singular values have to be non-negative.) In this paper, only LSA-based applications are of concern. In these rank-*r* problems, only the first *r* largest singular values are needed, where *r*<<*m*. Thus, we don't have to consider the behavior of the perturbation for the zero (or near zero) singular values.

3.3 New perturbation theory, corresponding to a multiple eigenvalue/ singular value

In this section, we present our result concerning the perturbation expansions, corresponding to the case where the matrix has at least one multiple eigenvalue/singular value.

First, we want to shed some light on the perturbation expansions concerning singular vectors that correspond to a multiple singular value. We do this by considering the perturbation expansions of the eigenvectors of a symmetric square matrix:

---

pointed this out to us.

Theorem 3*: Suppose* $\mathbf{M} \in R^{m,m}$, $\mathbf{M} = \mathbf{M}^T$, *and it has m eigenvalues* $\{\lambda_i\}$ *and m eigenvalues* $\{\mathbf{x}_i\}$, *which are orthogonal to each other*. Without loss of generality, suppose the first k eigenvalues of* **M** *are same,* $\lambda_i = \lambda$ *for* $i = 1,2,\cdots,k$. **M** *is corrupted with* **N**, *which, compared with* **M,** *is small enough. Define* $\mathbf{Q} = [\mathbf{x}_1,\cdots,\mathbf{x}_m]^T \mathbf{N} [\mathbf{x}_1,\cdots,\mathbf{x}_m]$. *Then, the first-order perturbation of the first k eigenvalues and eigenvectors of* **M+N** *are:*

$$\lambda_i' = \lambda + \delta_i \tag{13}$$

$$\mathbf{x}_i' = \sum_{j=1}^{k} S_{j,i}\mathbf{x}_j + \sum_{j=k+1}^{m} \frac{Q'_{j,i}}{\lambda - \lambda_j}\mathbf{x}_j \tag{14}$$

where δ_i *(supposing* $\delta_i \neq \delta_j$ *if* $i \neq j$*) and* $\mathbf{S}_i = [S_{1,i}, S_{2,i}, \cdots, S_{k,i}]^T$ *are the eigenvalues and eigenvectors of* $\mathbf{Q}_{1:k,1:k}$ *respectively, i.e.* $\mathbf{Q}_{1:k,1:k} = \mathbf{S}diag\{\delta_1,\cdots,\delta_k\}\mathbf{S}^{-1}$ *and* $\mathbf{S} = [\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_k]$. *Define* $\mathbf{Q}' = \begin{bmatrix} \mathbf{S}^{-1} & \\ & \mathbf{I}_{m-k} \end{bmatrix} \mathbf{Q} \begin{bmatrix} \mathbf{S} & \\ & \mathbf{I}_{m-k} \end{bmatrix}$. *The other m-k eigenvalues/eigenvectors can be obtained as in theorem 1.*

Following the same notation as used in theorem 2, we consider the perturbation expansion, where the matrix has at least one multiple *singular* value.

Theorem 4*:* **A**, **B**, **C** *and* $\mathbf{\Sigma}$ *are defined as those in theorem 2. Define* $\Omega = \mathbf{C} + \mathbf{\Sigma}$. *Without loss of generality, suppose the first k singular values of* **A** *are the same:*

---

* For an *r*-ple multiple eigenvalue, we, first, have its *r* eigenvectors, $\{\mathbf{x}_i \mid i = 1,\cdots,r\}$, which may not be orthogonal. Then, the *r* orthogonal eigenvectors can be obtained by applying Schmidt orthogonalization on $\{\mathbf{x}_i \mid i = 1,\cdots,r\}$.

$\{\kappa_i = \kappa \mid i = 1, \cdots, k\}$. *By SVD,* $\Omega_{1:k,1:k} = \mathbf{F}\mathbf{S}\mathbf{E}^T = \mathbf{F}diag\{S_1, \cdots, S_k\}\mathbf{E}^T$. *Let* $\mathbf{U}' = \begin{bmatrix} \mathbf{F} & \\ & \mathbf{I}_{m-k} \end{bmatrix}$, $\mathbf{V}' = \begin{bmatrix} \mathbf{E} & \\ & \mathbf{I}_{m-k} \end{bmatrix}$, *and* $\Omega' = \mathbf{U}'^T \Omega \mathbf{V}'$.

$$\mathbf{B} = (\mathbf{U}\mathbf{U}')\Omega'(\mathbf{V}\mathbf{V}')^T \tag{15}$$

The first order perturbation of the singular values, $\{\lambda_i'\}$, *right singular vectors* $\{\mathbf{x}_i'\}$, *and left singular vectors* $\{\mathbf{y}_i'\}$ *for* $0 \le i \le k$, *of* $\mathbf{\Omega}'$ *are respectively*

$$\lambda_i' = \Omega_{i,i}' = S_i \tag{16}$$

$$\mathbf{x}_i' = \mathbf{e}_i + \sum_{j=k+1}^{m} \frac{\kappa_j \Omega_{j,i}' + \kappa \Omega_{i,j}'}{\kappa^2 - \kappa_j^2} \mathbf{e}_j \tag{17}$$

$$\mathbf{y}_i' = \mathbf{e}_i + \sum_{j=k+1}^{m} \frac{\kappa \Omega_{j,i}' + \kappa_j \Omega_{i,j}'}{\kappa^2 - \kappa_j^2} \mathbf{e}_j \tag{18}$$

From (15), $\{\lambda_i'\}$ *are also the first k singular values of* **B**, *and, the right singular vectors* $\{\mathbf{x}_i\}$ *and left singular vectors* $\{\mathbf{y}_i\}$ *of* **B** *are respectively:* $\{\mathbf{V}\mathbf{V}'\mathbf{x}_i'\}$ *and* $\{\mathbf{U}\mathbf{U}'\mathbf{y}_i'\}$. *The perturbations, corresponding to other non-zero simple singular values, can be obtained as in theorem 2.*

4. Denoising capacity of SVD

In this and subsequent sections, we analyze the performance of SVD-related applications, as promised and sketched in the introduction and in section 2.2: (a) the *denoising* capacity of SVD; (b) and the *learning* capacity of LSA-based recognition system. We motivate our analysis by the perturbation theory concerning singular values and singular vectors, as outlined in section 3.2 and section 3.3.

4.1 Case of distinct singular values

First, we consider the simplest case: a square matrix with a few distinct non-zero singular values. **A**, **B**, **C**, and $\mathbf{\Sigma}$ are defined as in theorem 2: **A** is the signal matrix, **N** is the i.i.d. Gaussian noise matrix (with zero mean and σ^2 variance), **B**=**A**+**N**, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ and $\mathbf{C} = \mathbf{U}^\mathbf{T}\mathbf{N}\mathbf{V}$. Note **C** is still an i.i.d. Gaussian noise matrix (with zero mean and σ^2 variance). Further, define $\mathbf{\Omega} = \mathbf{C} + \mathbf{\Sigma}$. Then,

$$\mathbf{B} = \mathbf{U}\mathbf{\Omega}V^T \tag{19}$$

$\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$, defined as (11) and (12), are respectively the right and the left singular vectors of **B**; $\{\mathbf{x}'_i\}$ and $\{\mathbf{y}'_i\}$, defined as (A.4) and (A.5) in appendix A, are respectively the right and the left singular vectors of $\mathbf{\Omega}$. Obviously, from (19),

$$\mathbf{y}_i = \mathbf{U}\mathbf{y}'_i \text{ and } \mathbf{x}_i = \mathbf{V}\mathbf{x}'_i \tag{20}$$

And, also from (19), the singular values of **B**, $\{\lambda_i\}$, are same as the corresponding singular values of $\mathbf{\Omega}$, $\{\lambda'_i\}$.

Suppose that the noise-free matrix **A** should have a rank of *r*, i.e. $\mathbf{A} = \sum_{i=1}^{r} \kappa_i \mathbf{U}_i \mathbf{V}_i^T$.

Combining (5), (19) and (20), the closest rank-*r* approximation of **B** is

$$\mathbf{B}^r = \sum_{i=1}^{r} \lambda_i \mathbf{y}_i \mathbf{x}_i^T = \mathbf{U}(\sum_{i=1}^{r} \lambda'_i \mathbf{y}'_i \mathbf{x}'^T_i)\mathbf{V}^T = \mathbf{U}\Omega^r\mathbf{V}^T \tag{21}$$

Then

$$\| \mathbf{B}^r - \mathbf{A} \|_F^2 = \left\| \sum_{i,j} (\Omega^r_{i,j} - \Lambda_{i,j}) \mathbf{U}_i \mathbf{V}_j^T \right\|_F^2 \tag{22}$$

where $\mathbf{\Lambda} \in R^{m,m}$, $\Lambda_{i,j} = 0$ if $(i,j) \notin \{(1,1),(2,2),\cdots,(r,r)\}$ and $\Lambda_{i,i} = \kappa_i$ for $(i = 1,\cdots,r)$.

