Limits of Learning-Based Superresolution Algorithms

Zhouchen Lin[1] Junfeng He[2] Xiaoou Tang[1] Chi-Keung Tang[2]

[1]Microsoft Research Asia [2]Hong Kong University of Science and Technology

Technical Report
MSR-TR-2007-92

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Abstract[1]

Learning-based superresolution (SR) are popular SR techniques that use application dependent priors to infer the missing details in low resolution images (LRIs). However, their performance still deteriorates quickly when the magnification factor is moderately large. This leads us to an important problem: "Do limits of learning-based SR algorithms exist?" In this paper, we attempt to shed some light on this problem when the SR algorithms are designed for general natural images (GNIs). We first define an expected risk for the SR algorithms that is based on the root mean squared error between the superresolved images and the ground truth images. Then utilizing the statistics of GNIs, we derive a closed form estimate of the lower bound of the expected risk. The lower bound can be computed by sampling real images. By computing the curve of the lower bound w.r.t. the magnification factor, we can estimate the limits of learning-based SR algorithms, at which the lower bound of expected risk exceeds a relatively large threshold. We also investigate the sufficient number of samples to guarantee an accurate estimation of the lower bound. From our experiments, we have a key observation that the limits may be independent of the size of either the LRIs or the high resolution images.

1 Introduction

Superresolution (SR) is a technique that produces an image or video with a resolution higher than those of any of the input images or frames. Roughly speaking, SR algorithms can be categorized into four classes [3, 11, 5]. Interpolation-based algorithms register low resolution images (LRIs) with the high resolution image (HRI), then apply nonuniform interpolation to produce an improved resolution image which is further deblurred. Frequency-based algorithms try to dealias the LRIs by utilizing the phase difference among the LRIs. Reconstruction-based algorithms rely on the relationship between the LRIs and the HRI and assume various kinds of priors on the HRI in order to regularize this ill-posed inverse problem. Recently, many learning-based SR algorithms have attracted much attention.

1.1 Learning-Based SR Algorithms

Learning-based SR algorithms are new SR techniques that may have started from the seminal papers by Freeman and Pasztor [6] and Baker and Kanade [1]. Compared to traditional

[1] A short version of this technical report was accepted by International Conference on Computer Vision 2007.

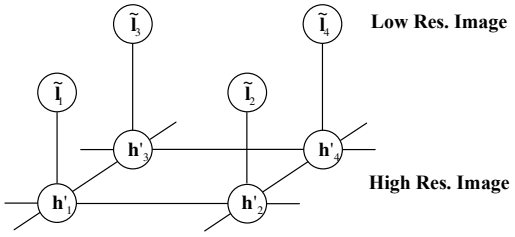

Figure 1: *The Markov network adopted by Freeman and Pasztor [6] (adapted from [6]).*

methods, which basically process images at the signal level, learning-based SR algorithms incorporate application dependent priors to infer the unknown HRI. For example, as a widely adopted framework, Freeman and Pasztor's Markov network [6] models the SR problem as an inference problem of the high frequency:

$$\mathbf{h}' = \arg\max_{\mathbf{h}'} P(\mathbf{h}'|\tilde{\mathbf{l}}) = \arg\max_{\mathbf{h}'} P(\tilde{\mathbf{l}}|\mathbf{h}')P(\mathbf{h}'),$$

where $\mathbf{h}'$ is the missing high frequency of the HRI $\mathbf{h}$, $\tilde{\mathbf{l}}$ is the mid-frequency of the input image $\mathbf{l}$ interpolated to the size of $\mathbf{h}$. Adding the inferred high frequency to the interpolated LRI gives the output HRI. Freeman and Pasztor defined the likelihood and the prior via image patches:

$$P(\tilde{\mathbf{l}}|\mathbf{h}') = \prod_k P(\tilde{\mathbf{l}}_k|\mathbf{h}'_k), \text{ and } P(\mathbf{h}') = \prod_{\mathbf{h}'_j \in \mathcal{N}(\mathbf{h}'_i)} P(\mathbf{h}'_i|\mathbf{h}'_j),$$

where $\mathbf{h}'_i$ and $\tilde{\mathbf{l}}_k$ are the patches in $\mathbf{h}'$ and $\tilde{\mathbf{l}}$, respectively, and $\mathcal{N}(\mathbf{h}'_i)$ is the set of neighboring high resolution patches of $\mathbf{h}'_i$. Figure 1 shows the Markov network that links the local patches. $P(\tilde{\mathbf{l}}_k|\mathbf{h}'_k)$ and $P(\mathbf{h}'_i|\mathbf{h}'_j)$ are learnt from training images and are approximated by a mixture of Gaussians (MoGs). The solution $\mathbf{h}'$ is found by belief propagation.

From the above example, one can see that the methodology of learning-based SR algorithms is quite different from traditional ones. Despite some drawbacks, such as the magnification factor is usually fixed and the performance often depends on how well the input LRI matches the training low resolution samples, learning-based SR algorithms have several advantages. For example, they work on fewer LRIs but can still achieve a higher magnification factor than traditional algorithms can. Most of them can even work on a *single* image. Moreover, it is possible to design fast learning-based SR algorithms, e.g., eigenface based face hallucination [4, 7], to achieve real-time SR. Finally, if we change the prior for learning-based SR algorithms, the HRIs may exhibit an artistic style [6, 12]. This may enable learning-based SR algorithms to perform style transfer. In contrast, traditional SR algorithms do not have such capability.