Due to the mutual orthonormality among any $\mathbf{U}_i \mathbf{V}_j^T$, we have the following formula:

$$\| \mathbf{B}^r - \mathbf{A} \|_F^2 = \left\| \mathbf{\Omega}^r - \mathbf{\Lambda} \right\|_F^2 \tag{23}$$

According to the perturbation theory in section 3.2, the first order perturbation of $\lambda_i' \mathbf{y}_i' \mathbf{x}_i'^T$ (to see the definition of $\{\lambda_i'\}$, $\{\mathbf{x}_i'\}$ and $\{\mathbf{y}_i'\}$ in (A.3-A.5), in appendix A), for example $\lambda_1' \mathbf{y}_1' \mathbf{x}_1'^T$, is:

$$\lambda_1' \mathbf{y}_1' \mathbf{x}_1'^T \approx \begin{bmatrix} \lambda_1' & \lambda_1' \mathbf{a}^T \\ \lambda_1' \mathbf{b} & \mathbf{0} \end{bmatrix} \approx \begin{bmatrix} \kappa_1 + C_{1,1} & \kappa_1 \mathbf{a}^T \\ \kappa_1 \mathbf{b} & \mathbf{0} \end{bmatrix} \tag{24}$$

where $\mathbf{a} = \begin{bmatrix} \frac{\kappa_2 C_{2,1} + \kappa_1 C_{1,2}}{\kappa_1^2 - \kappa_2^2} & \cdots & \frac{\kappa_r C_{r,1} + \kappa_1 C_{1,r}}{\kappa_1^2 - \kappa_r^2} & \frac{C_{1,r+1}}{\kappa_1} & \cdots & \frac{C_{1,m}}{\kappa_1} \end{bmatrix}^T$, and

$\mathbf{b} = \begin{bmatrix} \frac{\kappa_1 C_{2,1} + \kappa_2 C_{1,2}}{\kappa_1^2 - \kappa_2^2} & \cdots & \frac{\kappa_1 C_{r,1} + \kappa_r C_{1,r}}{\kappa_1^2 - \kappa_r^2} & \frac{C_{r+1,1}}{\kappa_1} & \cdots & \frac{C_{m,1}}{\kappa_1} \end{bmatrix}^T$. Note, 2-order and higher-order terms have been dropped. Similarly, the first-order perturbations of $\lambda_i' \mathbf{y}_i' \mathbf{x}_i'^T$, for $(i = 2,\cdots,r)$, can be obtained.

By combining such results as (24), it is easy to obtain

$$\mathbf{\Omega}^r - \mathbf{\Lambda} = \mathbf{Y} \tag{25}$$

where $\mathbf{Y} = \begin{bmatrix} C_{1,1} & \cdots & C_{1,r} & C_{1,r+1} & \cdots & C_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{r,1} & \cdots & C_{r,r} & C_{r,r+1} & \cdots & C_{r,m} \\ C_{r+1,1} & \cdots & C_{r+1,r} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{m,1} & \cdots & C_{m,r} & 0 & \cdots & 0 \end{bmatrix}$

$$E \| \mathbf{B}^r - \mathbf{A} \|_F^2 = E \left\| \mathbf{Y} \right\|_F^2 = \sum E Y_{i,j}^2 = (2rm - r^2)\sigma^2 \tag{26}$$

$$E \mid B^r_{i,j} - A_{i,j} \mid = \sigma \frac{\sqrt{2rm - r^2}}{m} \tag{27}$$

Obviously, (27) is a special case of (1) for square matrices, where n=m.

4.2 Case of multiple singular value

As in the theorem 4, suppose the first k ($k \le r$) singular values of **A** are same. Following the notation in theorem 4, we similarly have, as done in section 4.1:

$$\mathbf{B}^r = \sum_{i=1}^{r} \lambda_i \mathbf{y}_i \mathbf{x}_i^T = (\mathbf{U}\mathbf{U}')(\sum_{i=1}^{r} \lambda'_i \mathbf{y}'_i \mathbf{x}'^T_i)(\mathbf{V}\mathbf{V}')^T = (\mathbf{U}\mathbf{U}')\Omega'^r(\mathbf{V}\mathbf{V}')^T = \mathbf{U}\Omega'^r\mathbf{V}^T \tag{28}$$

By the same techniques as in section 4.1, the first-order perturbation of Ω'^r has the following form (please note the similar form between (17, 18) and (A.4, A.5) and the fact that the up-left $k \times k$ submatrix of Ω', $\Omega'_{1:k,1:k}$, is a diagonal matrix.):

$$\Omega'^r \approx \begin{bmatrix} \Omega'_{1,1} & \cdots & \Omega'_{1,r} & \Omega'_{1,r+1} & \cdots & \Omega'_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega'_{r,1} & \cdots & \Omega'_{r,r} & \Omega'_{r,r+1} & \cdots & \Omega'_{r,m} \\ \Omega'_{r+1,1} & \cdots & \Omega'_{r+1,r} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega'_{m,1} & \cdots & \Omega'_{m,r} & 0 & \cdots & 0 \end{bmatrix} \tag{29}$$

Then,

$$\Omega^r = \mathbf{U}'\Omega'^r\mathbf{V}'^T = \begin{bmatrix} \mathbf{F} & \\ & \mathbf{I}_{m-k} \end{bmatrix} \begin{bmatrix} \Omega'_{1,1} & \cdots & \Omega'_{1,r} & \Omega'_{1,r+1} & \cdots & \Omega'_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega'_{r,1} & \cdots & \Omega'_{r,r} & \Omega'_{r,r+1} & \cdots & \Omega'_{r,m} \\ \Omega'_{r+1,1} & \cdots & \Omega'_{r+1,r} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega'_{m,1} & \cdots & \Omega'_{m,r} & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mathbf{E}^T & \\ & \mathbf{I}_{m-k} \end{bmatrix}$$

$$= \begin{bmatrix} \Omega_{1,1} & \cdots & \Omega_{1,r} & \Omega_{1,r+1} & \cdots & \Omega_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega_{r,1} & \cdots & \Omega_{r,r} & \Omega_{r,r+1} & \cdots & \Omega_{r,m} \\ \Omega_{r+1,1} & \cdots & \Omega_{r+1,r} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Omega_{m,1} & \cdots & \Omega_{m,r} & 0 & \cdots & 0 \end{bmatrix} = \mathbf{\Lambda} + \mathbf{Y}$$

where $\mathbf{\Lambda}$ and $\mathbf{Y}$ are same as those in (25), and $\mathbf{E}$ and $\mathbf{F}$ are defined in theorem 4. Obviously, the same result, as (27), has been obtained.

4.3 Extension to the rectangular matrix

As stated in section 3.2, we only have to consider the first *r* largest singular values. Thus, in the cases of rectangular matrices, the perturbation theory concerning the singular values/vectors still holds and the performance analysis, in section 4.1 and section 4.2, can be easily extended to the rectangular matrices. Here, we only present the final result, omitting the tedious mathematical deduction, which is almost same as that in section 4.1 and section 4.2. Suppose the signal matrix, $\mathbf{A}$, and noise matrix, $\mathbf{N}$, lie in $R^{m,k}$ $(m,k \geq r)$. Other conditions stay same as in section 4.1.

$$\mathbf{\Omega}^r - \mathbf{\Lambda} = \mathbf{Y}$$

where $\mathbf{Y} = \begin{bmatrix} C_{1,1} & \cdots & C_{1,r} & C_{1,r+1} & \cdots & C_{1,k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{r,1} & \cdots & C_{r,r} & C_{r,r+1} & \cdots & C_{r,k} \\ C_{r+1,1} & \cdots & C_{r+1,r} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{m,1} & \cdots & C_{m,r} & 0 & \cdots & 0 \end{bmatrix}$

$$E \,||\, \mathbf{B}^r - \mathbf{A} \,||_F^2 = E\left\|\mathbf{Y}\right\|_F^2 = \sum EY_{i,j}^2 = (rm + rk - r^2)\sigma^2 \tag{30}$$

$$E \mid B_{i,j}^r - A_{i,j} \mid = \sigma\sqrt{\frac{rm + rk - r^2}{mk}} \tag{31}$$

which is the same as (1). As $m \to \infty$, while *k* is a constant, $E \mid B_{i,j}^r - A_{i,j} \mid \to \sigma\sqrt{\frac{r}{k}}$, a non-zero constant***. As suggested by (2), it is impossible to reconstruct 3D scene to arbitrary accuracy by the factorization method using an affine camera model, by only increasing the number of the frames (while keeping the number of the feature points unchanged).*** This contrasts with the claim that 3D scene could be reconstructed to arbitrary accuracy given enough frames [31]. However, we recognise the need for caution, our setting is not exactly the same as that of [31], where the perspective model was adopted.

5. Learning capacity of LSA-based recognition system

In this section, we analyze the performance of LSA-based recognition systems when the test image is correctly identified. Under such an assumption, there is still some error, as stated in the introduction, because of the noise in the basis images and the noise in the test image. In the following, we analyze the effect of this noise on the recognition system (also by the means of first-order perturbation theory).