Because of their advantages, learning-based SR algorithms have become popular. Bishop et al. [2], Pickup et al. [12], and Sun et al. [15] also adopted the Markov network as Freeman and Pasztor [6] did but they differed in the definition of priors and likelihoods. Baker and Kanade's hallucination algorithm [1] further inspired the work in this field. Gunturk et al. [7], Capel and Zisserman [4], Liu et al. [9], and Wang and Tang [18] all used face bases and inferred the combination coefficients of the bases, where the face bases are different. Liu et al.'s face hallucination algorithm [10] was a combination of [7] and [6] to infer the global face structure and the local details, respectively.

Despite different implementation details, in an abstract sense, a learning process picks a function $f(z, \alpha)$ from an admissible function set (by specifying the index parameter α) [17]. Then, a learning-based SR algorithm can be viewed as a function s that maps an LRI to an HRI,[2] where all prior knowledge has been used to specify s, and s is a function of the input LRI only (i.e., we only consider single-image SR in this paper, and after training, no additional information can be applied for SR).

1.2 What are the Limits of Learning-Based SR Algorithms?

Among the existing algorithms, those in [6, 15, 12, 2] can be applied to general images or videos [2]. In contrast, the algorithms in [1, 10, 4, 7, 9] are devoted only to face hallucination. The underlying reason that the second category of algorithms were proposed mainly because the first category of algorithms cannot produce good results when the magnification factors are only moderately large. Therefore, the application scenario needs to be narrowed down so that more specific prior knowledge, e.g., the strong structure of faces, can be used. However, even for the second category of algorithms, accurate alignment of faces has to be done. Otherwise, the hallucinated faces are still unsatisfactory even for magnification factors that are still not very large. This poses an important question: "Do limits exist for learning-based superresolution?", i.e., "Does there exist an upper bound for magnification factors such that *no* SR algorithm can produce satisfactory results?" This paper aims at presenting our preliminary work on this problem.

To investigate the problem quantitatively, we have to define the meaning of the "limit". In statistical learning theory, the performance of a learning function $f(z, \alpha)$ is usually evaluated by its expected risk [17]:

$$R(\alpha) = \int r(z, f(z, \alpha)) \mathrm{d}F(z), \tag{1}$$

[2] By compositing with the downsampling matrix we have a function that maps an HRI to another HRI. See Eqn. (2).

where $r(z, f(z, \alpha))$ is the risk function and $F(z)$ is the probability function of z. In our problem, z represents the HRI. If we can define what the risk function is, we can use the expected risk to evaluate the performance of learning-based SR algorithms, i.e., we have to look at the average performance of SR algorithms. It is possible that an SR algorithm performs well on a particular LRI. However, if the SR results on many other LRIs are poor, we still do not consider it a good SR algorithm.

As suggested in [8], a good SR algorithm should produce HRIs that are close to the ground truth. Otherwise, the produced HRI will not be what we desire, no matter how high its resolution is (e.g., a high resolution car image will not be considered as the HRI of a low resolution face image no matter how many details it presents). Therefore, we may define the risk function as the closeness between an HRI and its superresolved version. As the root mean squared error (RMSE) is a widely used measure of image similarity in the image processing community (e.g., the peak signal to noise ratio in image compression) and also in various kinds of error analysis, we may define the risk function using the RMSE between an HRI and its superresolved version.

Although small RMSEs do not necessarily guarantee good recovery of the HRIs, large RMSEs should nonetheless imply that the recovery is poor. Therefore, we may convert the problem to a tractable one: find the upper bound of the magnification factors such that the expected risk is below a relatively *large* threshold. Such an upper bound can be considered the limits of learning-based SR algorithms.

1.3 Previous Work and Our Contributions

Although many SR approaches have been proposed [1, 2, 6, 10, 15, 4, 12, 11, 5, 7, 9], theoretical analysis of SR algorithms has rarely been addressed. Only in [8] and [1], limits of reconstruction-based SR algorithms are discussed. For learning-based algorithms, no similar work has been done. In this paper, we provide some theoretical analysis on the limits of learning-based SR algorithms for general natural images, which is the first work on this problem according to the best of our knowledge. Our paper has two major contributions:

1. A closed form lower bound of the expected error between the superresolved and the ground truth images is proved. This formula only involves the covariance matrix and the mean of the prior distribution of HRIs. This lower bound is used to estimate the limits of learning-based SR algorithms.

2. A formula on the sufficient number of HRIs is provided to ensure the accuracy of the sample-based computation of the lower bound.

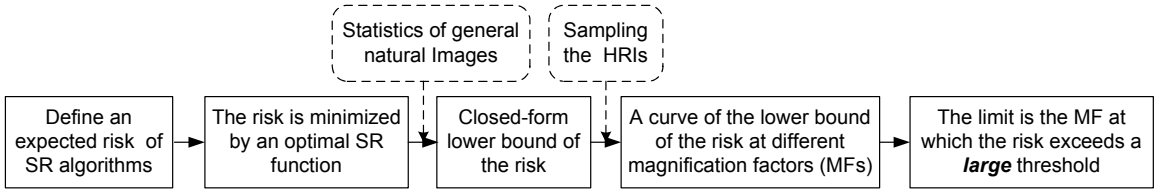

Figure 2: *Our methodology of finding the limits of learning-based SR algorithms. Please refer to Section 2 for the details.*

Moreover, from our experiments, we have observed that the limits may be independent of the sizes of both LRIs and HRIs.

Currently, we limit our analysis to general natural images, i.e., *the set of all natural images of given size*, because the statistics of general natural images have been studied for a long time [14] and there have been some pertinent observations on their characteristics that are useful for our analysis. In particular, we will use the following two properties:

1. The distribution of HRIs is not concentrated around several HRIs and the distribution of LRIs is not concentrated around several LRIs either. Noticing that general natural images cannot be classified into a small number of categories will justify this property.

2. Smoother LRIs have a higher probability than non-smooth ones. This property is actually called the "smoothness prior" that is widely used for regularization, for instance, when performing reconstruction-based SR.

In contrast, for specific class of images, e.g., face or text images, there is no similar work on their statistics to the best of our knowledge. So currently we have to focus on the SR of general natural images.

2 Analysis of Learning-Based SR Algorithms

Figure 2 outlines our analysis on the limits of learning-based SR algorithms. We first define the expected risk of a learning-based SR algorithm. The risk is minimized by an optimal SR function. Using the statistics of general natural images, we derive a closed form formula for the lower bound of the risk, which only involves the covariance matrix and the mean of the distribution of the HRIs. By sampling the real-world HRIs, we can obtain a curve of the lower bound of the risk w.r.t. the magnification factor. Finally, by choosing a relatively large threshold for the lower bound of the risk, we can roughly estimate the *limit* of the learning-based SR algorithms. We also estimate the sufficient number of image samples that indicates when to stop sampling. In the following subsections, we give the details of our analysis.

2.1 Problem Formulation

For simplicity, we present the arguments for the 1D case only. Those for the 2D case are similar but the derivation is much more complex.

As argued in Section 1.2, we use the RMSE between an HRI and the recovered HRI to evaluate the performance of a learning-based SR algorithm. This motivates us to define the following expected risk of the SR algorithm:[3]

$$\begin{aligned} g(N,m) &= \left(\frac{1}{mN}\tilde{g}(N,m)\right)^{\frac{1}{2}}, \text{ where} \\ \tilde{g}(N,m) &= \int_{\mathbf{h}} ||\mathbf{h} - s\left(\mathbf{Dh}\right)||^2 p_h(\mathbf{h})\mathrm{d}\mathbf{h}, \end{aligned} \tag{2}$$

in which s is the learnt SR function that maps N-dimensional images to mN-dimensional ones, $m > 1$ is the magnification factor and always makes mN an integer, p_h is the probability density functions of the HRIs, and $\mathbf{D}$ is the downsampling matrix that downsamples mN-dimensional signals to N-dimensional ones. The downsampling matrix is introduced here to simulate the image formation process. Although there might not be a uniform downsampling matrix for all the HRIs and some image formation process may even involve a nonlinear transform on the HRI, we may nevertheless throw all the discrepancy from our model into noise $\mathbf{n}$ by replacing $s\left(\mathbf{Dh}\right)$ with $s\left(\mathbf{Dh} + \mathbf{n}\right)$. However, the discussion on the effect of noise (see (22)) is deferred to our future work.

Eqn. (2) defines the expected risk of a particular SR algorithm s, which should be evaluated by running the algorithm on a large number of HRIs. This is very time consuming. Moreover, for a particular SR algorithm, its magnification factor is often fixed. Therefore, estimating the expected risk of a *particular* SR function does not help to find the limits of *all* learning-based SR algorithms. Consequently, we have to study the lower bound of (2).

Before going on, we first introduce the corresponding upsampling matrix $\mathbf{U}$ which upsamples N-dimensional signals to mN-dimensional ones. We expect that images are unchanged if they are upsampled and then downsampled. This implies that $\mathbf{DU} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This upsampling matrix is purely a mathematical tool to facilitate the derivation and the representation of our results. We also use $\mathbf{\Sigma}$ and $\bar{\mathbf{h}}$ to denote the covariance matrix and the mean of the HRIs $\mathbf{h}$, respectively.

2.2 Main Results

The central theorem of our paper is the following:

[3]Throughout our paper, vectors or matrices are written in boldface, while scalars are in normal fonts. Moreover, all the vectors without the transpose are column vectors.

Theorem 2.1 *(Lower Bound of the Expected Risk) When $p_h(\mathbf{h})$ is the distribution of general natural images, namely the set of* all *natural images, $\tilde{g}(N,m)$ is effectively lower bounded by $\tilde{b}(N,m)$, where*

$$\begin{aligned}\tilde{b}(N,m) = \ & \frac{1}{4}tr\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^t\right] \\ & +\frac{1}{4}\left|\left|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}\right|\right|^2,\end{aligned} \tag{3}$$

in which $tr(\cdot)$ is the trace operator and the superscript ^{t} represents the matrix or vector transpose. Hence $g(N,m)$ is lower bounded by

$$b(N,m) = \left(\frac{1}{mN}\tilde{b}(N,m)\right)^{\frac{1}{2}}. \tag{4}$$

As for an HRI $\mathbf{h}$, $(\mathbf{I}-\mathbf{UD})\mathbf{h} = \mathbf{h}-\mathbf{U}(\mathbf{Dh})$ is its high frequency. So Eqn. (3) is essentially related to the richness of the high frequency component in the HRIs. Hence Theorem 2.1 implies that the richer the high frequency component in the HRIs is, the more difficult the SR is.