Before we motivate the performance analysis of the LSA-based recognition system, we present a simple description of the LSA-based face recognition algorithm [3,7,8]. It consists of two steps: the off-line learning stage and the on-line recognition stage. In the learning stage, the image basis is obtained in this way: a set of learning images for one face is arranged as a learning matrix **A** so that each image is regarded as one column of the learning matrix **A**. Suppose the face image has a dimension of *m,* and *n* learning samples are collected. $\mathbf{A} \in R^{m,n}$. The *r* (*r*<<*m* and $r \le n$) basis images can be obtained as the first *r* left singular vectors of **A**, which correspond to the *r* largest singular values. In the on-line recognition stage, a test image is projected on the *r* basis images and its distance to the image basis is used for recognition.

5.1 Perturbation of the basis images

First, we analyze the learning stage, by using the matrix perturbation theory in sections 3.2 and 3.3. By SVD, the low-dimension subspaces, $\mathbf{U}'^{r} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_r]$ and $\mathbf{V}'^{r} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r]$, as defined in theorem 2, are obtained. In some cases, such as in face recognition, the consequent step is contingent on an accurate basis. Here, we only consider the subspace $\mathbf{U}'^{r}$: $\mathbf{U}'^{r} = \mathbf{UH}$, where

$$\mathbf{H} = \begin{bmatrix} 1 & \frac{\kappa_2 C_{1,2} + \kappa_1 C_{2,1}}{\kappa_2^2 - \kappa_1^2} & \cdots & \frac{\kappa_r C_{1,r} + \kappa_1 C_{r,1}}{\kappa_r^2 - \kappa_1^2} \\ \frac{\kappa_1 C_{2,1} + \kappa_2 C_{1,2}}{\kappa_1^2 - \kappa_2^2} & 1 & \vdots & \vdots \\ \vdots & \cdots & \ddots & \frac{\kappa_r C_{r-1,r} + \kappa_{r-1} C_{r,r-1}}{\kappa_r^2 - \kappa_{r-1}^2} \\ \frac{\kappa_1 C_{r,1} + \kappa_r C_{1,r}}{\kappa_1^2 - \kappa_r^2} & \cdots & \frac{\kappa_{r-1} C_{r,r-1} + \kappa_r C_{r-1,r}}{\kappa_{r-1}^2 - \kappa_r^2} & 1 \\ \frac{C_{r+1,1}}{\kappa_1} & \frac{C_{r+1,2}}{\kappa_2} & \cdots & \frac{C_{r+1,r}}{\kappa_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{C_{m,1}}{\kappa_1} & \frac{C_{m,2}}{\kappa_2} & \cdots & \frac{C_{m,r}}{\kappa_r} \end{bmatrix} \quad (32)$$

Note: From (32), we can roughly see that, for different singular vectors $\{\mathbf{U}_i\}$, their perturbations $\{\mathbf{y}_i\}$ have been corrupted, to a different extent, which depends on their *strength* (more formally, on their corresponding singular values). If *m*>>*r*, the corruption comes mostly from $\{\mathbf{U}_i \,|\, (i > r)\}$. Obviously, the corruption in $\mathbf{y}_i$ $(i \le r)$ is approximately inversely proportional to its corresponding singular value, κ_i. Thus, $\mathbf{y}_1$ can be considered *cleanest*, while $\mathbf{y}_r$ the *dirtiest*. In section 5.2, we will return to this point when the projection error is analyzed.

Furthermore, to decompose $\mathbf{H}$ into: $\mathbf{H} = \mathbf{E} + \mathbf{F} + \mathbf{G}$, where $\mathbf{E} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{0}_{(m-r)\times r} \end{bmatrix}$

$$\mathbf{F}=\begin{bmatrix} 0 & \frac{\kappa_2 C_{1,2}+\kappa_1 C_{2,1}}{\kappa_2^2-\kappa_1^2} & \cdots & \frac{\kappa_r C_{1,r}+\kappa_1 C_{r,1}}{\kappa_r^2-\kappa_1^2} \\ \frac{\kappa_1 C_{2,1}+\kappa_2 C_{1,2}}{\kappa_1^2-\kappa_2^2} & 0 & \vdots & \vdots \\ \vdots & \cdots & \ddots & \frac{\kappa_r C_{r-1,r}+\kappa_{r-1} C_{r,r-1}}{\kappa_r^2-\kappa_{r-1}^2} \\ \frac{\kappa_1 C_{r,1}+\kappa_r C_{1,r}}{\kappa_1^2-\kappa_r^2} & \cdots & \frac{\kappa_{r-1} C_{r,r-1}+\kappa_r C_{r-1,r}}{\kappa_{r-1}^2-\kappa_r^2} & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad (33)$$

$$\mathbf{G}=\begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \vdots & \vdots \\ \vdots & \cdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \\ \frac{C_{r+1,1}}{\kappa_1} & \frac{C_{r+1,2}}{\kappa_2} & \cdots & \frac{C_{r+1,r}}{\kappa_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{C_{m,1}}{\kappa_1} & \frac{C_{m,2}}{\kappa_2} & \cdots & \frac{C_{m,r}}{\kappa_r} \end{bmatrix} \quad (34)$$

5.2 Projection of a new test image on the basis images

The underlying noise-free subspace $\mathbf{U}^r=\mathbf{U}\begin{bmatrix}\mathbf{I}_r \\ \mathbf{0}_{(m-r)\times r}\end{bmatrix}=\mathbf{UE}$. Suppose a noise corrupted test image **p**, to be identified, is observed, and the underlying truth is **q**: $\mathbf{q}=\mathbf{Uf}$ and $\mathbf{p}=\mathbf{U}(\mathbf{f}+\mathbf{g})$. Because $\mathbf{q}\in\mathbf{U}^r$, only the first r components of **f** are possibly non-zeroes, i.e. $\mathbf{f}=[f_1,f_2,\cdots,f_r,0,\cdots,0]^T$. In practice, the noise-corrupted test image has to be projected on the noise-corrupted basis in the recognition system because the noise free basis is always unknown. More formally, the projection error of **p** on $\mathbf{U}'^r$ is used:

$$\begin{aligned}\mathbf{p}-\mathbf{U}''^{r}\mathbf{U}''^{r\,T}\mathbf{p} &= \mathbf{U}(\mathbf{f}+\mathbf{g})-\mathbf{U}(\mathbf{E}+\mathbf{F}+\mathbf{G})(\mathbf{E}+\mathbf{F}+\mathbf{G})^{T}\mathbf{U}^{T}\mathbf{U}(\mathbf{f}+\mathbf{g}) \\ &\approx \mathbf{U}[\mathbf{g}-\mathbf{E}\mathbf{E}^{T}\mathbf{g}-(\mathbf{E}\mathbf{F}^{T}+\mathbf{E}\mathbf{G}^{T}+\mathbf{F}\mathbf{E}^{T}+\mathbf{G}\mathbf{E}^{T})\mathbf{f}] \\ &= \mathbf{U}[\mathbf{g}-\mathbf{E}\mathbf{E}^{T}\mathbf{g}-(\mathbf{E}\mathbf{G}^{T}+\mathbf{G}\mathbf{E}^{T})\mathbf{f}] \\ &= \mathbf{U}[\mathbf{g}'-\mathbf{G}\mathbf{f}']\end{aligned} \tag{35}$$

where $\mathbf{g}'$ has same components as $\mathbf{g}$, except its first r zeroes, i.e. $\mathbf{g}'=[0,0,\cdots,0,g_{r+1},\cdots,g_m]^T$. And $\mathbf{f}'=[f_1,f_2,\cdots,f_r]^T$. Note, in (35), the 2-order and higher-order terms have been dropped: **F**, **G**, and **g** can possibly approach **0**. From (35),

$$\mathbf{p}-\mathbf{U}''^{r}\mathbf{U}''^{r\,T}\mathbf{p}=\mathbf{U}[\mathbf{g}',\mathbf{C}'_1,\mathbf{C}'_2,\cdots,\mathbf{C}'_r][1,-\mathbf{h}^T]^T \tag{36}$$

$$\left\|\mathbf{p}-\mathbf{U}''^{r}\mathbf{U}''^{r\,T}\mathbf{p}\right\|_F=\left\|[\mathbf{g}',\mathbf{C}'_1,\mathbf{C}'_2,\cdots,\mathbf{C}'_r][1,-\mathbf{h}^T]^T\right\|_F \tag{37}$$

where $\mathbf{C}'_i=[0,\cdots,0,C_{r+1,i},C_{r+2,i},\cdots,C_{m,i}]^T$ and $\mathbf{h}=[f_1/\kappa_1,\cdots,f_r/\kappa_r]^T$.