Note that Theorem 2.1 holds for all possible SR functions s as it gives the lower bound of the risk, which we have derived conservatively. Consequently, the estimate on the limits of learning-based SR algorithms using (4) is also conservative. And also note that $p_h(\mathbf{h})$ being the distribution of the set of all natural images is important for us to arrive at (3). Otherwise, we will not come up with the coefficient 1/4 therein and $\tilde{g}(N,m)$ may be arbitrarily close to 0. For example, if there is only one HRI, we can always recover the HRI no matter how low resolution the LRI is.

As a simple yet effective analytical model for $p_h(\mathbf{h})$ of general natural images is unavailable, we sample real HRIs to estimate $\tilde{b}(N,m)$. To make sure that sufficient images have been sampled to achieve an accurate estimate of $\tilde{b}(N,m)$, we further prove the following theorem:

Theorem 2.2 *(Sufficient Number of Samples) If we sample $M(p,\varepsilon)$ HRIs independently, then with probability of at least $1-p$, $|\hat{\tilde{b}}(N,m)-\tilde{b}(N,m)|<\varepsilon$, where $\hat{\tilde{b}}(N,m)$ is the value of $\tilde{b}(N,m)$ estimated from real samples,*[4]

$$M(p,\varepsilon) = \frac{(C_1+2C_2)^2}{16p\varepsilon^2}, \tag{5}$$

[4]Throughout our paper, we use the embellishment $\wedge$ above a value to represent the sampled or estimated quantities.

$C_1 = \sqrt{E\left(\left\|(\mathbf{I}-\mathbf{UD})(\mathbf{h}-\bar{\mathbf{h}})\right\|^4\right) - tr^2\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^t\right]}$, *and* $C_2 = \sqrt{\bar{\mathbf{b}}^t\mathbf{\Sigma}\bar{\mathbf{b}}}$, *in which* $E(\cdot)$ *is the expectation operator and* $\bar{\mathbf{b}} = (\mathbf{I}-\mathbf{UD})^t(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}$.

Note that both C_1 and C_2 are related to the variance of the high frequency component of the HRIs. So Theorem 2.2 implies that the larger the variance is, the more samples are required.

In subsections 2.3 and 2.6, we will provide the sketches of proving the above two theorems.

2.3 Lower Bound of the Expected Risk

In this subsection, we present the idea of proving Theorem 2.1. Now that different HRIs can result in the same LRI ($\mathbf{Dh}$ can be identical for different $\mathbf{h}$), it may be easier to analyze (2) by fixing $\mathbf{Dh}$. This can be achieved by performing a variable transform in (2). To do so, we find a complementary matrix (not unique) $\mathbf{Q}$ such that $\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}$ is a non-singular square matrix and $\mathbf{QU} = \mathbf{0}$. Such a $\mathbf{Q}$ exists. The proof can be found in Appendix. Denote $\mathbf{M} = (\mathbf{R} \quad \mathbf{V}) = \begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}^{-1}$. From $\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix}(\mathbf{R} \quad \mathbf{V}) = \mathbf{I}$, we know that $\mathbf{R} = \mathbf{U}$.

Now we perform a variable transform $\mathbf{h} = \mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}$, then (2) becomes

$$\begin{aligned}\tilde{g}(N,m) = & \int_{\mathbf{x},\mathbf{y}} \left\|(\mathbf{U} \quad \mathbf{V})\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} - s(\mathbf{x})\right\|^2 \\ & \times p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{dxdy} \\ = & \int_{\mathbf{x}} p_x(\mathbf{x})V(\mathbf{x})\mathrm{dx},\end{aligned} \tag{6}$$

where

$$\begin{aligned}p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right) &= |\mathbf{M}|p_h\left(\mathbf{M}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right), \\ V(\mathbf{x}) &= \int_{\mathbf{y}} ||\mathbf{Vy} - \phi(\mathbf{x})||^2 \tilde{p}_y(\mathbf{y}|\,\mathbf{x})\,\mathrm{dy}.\end{aligned} \tag{7}$$

$p_x(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$, $\tilde{p}_y(\mathbf{y}|\,\mathbf{x})$ is the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, and $\phi(\mathbf{x}) = s(\mathbf{x}) - \mathbf{Ux}$ is the recovered high frequency component of the HRI given the LRI $\mathbf{x}$. For this reason, we call $\phi(\mathbf{x})$ the high frequency (HF) function. Note that $\mathbf{x} = \mathbf{Dh}$, and $\mathbf{Vy} = \mathbf{h} - \mathbf{Ux}$. So $\mathbf{x}$ is the LRI downsampled from $\mathbf{h}$, and $\mathbf{Vy}$ is the high frequency of $\mathbf{h}$.

One can see that there is an optimal HF function such that $V(\mathbf{x})$ (hence $g(N,m)$) is minimized:

$$\phi_{opt}(\mathbf{x};\tilde{p}_y) = \mathbf{V}\int_{\mathbf{y}} \mathbf{y}\tilde{p}_y\left(\mathbf{y}|\,\mathbf{x}\right)\mathrm{d}\mathbf{y}, \tag{8}$$

where $\phi_{opt}(\mathbf{x};p)$ denotes the optimal ϕ w.r.t. the distribution p. This means that the optimal high frequency component should be the expectation of all possible high frequencies associated to the LRI $\mathbf{x}$.