We suppose the basis is obtained from n learning samples, i.e., the learning matrix is $\mathbf{A}\in R^{m,n}$, and each entry of **A** has energy of σ_s^2, and is corrupted with i.i.d. Gaussian noise with energy of σ_l^2. It is also assumed that the test image has energy of σ_s^2 and is corrupted with noise of σ_t^2. $\sum_{i=1}^{r}\kappa_i^2\cong mn\sigma_s^2$, $\sum_{i=1}^{m}\sum_{j=1}^{n}C_{i,j}^2\cong mn\sigma_l^2$, $\sum_{i=1}^{r}f_i^2\cong m\sigma_s^2$, and $\sum_{i=1}^{m}g_i^2\cong m\sigma_t^2$. $\|\mathbf{g}'\|_F\cong\sqrt{m-r}\,\sigma_t$ and $\|\mathbf{C}'_i\|_F\cong\sqrt{m-r}\,\sigma_l$. Due to the independence among $\{\mathbf{g}',\{\mathbf{C}'_i\mid i=1,\cdots,r\}\}$, (37) becomes

$$\left\|\mathbf{p}-\mathbf{U}''^{r}\mathbf{U}''^{r\,T}\mathbf{p}\right\|_F^2\cong(m-r)\sigma_t^2+(m-r)\sigma_l^2\sum_{i=1}^{r}\frac{f_i^2}{\kappa_i^2} \tag{38}$$

$$\|\mathbf{p}\|_F^2\cong m(\sigma_s^2+\sigma_t^2) \tag{39}$$

Obviously, from (38), the projection error is contingent on the relationship between $\{f_i\}$ and $\{\kappa_i\}$. From **G** in (34), and (38), it can be concluded that the basis $\mathbf{y}_1$ that corresponds to the largest singular value is the *cleanest*, and that the basis $\mathbf{y}_r$ that corresponds to the least singular value is the *dirtiest*. The *cleanness* of the j^{th} basis $\mathbf{y}_j$, here, is measured by the projection error, in (38), which is introduced by the j^{th} unit-norm basis image. For a random test image, the best and worst performance is:

$$(m-r)\sigma_t^2 + (m-r)\sigma_l^2 \frac{\sum_{i=1}^{r} f_i^2}{\kappa_1^2} \le \left\| \mathbf{p} - \mathbf{U}'^r \mathbf{U}'^{r\,T} \mathbf{p} \right\|_F^2 \le (m-r)\sigma_t^2 + (m-r)\sigma_l^2 \frac{\sum_{i=1}^{r} f_i^2}{\kappa_r^2} \quad (40)$$

$$(m-r)\sigma_t^2 + (m-r)\sigma_l^2 \frac{m\sigma_s^2}{\kappa_1^2} \le \left\| \mathbf{p} - \mathbf{U}'^r \mathbf{U}'^{r\,T} \mathbf{p} \right\|_F^2 \le (m-r)\sigma_t^2 + (m-r)\sigma_l^2 \frac{m\sigma_s^2}{\kappa_r^2} \quad (41)$$

where $\kappa_r^2 \le \frac{mn\sigma_s^2}{r} \le \kappa_1^2$. Define, furthermore, $\kappa_i^2 = c_i mn\sigma_s^2$:

$$(m-r)\sigma_t^2 + \frac{m-r}{nc_1}\sigma_l^2 \le \left\| \mathbf{p} - \mathbf{U}'^r \mathbf{U}'^{r\,T} \mathbf{p} \right\|_F^2 \le (m-r)\sigma_t^2 + \frac{m-r}{nc_r}\sigma_l^2 \quad (42)$$

5.3 Performance analysis over the learning samples

We have given the best and the worst performance analysis of the recognition system. Next, we want to analyze the average performance of the system when we test the basis on the whole learning examples, i.e. all the images that are used to obtain the basis images.

$$\underset{\mathbf{q}\in\{\mathbf{A}_i | i=1,\cdots,n\}}{E} \left\| \mathbf{p} - \mathbf{U}'^r \mathbf{U}'^{r\,T} \mathbf{p} \right\|_F^2 = (m-r)\sigma_t^2 + (m-r)\sigma_l^2 \sum_{i=1}^{r} \frac{\underset{\mathbf{q}\in\{\mathbf{A}_i | i=1,\cdots,n\}}{E} f_i^2}{\kappa_i^2} \quad (43)$$

From (5),

$$\mathbf{A} = [\mathbf{A}_1, \cdots, \mathbf{A}_n] = [\mathbf{U}_1, \cdots, \mathbf{U}_r] diag(\kappa_1, \cdots, \kappa_r)[\mathbf{V}_1, \cdots, \mathbf{V}_r]^T = [\mathbf{U}_1, \cdots, \mathbf{U}_r][\kappa_1 \mathbf{V}_1, \cdots, \kappa_r \mathbf{V}_r]^T$$

Surprisingly,

$$\sum_{\mathbf{q} \in \{\mathbf{A}_i | i=1,\cdots,n\}} f_i^2 = \left\| \kappa_i \mathbf{V}_i \right\|_F^2 = \kappa_i^2 \text{ and } \underset{\mathbf{q} \in \{\mathbf{A}_i | i=1,\cdots,n\}}{E} f_i^2 = \frac{\kappa_i^2}{n} \tag{44}$$

Then

$$\underset{\mathbf{q} \in \{\mathbf{A}_i | i=1,\cdots,n\}}{E} \left\| \mathbf{p} - \mathbf{U}'^r \mathbf{U}'^{r^T} \mathbf{p} \right\|_F^2 \cong (m-r)\sigma_t^2 + {(m-r) r \sigma_l^2} \Big/ {n} \tag{45}$$

It can be easily proved that $(m-r)\sigma_t^2 + {(m-r) r \sigma_l^2} \Big/ {n}$ is the expectation for any test sets when the r largest singular values of the learning matrix $\mathbf{A}$ are equivalent. Moreover, from (48), this is also the best expectation performance over a *random* sample set, where the *randomness* means that Ef_i^2 in (38) should be statistically equivalent.

From this formula, (45), we can see clearly the effects of all the parameters in the recognition system. Given that the noise in the learning samples and in the test image, compared with the signal, is small, the performance can be regarded to be independent of the signal level. As m approaches a very large number, compared with r, the SSD is almost linearly dependent on m. As the number of the learning samples, n, increases, the recognition system improves: the error from the basis images decreases, and as n approaches infinite, the error from the basis images approaches zero. However, the error from the test image can't be reduced except by having a cleaner image.

Another measure, used in the recognition system, is the angle between the test image and the basis images:

$$\left\|\mathbf{p}-\mathbf{U}'^{r}\mathbf{U}'^{rT}\mathbf{p}\right\|_F^2 \Big/ \|\mathbf{p}\|_F^2 = \frac{(m-r)\sigma_t^2+(m-r)r\sigma_l^2/n}{m(\sigma_s^2+\sigma_t^2)} \underset{m\to\infty}{\rightarrow} \frac{\sigma_t^2+r\sigma_l^2/n}{\sigma_s^2+\sigma_t^2} = \frac{\sigma_t^2}{\sigma_s^2+\sigma_t^2}+\frac{r\sigma_l^2}{n(\sigma_s^2+\sigma_t^2)} \quad (46)$$

Supposing $m>>r$, the angle is independent of the size of the object, and depends on the energy level of the signal and the noise (in the learning samples and in the test image). As the size of the learning samples, n, increases, the system improves: the error from the basis images approaches zero and the error from the test image gradually dominates in the total error.

5.4 The optimal learning set

Suppose that the expectation of the test images, i.e. $\{f_i^2\}$, in (38), is known. How should we design the recognition system: specifically, how to select the learning samples, so that the system, concerning the expectation, has the best performance? Obviously, only the second term in (38) is dependent on the learning samples. The problem is:

$$\min \quad \sum \frac{f_i^2}{\kappa_i^2}, \text{ subject to } \sum \kappa_i^2 = C \quad (47)$$

$\sum \kappa_i^2 = C$ means that, when the dimension, m, and the size, n, of the learning samples is large enough, the signal energy, $\sum \kappa_i^2$, should be approximately $mn\sigma_s^2$. By using a Lagrange multiplier, the minimum can be obtained *iff*

$$\frac{f_i}{\kappa_i^2} \equiv Cons \quad (48)$$

From (48), we can draw such a conclusion, however it is a little surprising that the basis images, obtained from the n samples of $\mathbf{A}$ are not optimal when the test image set is also $\{\mathbf{A}_i\}$. The reason is that, the basis, corresponding to the largest singular value, is

overlearned in the *learning* process: from (48), the optimal *learning ability*, κ_i^2, should be proportional to f_i, while κ_i^2 is actually proportional to f_i^2, as in (44).

6. Simulation results

Here, we have to note that it is very difficult to have real data with high precision ground truth. Thus, in this section, we present some *simulations* to verify result 1 and result 2*.

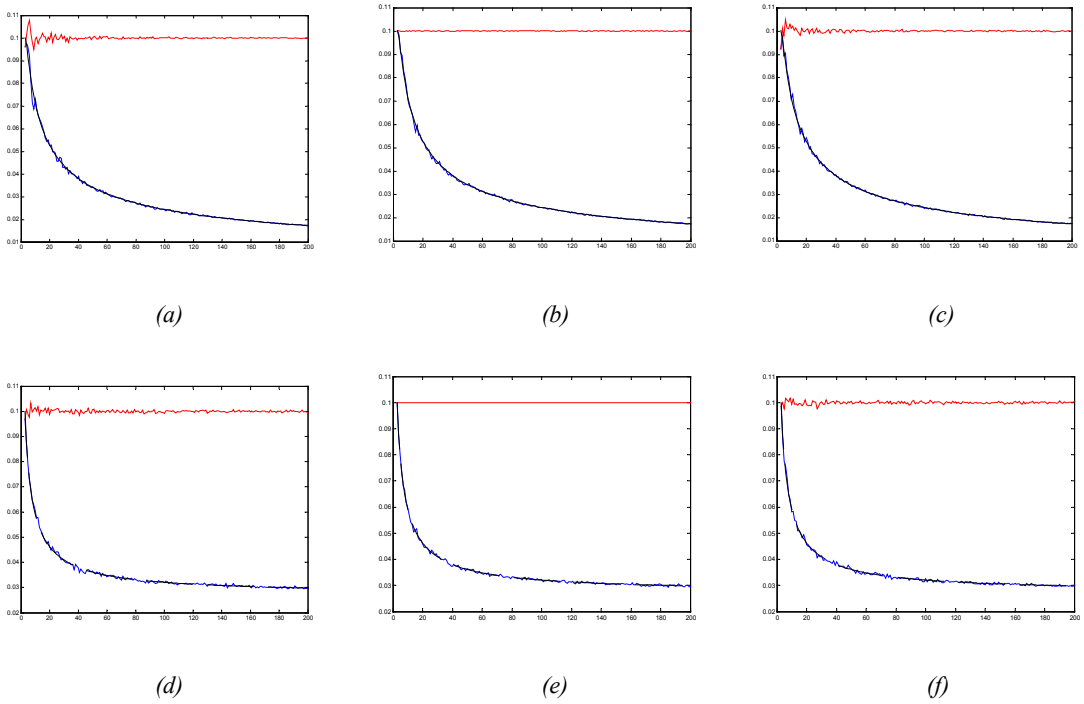
6.1 Simulation of the denoising capacity of SVD

In a recent paper, an experimental result related the SVD's *denoising* performance has been reported [5]. In that example, noise with amplitude of 1.5/40=0.037 still resides in the approximation matrix: where the noise-free 40×40 matrix, with a rank of 3, had been corrupted with zero-mean-and-0.01-variance Gaussian noise. From result 1 we have derived, the value should be 0.038. That this is pretty close to the result in [5], confirming the theory present here.

To provide further evidence, we have carried out our own simulations. Here, we work on a set of rank-3 matrices. For square matrices, the size of the matrices increases from 3 to 200; while for rectangular matrices, the number of the columns remains unchanged, staying at 40. The noise level is 0.1. In Matlab notation, $\mathbf{M} = randn(rows,3) * randn(3, columns) + 0.1 * randn(rows, columns)$ is the noise-corrupted matrix. Fig. 1 shows the simulation results of SVD's denoising performance, compared with the expectation from result 1. It can be easily observed that the expected

* The Matlab code for the simulation in this section is available in [12].

curve almost coincides with the simulation result. In contrast with fig. 1(d-f) (rectangular matrices), the curves in fig. 1(a-c) (square matrices) can be observed to continue towards zero error, while the error for the rectangular matrices changes little after the number of the rows increases to 20 or 40.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

Fig. 1: The average error that still resides in the approximation matrix. The abscissa denotes the number of the rows of the matrices, and the error is on the ordinate. (a-c) are for the square matrices, and (d-f) are for the rectangular matrices, who have a constant, 40, columns. There are three curves in every sub-figure: the (approximately) straight curve in the upper part denotes the original noise in the noise corrupted matrix, and the smooth/unsmooth curves are the expectation/actual error in the approximation matrix respectively. In (a) and (d), the signal and the noise are randomly generated. In (b) and (e), the noise levels are normalized, so that the average energy in each entry of the matrices is 0.01. In (c) and (f), the signal matrices have 3 equivalent singular values, while the energy level remains same.

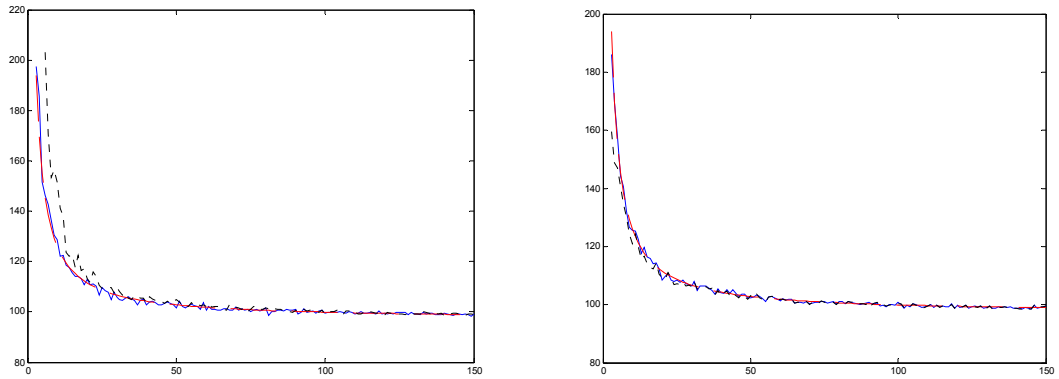
6.2 Simulation of the *learning* capacity for LSA-based recognition

In this section, we present some simulation results concerning the SSD performance of the LSA-based recognition system, as stated in section 5. Suppose we work on a set of rank 3 subspaces but in a dimension of 100. In this section, the

parameters are set as follows: *m=100*, *r=3*, $\sigma_s = 100$, and $\sigma_l = \sigma_t = 1$. First, the SSD performance of a set of basis images is analyzed, over two test sets: the learning set, from which the basis images are obtained, and another random set where its 3 singular values have been artificially equalized. Obviously, as the learning sample size approaches infinite, the SSD, over two sets, approaches a stable value, as shown in fig. 2-a. Over the learning set, the performance, denoted by solid curve, almost coincides with the expectation, denoted by dashed curve. Over the random set, the performance is denoted by dotted curve. Because the 3 singular values of the random test set have been artificially equalized, the best performance over this random set can be obtained only if the learning set has 3 equal singular values, from (48). However, the random learning set always has 3 distinct singular values. Thus, the performance over the random test set is worse than the optimal curve, denoted by dashed curve, especially for the small-size learning samples; in fact, the performance for the recognition system is very bad, at 5,771.6, 788.1 and 588.1 respectively, when the learning sample sizes are only 3, 4 and 5; in order to make the curves clear, these points have been omitted in fig. 2-a.

Conversely, next, we first have a random test set, and show the performance of different learning sets (different basis images): an optimal learning set, which complies with (48), and a random learning set, who has 3 equal singular values. For the random learning set, with 3 equal singular values, its performance, denoted by the solid curve, can be expected to coincide with the expectation (45), denoted by the dashed curve, as shown in fig. 2-b. Obviously, the optimal learning set, complying with (48), has a better performance than the random learning set, especially for small learning sizes. Note, if the learning set is *truly* randomly generated, it probably has a very bad SSD performance,

especially for a small-size learning set. For example, the r^{th} basis image may be very *dirty*, because the r^{th} singular value of the learning set is comparatively small; while most of the energy of the test image comes from this basis image. In such cases, the error from the basis images, especially from the r^{th} basis image, will dominate the total error, as can be seen from (38).



(a) *(b)*

*Fig. 2, The dependency of SSD on the size of learning samples.(a) for a learning set, over two test sets: **(solid)** the learning set from which the basis images are obtained, and **(dotted)** another random set that has 3 equal singular values; (b) for a test set, by two learning sets: **(dotted)** the optimal learning set and **(solid)** another random learning set that has 3 equal singular values. In both subfigures, the dashed curves denote the expectation from (45).*

In fig. 3, we show the effects of the three parameters in (45), the size of the learning samples, *n*, the noise level in the learning set, σ_l, and the noise level in the test set, σ_t; on SSD when the recognition system works over the *learning* samples. Fig. 3 (a) shows the performance of SSD when the noise level in *test* image is 0.5 (very small). It can be easily observed: the *square* dependency on the noise level in the learning set and the decreased effects of the noise in the learning set as the learning size increases. Fig. 3 (b) shows the performance of SSD when the noise level in *learning* samples is 0.5 (very small). It can be easily observed: the *square* dependency on the noise level in the test set

and its effect is almost independent of the learning size. Fig 3 (c) and (d) show the effect of the noise levels of the learning set and the test set when the learning sizes are 3 and 125 respectively. When the learning size is 3, the noise in learning set has almost a same effect on SSD as the noise in test set; when the learning size is 125 (>>3), the noise in learning set can be almost neglected if the level is not much higher than that in the test set.