Then one can easily verify that

$$V(\mathbf{x}) = \int_{\mathbf{y}} ||\mathbf{V}\mathbf{y}||^2 \tilde{p}_y\left(\mathbf{y}|\,\mathbf{x}\right)\mathrm{d}\mathbf{y} - ||\phi_{opt}(\mathbf{x};\tilde{p}_y)||^2. \tag{9}$$

In Appendix, we show that for general natural images,

$$\int_{\mathbf{x}} p_x(\mathbf{x})\,||\phi_{opt}(\mathbf{x};\tilde{p}_y)||^2\,\mathrm{d}\mathbf{x} \le \frac{3}{4}\int_{\mathbf{x}}\int_{\mathbf{y}} ||\mathbf{V}\mathbf{y}||^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x}. \tag{10}$$

Therefore, from (6) and (9) we have that

$$\begin{aligned}
\tilde{g}(N,m) &= \int_{\mathbf{x}} p_x(\mathbf{x})\left(\int_{\mathbf{y}} ||\mathbf{V}\mathbf{y}||^2 \tilde{p}_y\left(\mathbf{y}|\,\mathbf{x}\right)\mathrm{d}\mathbf{y} - ||\phi_{opt}(\mathbf{x};\tilde{p}_y)||^2\right)\mathrm{d}\mathbf{x}\\
&\ge \frac{1}{4}\int_{\mathbf{x},\mathbf{y}} ||\mathbf{V}\mathbf{y}||^2 p_{x,y}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\
&= \frac{1}{4}\int_{\mathbf{h}} ||\mathbf{V}\mathbf{Q}\mathbf{h}||^2 p_h(\mathbf{h})\mathrm{d}\mathbf{h}\\
&= \frac{1}{4}tr\left((\mathbf{I}-\mathbf{U}\mathbf{D})\mathbf{\Sigma}(\mathbf{I}-\mathbf{U}\mathbf{D})^t\right) + \frac{1}{4}\left|\left|(\mathbf{I}-\mathbf{U}\mathbf{D})\bar{\mathbf{h}}\right|\right|^2,
\end{aligned} \tag{11}$$

where we have used $\mathbf{V}\mathbf{Q} = \mathbf{I} - \mathbf{U}\mathbf{D}$, which comes from $\begin{pmatrix}\mathbf{U} & \mathbf{V}\end{pmatrix}\begin{pmatrix}\mathbf{D}\\ \mathbf{Q}\end{pmatrix} = \mathbf{I}$. This proves Theorem 2.1.

We see that the variance and the mean of the HRIs plays a key role in lower bounding $g(N,m)$. Although it is intuitive that $p_h(\mathbf{h})$ is critical for the limits of learning-based SR algorithms, Theorem 2.1 exactly depicts how $p_h(\mathbf{h})$ influences the SR performance.

2.4 Limits of Learning-Based SR Algorithms

The introduction of the optimal HF function (or equivalently, the optimal SR function, as $s_{opt}(\mathbf{x}) = \phi_{opt}(\mathbf{x}) + \mathbf{U}\mathbf{x}$) frees us from dealing with the details of different learning-based

SR algorithms, as s_{opt} attains the minimum of the expected risk. In other words, if at a particular magnification factor, $b(N, m)$ (see Eqn. (4)) is larger than a threshold T, i.e., the expected RMSE between $\mathbf{h}$ and $s_{opt}(\mathbf{Dh})$ is larger than T, then for *any* SR function s, the RMSE between $\mathbf{h}$ and $s(\mathbf{Dh})$ is also expected to be larger than T. This will imply that at this magnification factor no SR function can effectively recover the original HRI.

Therefore, if we have full knowledge of the variance and the mean of the prior distributions $p_h(\mathbf{h})$ at different magnification factors, we can define a curve of $b(N, m)$ as a function of m. Then the limit of learning-based SR algorithms is upper bounded by $b^{-1}(T)$.

2.5 Estimating the Lower Bound from Real Samples

To compute $b(N, m)$, we have to know the covariance matrix and the mean of HRIs $\mathbf{h}$ for a wide range of mN. There has been a long history of natural image statistics [14]. Unfortunately, all the existing models only solve the problem partially: the natural images fit some models, but not all images that are sampled from these models are natural images. On the other hand, we do not need full knowledge of $p_h(\mathbf{h})$: its covariance matrix and mean already suffice. This motivates us to sample HRIs from real data.

Thanks to the fine property of covariance matrices and means that they can be computed incrementally and in parallel, we can easily sample a huge number of HRIs at a low memory cost.

2.6 The Sufficient Number of HRI Samples

Now that we have estimated the lower bound from HRI samples, we have to know how many samples are sufficient to achieve the required accuracy. We first denote $\hat{\bar{\mathbf{\Sigma}}}_M = \frac{1}{M}\sum_{k=1}^{M}(\hat{\mathbf{h}}_k - \bar{\mathbf{h}})(\hat{\mathbf{h}}_k - \bar{\mathbf{h}})^t$ and the estimated covariance matrix $\hat{\mathbf{\Sigma}}_M = \frac{1}{M}\sum_{k=1}^{M}(\hat{\mathbf{h}}_k - \hat{\bar{\mathbf{h}}})(\hat{\mathbf{h}}_k - \hat{\bar{\mathbf{h}}})^t$,[5] where $\hat{\mathbf{h}}_k$'s are i.i.d. samples and $\hat{\bar{\mathbf{h}}} = \frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{h}}_k$ is the estimated mean. Then one may check that

$$\hat{\mathbf{\Sigma}}_M = \hat{\bar{\mathbf{\Sigma}}}_M - (\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}})(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}})^t. \tag{12}$$

[5]For unbiased estimation of the covariance matrix, the coefficient before the summation should be $1/(M-1)$. However, we are more interested in the error of $\tilde{b}(N, m)$, rather than the covariance matrix itself. If $1/(M-1)$ is used instead of $1/M$, there will be an additional $O(1/M)$ term at the right hand side of (19), indicating that the convergence might be slightly slower. Nonetheless, when M is large, the difference is negligible.

In the following, we denote $\mathbf{B} = (\mathbf{I} - \mathbf{UD})^t(\mathbf{I} - \mathbf{UD})$ and $\bar{\mathbf{b}} = \mathbf{B}\bar{\mathbf{h}}$ for brevity, and denote the i-th entry of a vector $\mathbf{a}$ as a_i and the (i, j)-th entry of a matrix $\mathbf{A}$ as A_{ij}.

With some calculation we have

$$\begin{aligned} & |\hat{\tilde{b}}(N, m) - \tilde{b}(N, m)| \\ \leq \ & \frac{1}{4}\left|\sum_{i,j=1}^{mN} B_{ij}\left(\hat{\bar{\Sigma}}_{M;ij} - \Sigma_{ij}\right)\right| + \frac{1}{2}\left|\sum_{i=1}^{mN} \bar{b}_i(\hat{\bar{h}}_{M;i} - \bar{h}_i)\right|. \end{aligned} \tag{13}$$

The details can be found in Appendix. So we have to estimate the convergence rates of both terms.

Therefore, we define $\xi = \sum_{i,j=1}^{mN} B_{ij}\hat{\bar{\Sigma}}_{M;ij} = tr(\mathbf{B}\hat{\bar{\mathbf{\Sigma}}})$ and $\eta = \sum_{i=1}^{mN} \bar{b}_i\hat{\bar{h}}_{M;i} = \bar{\mathbf{b}}^t\hat{\bar{\mathbf{h}}}_M$. Then their expectations are

$$E\left(\xi\right) = tr(\mathbf{B\Sigma}), \text{ and } E(\eta) = \bar{\mathbf{b}}^t\bar{\mathbf{h}}, \tag{14}$$

respectively. And their variances can be found to be

$$var(\xi) = \frac{C_1^2}{M}, \tag{15}$$

and

$$var(\eta) = \frac{C_2^2}{M}, \tag{16}$$

respectively. The proofs can be found in Appendix.

Then by Chebyshev's inequality [13],

$$\begin{aligned} P\left(\left|\sum_{i,j=1}^{mN} B_{ij}\left(\hat{\bar{\Sigma}}_{M;ij} - \Sigma_{ij}\right)\right| \geq \delta\right) &= P\left(|\xi - E(\xi)| \geq \delta\right) \leq \frac{var(\xi)}{\delta^2} = \frac{C_1^2}{M\delta^2}, \\ P\left(\left|\sum_{i=1}^{mN} \bar{b}_i\left(\hat{\bar{h}}_{M;i} - \bar{h}_i\right)\right| \geq \delta\right) &= P\left(|\eta - E(\eta)| \geq \delta\right) \leq \frac{var(\eta)}{\delta^2} = \frac{C_2^2}{M\delta^2}. \end{aligned} \tag{17}$$

Therefore, at least at a probability of $1 - p$,

$$\begin{aligned} \left|\sum_{i,j=1}^{mN} B_{ij}\left(\hat{\bar{\Sigma}}_{M;ij} - \Sigma_{ij}\right)\right| &\leq \frac{C_1}{\sqrt{Mp}}, \\ \left|\sum_{i=1}^{mN} \bar{b}_i\left(\hat{\bar{h}}_{M;i} - \bar{h}_i\right)\right| &\leq \frac{C_2}{\sqrt{Mp}}. \end{aligned} \tag{18}$$

Then by (13) and (18), with probability at least $1 - p$, we have

$$\left|\hat{\tilde{b}}(N, m) - \tilde{b}(N, m)\right| \leq \frac{C_1}{4\sqrt{Mp}} + \frac{C_2}{2\sqrt{Mp}}. \tag{19}$$

Now one can check that Theorem 2.2 is true.

Note that

$$\begin{aligned}
& \left|\hat{b}(N,m) - b(N,m)\right| \\
= & \left|\sqrt{\frac{1}{mN}\hat{\tilde{b}}(N,m)} - \sqrt{\frac{1}{mN}\tilde{b}(N,m)}\right| \\
= & \frac{\frac{1}{mN}\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{\sqrt{\frac{1}{mN}\hat{\tilde{b}}(N,m)} + \sqrt{\frac{1}{mN}\tilde{b}(N,m)}} \\
\approx & \frac{\frac{1}{mN}\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{2\sqrt{\frac{1}{mN}\tilde{b}(N,m)}} \\
= & \frac{\left|\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)\right|}{2mNb(N,m)}.
\end{aligned} \tag{20}$$

So in practice, we may choose $p = 0.01$ and $\varepsilon = \frac{1}{2}mNb(N,m)$ in (5) in order to make $\left|\hat{b}(N,m) - b(N,m)\right| \leq 0.25$ at above 99% certainty. Here we choose 0.25 as the threshold because it is roughly the mean of the graylevel quantization error.