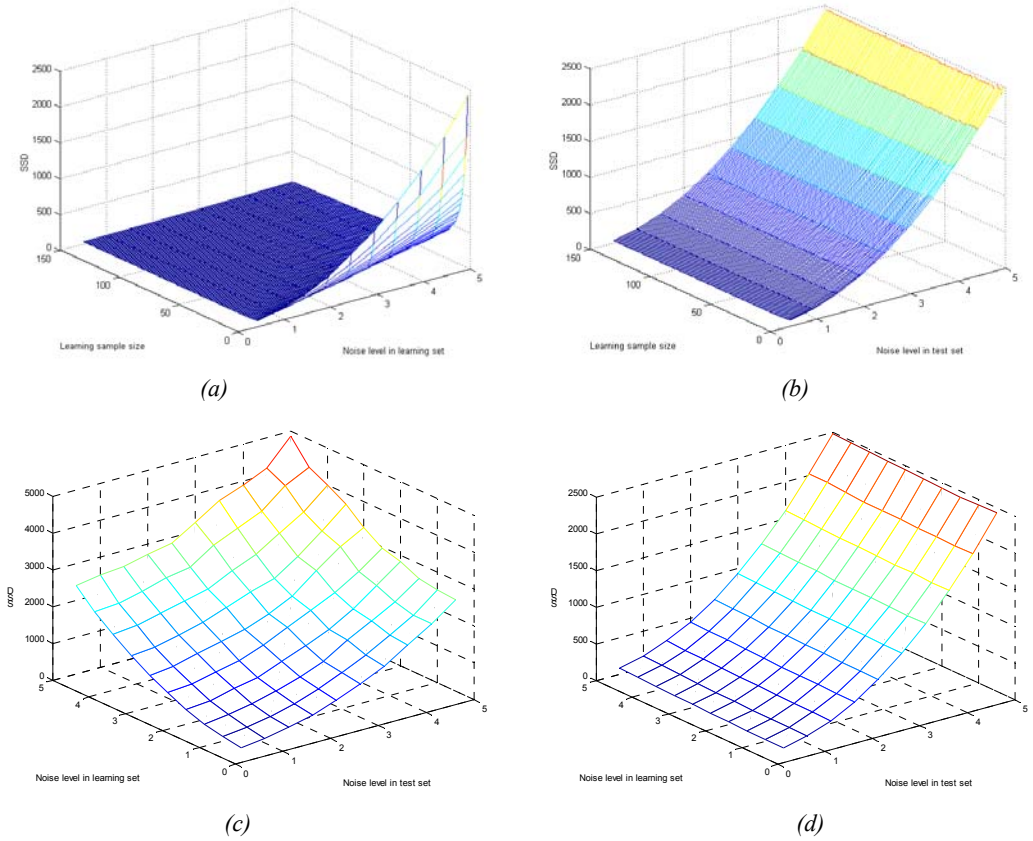


Fig. 3: The effects of three parameters in (43) on SSD. To see the description in the text.

6.3 Relationship with some experimental observations

Here, we can explain such phenomena previously reported in the computer vision literature, by using the analysis in section 4 and section 5. For example, in SFM, the root-mean-square error of the recovered shape with respect to the true shape was reported in

[18]. Fig. 6 in that paper [18] shows that the error approaches a constant value after the number of the frames increases to 20 or 40, as almost coincides with the result 1, in section 2.2 or the fig. 1 in section 6.1.

Another two observations are related to result 2, in section 2.2. In [2], it was reported that no significant deterioration of the performance was found for LSA-based face recognition, if the images were subsampled by 16×16 squares, which means that, m, the number of the rows of **A**, decreases by 1/256. However, the reduced m is still very large, about 1000 (>>4 or 9). We can find the explanation from (46): the performance, measured by the angel between the test image and the basis images, is almost independent of m if $m>>r$.

The last, but not the least, (maybe even the most important), observation was that "recognition of an object under a particular lighting and pose can be performed reliably provided the object has been previously seen under similar circumstances" [8]. A very reliable explanation can be found from (38) and (48). For a test image, if it or its similar cases have been observed in the learning samples, its $\{f_i^2\}$ will probably have a *good* relationship with $\{\kappa_i^2\}$, i.e., for a larger κ_i^2, f_i^2 is also larger, and vice versa. More formally, if (48) holds, the recognition system has a best performance. However, for a test image, which is produced under very different lighting conditions from those in the learning set, its $\{f_i^2\}$ probably has very *bad* relationship with $\{\kappa_i^2\}$. If most of its energy comes from the *dirtiest* basis, which corresponds to the r^{th} singular value of the learning matrix, from (38), the recognition error is probably very large. This not only explains the drawback of PCA-based face recognition, pointed out in [8], but also gives a possible

solution, as suggested by (48). For a *random* test set, the best learning samples should be selected this way: to equalize the first *r* largest singular values as possible. However, we do not present any specific strategies for this *open*, and probably promising, issue.

6. Conclusion

The main contribution of this paper is the presentation of a theoretical analysis of SVD-based low rank projections: specifically the *denoising* capacity of SVD (where we characterized the error that still resides in the SVD-*denoised* matrix) and the *learning* capacity of LSA-based recognition systems (where we showed that the projection error can be decomposed into two independent sources, one from the test image and the other from the basis image). Another contribution of this paper is to fill an apparent gap in the literature: the perturbation theory concerning multiple eigenvalues (singular values).

Appendix A

In this appendix, we present the proofs of theorem 1-4. Here, we do not follow the notation in [34], where an arbitrarily small positive number, ε, was introduced. Because we only consider the first-order perturbation, a simpler and straightforward form is used. Suppose **M** has a simple eigenvalue λ, and the corresponding eigenvector is **x**. If **M** is corrupted with $\Delta\mathbf{M}$ and $\Delta\mathbf{M}$ is small enough, the first-order perturbations of the eigenvalue and the eigenvector, denoted as $\Delta\lambda$ and $\Delta\mathbf{x}$ respectively, will be small enough, from Ostrowski's continuity theorem [34]. Suppose their higher-order terms are $\delta\lambda$ and $\delta\mathbf{x}$, respectively. From $(\mathbf{M}+\Delta\mathbf{M})(\mathbf{x}+\Delta\mathbf{x}+\delta\mathbf{x}) = (\lambda+\Delta\lambda+\delta\lambda)(\mathbf{x}+\Delta\mathbf{x}+\delta\mathbf{x})$, we have the first-order perturbation, by dropping the higher-order terms:

$$\mathbf{M}\cdot\mathbf{x} + \mathbf{M}\cdot\Delta\mathbf{x} + \Delta\mathbf{M}\cdot\mathbf{x} \approx \lambda\cdot\mathbf{x} + \lambda\cdot\Delta\mathbf{x} + \Delta\lambda\cdot\mathbf{x} \qquad \text{(A.1)}$$

Of course, this first-order perturbation is same as that in [34], despite the difference in notation.

A.1 Proof of theorem 1

Proof: Suppose $\mathbf{x}_i' = \mathbf{x}_i + \sum_{j \neq i} c_{j,i}\mathbf{x}_j$ and $\lambda_i' = \lambda_i + b_i$. From the first-order perturbation, we have $\mathbf{M}\mathbf{x}_i + \mathbf{M}\sum_{j \neq i} c_{j,i}\mathbf{x}_j + \mathbf{N}\mathbf{x}_i \approx \lambda_i \mathbf{x}_i + \lambda_i \sum_{j \neq i} c_{j,i}\mathbf{x}_j + b_i \mathbf{x}_i$, and

$$\sum_{j \neq i} c_{j,i}(\lambda_j - \lambda_i)\mathbf{x}_j + \mathbf{N}\mathbf{x}_i = b_i \mathbf{x}_i \tag{A.2}$$

Because $\mathbf{M}$ is symmetric and has m distinct eigenvalues, $\{\mathbf{x}_i\}$ are orthogonal to each other. Pre-multiplying (A.2) by $\mathbf{x}_i^T$, we obtain $b_i = \mathbf{x}_i^T \mathbf{N}\mathbf{x}_i = \beta_{i,i}$. Pre-multiplying $\mathbf{x}_j^T$, we have $c_{j,i} = \dfrac{\beta_{j,i}}{\lambda_i - \lambda_j}$.

A.2 Proof of theorem 2

Proof: Suppose $\mathbf{\Omega} = \mathbf{\Sigma} + \mathbf{C}$. Obviously, $\{\kappa_j\}$ and $\{\mathbf{e}_j\}$ are respectively the singular values and the right/left singular vectors of $\mathbf{\Sigma}$. First, about Ω, we prove the first order perturbations of the singular values, λ_i', the right singular vectors $\mathbf{x}_i'$, and the left singular vectors $\mathbf{y}_i'$ are respectively

$$\lambda_i' = \kappa_i + C_{i,i} \tag{A.3}$$

$$\mathbf{x}_i' = \mathbf{e}_i + \sum_{j \neq i} \frac{\kappa_j C_{j,i} + \kappa_i C_{i,j}}{\kappa_i^2 - \kappa_j^2}\mathbf{e}_j \tag{A.4}$$

$$\mathbf{y}_i' = \mathbf{e}_i + \sum_{j \neq i} \frac{\kappa_i C_{j,i} + \kappa_j C_{i,j}}{\kappa_i^2 - \kappa_j^2}\mathbf{e}_j \tag{A.5}$$

Suppose $\lambda_i' = \kappa_i + \Delta\kappa_i$, $\mathbf{x}_i' = \mathbf{e}_i + \sum_{j\neq i} f_{j,i}\mathbf{e}_j$, and $\mathbf{y}_i' = \mathbf{e}_i + \sum_{j\neq i} g_{j,i}\mathbf{e}_j$