3 Experiments

Collecting Samples. We crawled images from the web and collected 100,000+ images. They are of various kinds of scenes: cityscape, landscape, sports, portraits, etc. Therefore, our image library could be viewed as an i.i.d. sampling of general natural images. To sample $mN \times mN$ sized HRIs, we convert each image into graylevel, break it into non-overlapping patches of size $mN \times mN$ (with at least one pixel gap among them in order to ensure independence among them), and view each patch as a sample of HRIs of size $mN \times mN$. Then we blindly run our program to estimate the covariance and mean of the HRIs, where mN varies from 8 to 48 at a step size of 4. The number of samples is at the scale of 10^6 to 10^8. Note that such a scale may not be enough in estimating $p_h(\mathbf{h})$. But estimating $p_h(\mathbf{h})$ is not our goal at all. We are interested in the values of $b(N,m)$ only.

Characteristics of $\mathbf{b(N,m)}$. Next, we have to specify a downsampling matrix in order to compute the lower bound $b(N,m)$ by (4) (The upsampling matrix $\mathbf{U}$ is determined by $\mathbf{D}$. See Section 2.1.). We simply choose a downsampling matrix that corresponds to the bicubic

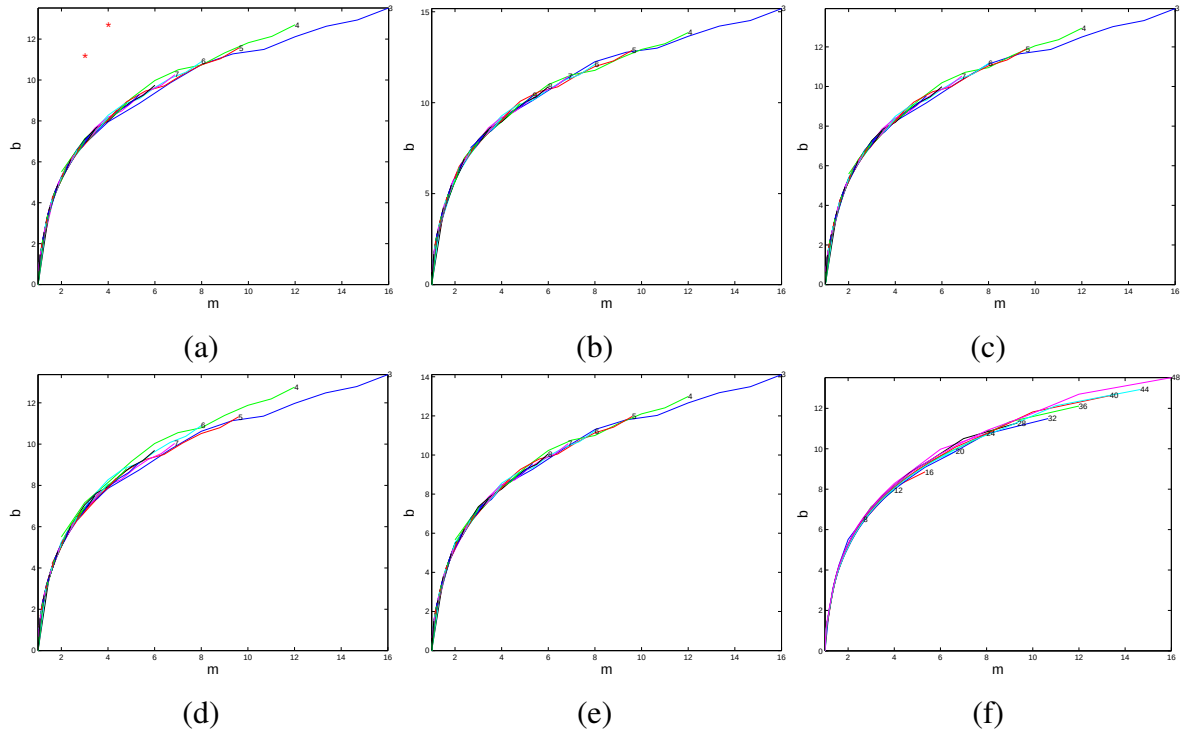

Figure 3: *(a)∼(e) are curves of $b(N, m)$ using different $\mathbf{D}$'s, drawn with N fixed for each individual curve. The corresponding N's are labelled at the tails of the curves (in order not to make the graph crowded, large N's for short curves are not shown). (a) uses a bicubic filter with $a = -1$. The asterisks at (3, 11.1) and (4,12.6) represent the expected risks of Sun et al.'s [15] and Freeman et al.'s [6] SR algorithms, respectively. (b) uses a bicubic filter with $a = 0.5$. (c) uses a Gaussian filter with $\sigma = 0.5$. (d) uses a Gaussian filter with $\sigma = 1.5$. (e) uses the bilinear filter. (f) are the curves of $b(N, m)$ with mN fixed. The corresponding mN's are labelled at the tails of the curves. The filter used is the same as that in (a).*

B-spline filter.[6] Then the curves of $b(N, m)$ w.r.t. m are shown in Figure 3(a), where for each individual curve N is fixed.