According the property of SVD, we have $\Omega\mathbf{x}_i' = \lambda_i'\mathbf{y}_i'$ and $\Omega^T\mathbf{y}_i' = \lambda_i'\mathbf{x}_i'$. Equating their first order, we have:

$$\Sigma\mathbf{e}_i + \mathbf{C}\mathbf{e}_i + \Sigma\sum_{j\neq i} f_{j,i}\mathbf{e}_j \approx \kappa_i\mathbf{e}_i + \Delta\kappa_i\mathbf{e}_i + \kappa_i\sum_{j\neq i} g_{j,i}\mathbf{e}_j \tag{A.6}$$

$$\Sigma^T\mathbf{e}_i + \mathbf{C}^T\mathbf{e}_i + \Sigma^T\sum_{j\neq i} g_{j,i}\mathbf{e}_j \approx \kappa_i\mathbf{e}_i + \Delta\kappa_i\mathbf{e}_i + \kappa_i\sum_{j\neq i} f_{j,i}\mathbf{e}_j \tag{A.7}$$

Then

$$\mathbf{C}\mathbf{e}_i + \sum_{j\neq i}\kappa_j f_{j,i}\mathbf{e}_j = \Delta\kappa_i\mathbf{e}_i + \kappa_i\sum_{j\neq i} g_{j,i}\mathbf{e}_j \tag{A.8}$$

$$\mathbf{C}^T\mathbf{e}_i + \sum_{j\neq i}\kappa_j g_{j,i}\mathbf{e}_j = \Delta\kappa_i\mathbf{e}_i + \kappa_i\sum_{j\neq i} f_{j,i}\mathbf{e}_j \tag{A.9}$$

First, by equating $\mathbf{e}_i$, we have $\Delta\kappa_i = C_{i,i}$. And from $\mathbf{e}_j$ $(j \neq i)$,

$$\begin{cases} \kappa_i g_{j,i} - \kappa_j f_{j,i} = C_{j,i} \\ -\kappa_j g_{j,i} + \kappa_i f_{j,i} = C_{i,j} \end{cases} \tag{A.10}$$

$$\begin{cases} g_{j,i} = (\kappa_i C_{j,i} + \kappa_j C_{i,j})/(\kappa_i^2 - \kappa_j^2) \\ f_{j,i} = (\kappa_j C_{j,i} + \kappa_i C_{i,j})/(\kappa_i^2 - \kappa_j^2) \end{cases} \tag{A.11}$$

So far, (A.3-A.5) have been proved. From

$$\mathbf{B} = \mathbf{U}\Omega\mathbf{V}^T \approx \mathbf{U}[\mathbf{y}_1', \cdots, \mathbf{y}_m']diag\{\lambda_1', \cdots, \lambda_m'\}[\mathbf{x}_1', \cdots \mathbf{x}_m']^T\mathbf{V}^T \tag{A.12}$$

$\mathbf{B}$ has λ_i', $\mathbf{V}\mathbf{x}_i'$, and $\mathbf{U}\mathbf{y}_i'$ respectively as its singular values, right and left singular vectors.

A.3 Proof of theorem 3

Proof: From the perturbation expansion about the eigenvectors corresponding to a multiple eigenvalue [34], we can suppose $\mathbf{x}_i' = \sum_{j=1}^{k} c_{j,i}\mathbf{x}_j + \sum_{j=k+1}^{m} f_{j,i}\mathbf{x}_j$ and $\lambda_i' = \lambda + \Delta\lambda_i$.

Note: $c_{j,i}$ are different from $f_{j,i}$. $c_{j,i}$ can possibly take any value within [0,1], while $f_{j,i}$ approach zeroes if **N** is small enough.

$$(\mathbf{M}+\mathbf{N})\mathbf{x}'_i = \lambda'_i\mathbf{x}'_i \tag{A.13}$$

Equating the first order:

$$\mathbf{M}\sum_{j=1}^{k} c_{j,i}\mathbf{x}_j + \mathbf{M}\sum_{j=k+1}^{m} f_{j,i}\mathbf{x}_j + \mathbf{N}\sum_{j=1}^{k} c_{j,i}\mathbf{x}_j \approx \lambda\sum_{j=1}^{k} c_{j,i}\mathbf{x}_j + \lambda\sum_{j=k+1}^{m} f_{j,i}\mathbf{x}_j + \Delta\lambda_i\sum_{j=1}^{k} c_{j,i}\mathbf{x}_j \tag{A.14}$$

Then

$$\sum_{j=k+1}^{m}\lambda_j f_{j,i}\mathbf{x}_j + [\mathbf{x}_1,\cdots,\mathbf{x}_m]\mathbf{Q}_{1:m,1:k}\mathbf{c}_i \approx \lambda\sum_{j=k+1}^{m} f_{j,i}\mathbf{x}_j + \Delta\lambda_i\sum_{j=1}^{k} c_{j,i}\mathbf{x}_j \tag{A.15}$$

where $\mathbf{c}_i = [c_{1,i}, c_{2,i}, \cdots, c_{k,i}]^T$. Equating the coefficients of $\mathbf{x}_j$ for $(j = 1,\cdots,k)$, we have

$$\mathbf{Q}_{1:k,1:k}\mathbf{c}_i = \Delta\lambda_i\mathbf{c}_i \tag{A.16}$$

where $\mathbf{Q}_{1:k,1:k}$ is the left-up $k \times k$ submatrix of **Q**. If $\mathbf{Q}_{1:k,1:k}$ has *k* distinct eigenvalues, the solution of $\Delta\lambda_i$ and $\mathbf{c}_i$ is unique, as (A.16). Obviously, **c** is same as **S**, as defined in the theorem. After substituting $\Delta\lambda_i$ and $\mathbf{c}_i$ in (A.14), the equality of $\mathbf{x}_j$ for $(j = k+1,\cdots,m)$ produces the first order perturbations of $f_{j,i}$ as in the theorem.

A.4 Proof of theorem 4

Proof: Let **Ω** has $\{\lambda''_i\}$, $\{\mathbf{x}''_i\}$ and $\{\mathbf{y}''_i\}$ as its first *k* singular values, right singular vectors and left singular vectors respectively. For $i > k$, $\{\lambda''_i\}$, $\{\mathbf{x}''_i\}$ and $\{\mathbf{y}''_i\}$ can be obtained as in theorem 2. Thus, we concentrate on the first-order perturbation of $\{\lambda''_i\}$, $\{\mathbf{x}''_i\}$ and $\{\mathbf{y}''_i\}$, for $i \le k$.

First, we only consider one singular value and the corresponding singular vector. Combining the techniques in the proof of theorem 2 and theorem 3, we assume that the first-order perturbations of the right and the left singular vectors, $\mathbf{x}''$ and $\mathbf{y}''$ respectively, have the following forms:

$$\mathbf{x}'' = \sum_{i=1}^{k} p_i \mathbf{e}_i + \sum_{i=k+1}^{m} q_i \mathbf{e}_i \tag{A.17}$$

$$\mathbf{y}'' = \sum_{i=1}^{k} f_i \mathbf{e}_i + \sum_{i=k+1}^{m} g_i \mathbf{e}_i \tag{A.18}$$

Note: p_i and f_i can possibly take any value within [0,1], while q_i and g_i approach zeroes if **N** is small enough. Because the singular values of the matrix, **M**, are the square roots of the eigenvalues of $\mathbf{M}\mathbf{M}^T$. From the continuity of the eigenvalues of $\mathbf{M}\mathbf{M}^T$, the singular values of $\mathbf{M}$ also obey Ostrowski's continuity rule. Supposing the corresponding singular value is $\lambda'' = \kappa + \Delta\kappa$, equality of the first order of $\mathbf{\Omega}\mathbf{x}'' = \lambda''\mathbf{y}''$ and $\mathbf{\Omega}^T\mathbf{y}'' = \lambda''\mathbf{x}''$ produces:

$$\kappa \sum_{i=1}^{k} p_i \mathbf{e}_i + \sum_{i=k+1}^{m} q_i \kappa_i \mathbf{e}_i + \sum_{i=1}^{k} p_i \mathbf{C}_i = (\kappa + \Delta\kappa) \sum_{i=1}^{k} f_i \mathbf{e}_i + \kappa \sum_{i=k+1}^{m} g_i \mathbf{e}_i \tag{A.19}$$

$$\kappa \sum_{i=1}^{k} f_i \mathbf{e}_i + \sum_{i=k+1}^{m} g_i \kappa_i \mathbf{e}_i + \sum_{i=1}^{k} f_i (\mathbf{C}^T)_i = (\kappa + \Delta\kappa) \sum_{i=1}^{k} p_i \mathbf{e}_i + \kappa \sum_{i=k+1}^{m} q_i \mathbf{e}_i \tag{A.20}$$

From (A.19) and (A.20), we have, by equating $\mathbf{e}_s$ (for $s = 1, \cdots, k$):

$$\kappa p_s + \sum_{i=1}^{k} p_i C_{s,i} = (\kappa + \Delta\kappa) f_s \tag{A.21}$$

$$\kappa f_s + \sum_{i=1}^{k} f_i C_{i,s} = (\kappa + \Delta\kappa) p_s \tag{A.22}$$

In matrix form, they are:

$$(\mathbf{C}_{k\times k} + \kappa\mathbf{I})\mathbf{p} = (\kappa + \Delta\kappa)\mathbf{f} \tag{A.23}$$

$$(\mathbf{C}^{T}_{k\times k} + \kappa\mathbf{I})\mathbf{f} = (\kappa + \Delta\kappa)\mathbf{p} \tag{A.24}$$

where $\mathbf{C}_{k\times k}$ is the left-up k *by* k submatrix of $\mathbf{C}$, $\mathbf{p} = [p_1, p_2, \cdots, p_k]^T$ and $\mathbf{f} = [f_1, f_2, \cdots, f_k]^T$. Obviously, $\kappa + \Delta\kappa$, $\mathbf{p}$ and $\mathbf{f}$ are respectively the singular value, the right and the left singular vectors of $\mathbf{C}_{k\times k} + \kappa\mathbf{I}$; and $\mathbf{p}$ and $\mathbf{f}$ correspond to the columns of $\mathbf{E}$ and $\mathbf{F}$ in the theorem. $\mathbf{C}_{k\times k} + \kappa\mathbf{I}$ just has k singular values, right and left singular vectors, which correspond to $\{\lambda_i''\}$, $\{\mathbf{x}_i''\}$ and $\{\mathbf{y}_i''\}$, for $i \le k$, of $\mathbf{\Omega}$.

Equating the $\mathbf{e}_t$ in (A.19) and (A.20), for $t > k$, we have

$$\sum_{l=1}^{k} C_{t,l} p_l = \kappa g_t - \kappa_t q_t \tag{A.25}$$

$$\sum_{l=1}^{k} C_{l,t} f_l = \kappa q_t - \kappa_t g_t \tag{A.26}$$

Note for $t>k$ and $l \le k$, $C_{t,l} = \Omega_{t,l}$ and $C_{l,t} = \Omega_{l,t}$. And, suppose we work on the i^{th} (for $1 \le i \le k$) singular value:

$$\kappa g_t - \kappa_t q_t = \sum_{j=1}^{k} \Omega_{t,j} p_{j,i} = \Omega'_{t,i} \tag{A.27}$$

$$\kappa q_t - \kappa_t g_t = \sum_{j=1}^{k} \Omega_{j,t} f_{j,i} = \Omega'_{i,t} \tag{A.28}$$

Combining (A.23-24) and (A.27-28),

$$\mathbf{x}'' = \sum_{i=1}^{k} p_i \mathbf{e}_i + \sum_{j=k+1}^{m} \frac{\kappa_j \Omega'_{j,i} + \kappa \Omega'_{i,j}}{\kappa_i^2 - \kappa_j^2} \mathbf{e}_j \tag{A.29}$$

$$\mathbf{y}'' = \sum_{i=1}^{k} f_i \mathbf{e}_i + \sum_{j=k+1}^{m} \frac{\kappa \Omega'_{j,i} + \kappa_j \Omega'_{i,j}}{\kappa_i^2 - \kappa_j^2} \mathbf{e}_j \qquad \text{(A.30)}$$

are respectively the right and left singular vectors of $\mathbf{\Omega}$. After the system transformation as (15), the first *k* left/right singular vectors of $\mathbf{\Omega}'$ are as defined in (17, 18).

References:

1. R. Basri, and D. W. Jacobs,Lambertian Reflectance and Linear Subspaces, *ICCV01*, pp. 383-390, 1999.
2. R. Basri, and D. W. Jacobs,Lambertian Reflectance and Linear Subspaces, *IEEE PAMI*, vol. 25, no. 2, pp. 218-233, 2003.
3. P.N. Belhumeur, J. Hespanha, and D. Kriegman Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE PAMI*, vol. 19, no. 7, pp. 711-720, 1997.
4. P.N. Belhumeur, and D. Kriegman, What is the Set of Images of an Object under all Possible Illumination Conditions? *IJCV*, 28(3), 245-260, 1998.
5. P. Chen and D. Suter, Recovering the missing components in a large noisy low-rank matrix: application to SFM, submitted to *IEEE PAMI*.
6. R. Eipstein, P. Hallinan, and A. Yuille, 5±2 eigenimages suffices: An empirical investigation of low-dimensional lighting models, *Proc. IEEE Workshop Physics-Based vision*, pp. 108-116, 1995.
7. A. Georghiades, D. Kriegman, and P.N. Belhumeur, Illumination cones for recognition under variable lighting: faces, *CVPR98*, pp. 52-58, 1998.
8. A. Georghiades, P.N. Belhumeur, and D. Kriegman, From Few to Many: Generative Models of Object Recognition, *IEEE PAMI*, vol. 23, no. 6, pp. 643-660, 2001.

9. G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
10. P. Hallinan, A low-dimensional representation of human faces for arbitrary lighting conditions, *CVPR94*, pp. 995-999, 1994.
11. R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2000.
12. http://www.ds.eng.monash.edu.au/dpl/pei/MatlabCodeForLSA.zip
13. M. Irani, Multi-frame optical flow estimation using subspace constraints, *ICCV99*(I), pp. 626-633, 1999.
14. M. Irani, Multi-frame correspondence estimation using subspace constraints, *IJCV*, vol. 48 (3), pp. 173-194, 2002.
15. D. Jacobs, P.N. Belhumeur and R. Basri, Comparing Images Under Variable Illumination," *CVPR98*, pp. 610-617, 1998.
16. F. Kahl and A. Heyden, Affine structure and motion from points, lines and conics, *IJCV*, vol. 33(3), pp. 163-180, 1999.
17. K. Kanatani, Motion segmentation by subspace separation and model selection, *ICCV01*(II), pp. 301-306, 2001.
18. T. Morita and T. Kanade, A sequential factorization method for recovering shape and motion from image streams, *IEEE PAMI*, vol. 19, no. 8, pp. 858-867, 1997.
19. Y. Moses, Y. Adini, and S. Ullman, Face recognition: The problem of compensating for changes in illumination direction, *ECCV94*, pp.286-296, 1994.
20. H. Murase and S. K. Nayar, Illumination planning for object recognition using parametric eigenspaces, *IEEE PAMI*, vol. 16, no. 12, pp. 1219-1227, 1994.

21. H. Murase and S. K. Nayar, Visaul learning and recognition of 3-D objects from appearance, *IJCV*, vol. 14, pp. 5-24, 1995.
22. S. A. Nene, S. K. Nayar and H. Murase, Software library for appearance matching (SLAM), *ARPA Image understanding workshop*, Montery, Nov., 1994.
23. C. Poelman and T. Kanade, A paraperspective factorization method for shape and motion recovery, *IEEE PAMI*, vol. 19(3), pp. 206–219, 1997.
24. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C (2nd edition)*, Cambridge University Press, 1992.
25. R. Ramamoorthi, Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertial Object, *IEEE PAMI*, vol. 24, no. 10, pp. 1322-1333, 2002.
26. R. Ramamoorthi and P. Hanrahan, On the relationship between radiance and irrandiance: Determining the illumination from images of a convex Lambertian object, *Journal of the Optical Society of America* (*JOSA A*), vol. 18, no. 10, pp. 2448-2459, 2001.
27. I. D. Reid and D. W. Murray, Active tracking of foveated feature clusters using affine structure, *IJCV*, vol. 18, no. 1, pp. 41-60, 1996.
28. *A.* Shashua, On photometric issues in 3D visual recognition from a single 2D image, *IJCV*, vol. 21, no. 1/2, pp. 99-122, 1997.
29. *A.* Shashua and S. Avidan, The Rank 4 Constraint in Multiple (>2) View Geometry, *ECCV96*, pp. 196-206, 1996.
30. G. W. Stewart and J. G. Sun, *Matrix perturbation theory*, Acedemic press, 1990.

31. J. I. Thomas and J. Oliensis, Dealing with noise in multiframe structure from motion, *CVIU*, vol. 76, no. 2, pp. 109-124, 1999.
32. C. Tomasi and T. Kanade, Shape and motion from image streams under orthography: A factorization method, *IJCV*, vol. 9(2), pp. 137–154, 1992.
33. M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cognitive Neuroscience*, vol. 3, no. 1, pp.71-96, 1991.
34. J. H. Wilkinson, *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965.
35. A. L. Yuille, D. Snow, R. Epstein, 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), pp. 203—22, 1999.
36. L. Zelnik-Manor M. Irani, Multi-View Subspace Constraints on Homographies, *ICCV99*, pp.710-715, 1999.
37. L. Zelnik-Manor M. Irani, Multi-View Subspace Constraints on Homographies, *IEEE PAMI*, vol. 24, no. 2, pp. 214-223, 2002.
38. L. Zhao and Y. H. Yang, Theoretical analysis of illumination in PCA-based vision systems, *Pattern recognition*, vol. 32, pp. 547-564, 1999.

***IJCV**: **I**nternational **J**ournal of **C**omputer **V**ision*
***IEEE PAMI**: **IEEE** Transactions on **P**attern **A**nalysis and **M**achine **I**ntelligence*
***ICCV**: **I**nternational **C**onference on **C**omputer **V**ision*
***CVPR**: IEEE Computer Society Conference on **C**omputer **V**ision and **P**attern **R**ecognition*
***ECCV**: **E**uropean **C**onference on **C**omputer **V**ision*