We can see that for fixed N, $b_N^{(1)}(m) = b(N, m)$ increases with m. A remarkable observation is that for different N's, the curves in Figure 3(a) coincide well with each other. This suggests that for general natural images $b(N, m)$ may be *independent* of N. Another interest-

[6] In the 1D case, a cubic filter can be written as:

$$k(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1, & \text{if } 0 \le |x| \le 1, \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a, & \text{if } 1 \le |x| \le 2, \\ 0, & \text{if } |x| > 2. \end{cases} \tag{21}$$

When $a = -1$, it is the cubic B-spline filter. The downsampling matrix for 2D images is the Kronecker product of the 1D downsampling matrices.

Figure 4: *Part of the SR results using Sun et al.'s algorithm [15]. The magnification factor is 3.0. On the top are the LRIs of* 16×16*, interpolated to* 48×48 *using bicubic interpolation. In the middle are the SR results. At the bottom are the ground truth HRIs.*

ing observation on Figures 3(a) is that $b_N^{(1)}(m)$ seems to grow at the rate of $(m-1)^{1/2}$. The important implication from these observations is: we may estimate the limits of learning-based SR by trying relatively small sized images and small magnification factors, rather than trying large sized images and large magnification factors, which saves computation and memory without compromising the estimation accuracy.

However, one should be cautious that strictly speaking the $\mathbf{D}$ in (2) should be estimated from real cameras. Fortunately, we have found that our lower bound does not seem to be very sensitive to the choice of $\mathbf{D}$. We have tried the bilinear filter, Gaussian filters (with the variance varying from 0.5^2 to 1.5^2), and bicubic filters (with the parameter a varying from -1 to 0.5, see (21)), and have found that the lower bounds are fairly close to each other. The curves in Figures 3(a)$\sim$(e) testify to this observation. Moreover, what we have observed in the last paragraph is still true.

When training learning based SR algorithms, one usually collects HRIs and downsamples them to LRIs. So it is also helpful to draw the curves by fixing mN instead. The same phenomenon mentioned above can also be observed (Figure 3(f)). And the curves of $b_{mN}^{(2)}(m) = b(N, m)$ by fixing mN also coincide well with those of $b_N^{(1)}(m)$ (Please compare Figures 3(a) and (f)), implying that $b(N, m)$ is also independent of the size of HRIs. This can be easily proved: if $b(N, m) = c(m)$ for some function c, then $b_{mN}^{(2)}(m) = b(N, m) = b_N^{(1)}(m) = c(m)$.

Testing Theorem 2.1. We run the SR algorithm by Sun et al. [15] on over 50,000 16×16 LRIs that are downsampled from 48×48 HRIs and that by Freeman et al. [6] on over 40,000 12×12 LRIs that are downsampled from 48×48 HRIs. Both algorithms are designed for general images and they work at magnification factors of 3.0 and 4.0, respectively. A few sample results are shown in Figure 4. The expected risks of Sun et al.'s algorithm and

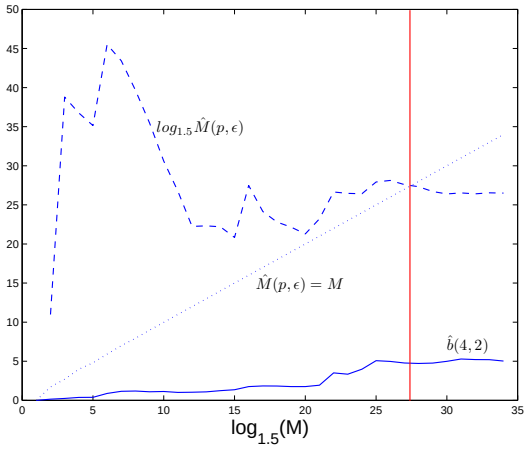

Figure 5: *The evolution of $\hat{b}(4,2)$ w.r.t. the number M of HRI samples. The dashed curve is the log of the estimated sufficient samples using the currently available covariance matrix and mean. The dotted line is used to identify when the estimated number of samples is enough. The solid curve at the bottom is the estimated $b(4,2)$ using M samples. The horizontal axis is in log scale.*

Freeman et al.'s are about 11.1 and 12.6, respectively, which are both above our curves (Figure 3(a)). Therefore, these results are consistent with Theorem 2.1.

Estimating the Limits. With the curves of $b(N,m)$, we can find the limits of learning based algorithms by choosing an appropriate threshold T (see Section 2.4). Unfortunately, there does not seem to exist a benchmark threshold. So every practitioner can choose a threshold that he/she deems appropriate and estimate the limits on his/her own. For example, from the SR results of Sun et al.'s algorithm [15] (Figure 4), we see that the fine details are already missing. Therefore, we deem that the estimated risk 11.1 of their algorithm is a large enough threshold. Using $T = 11.1$ we can expect that the limit of learning-based SR algorithms for general natural images is roughly 10 (Figure 3(a)). This limit is a bit loose but it can be enhanced when the noise in LRIs (see Section 2.1) is considered.[7]

Testing Theorem 2.2. Finally, we present an experiment to test Theorem 2.2. We sample over 1.5 million 8×8 images and set $m = 2$ (hence $N = 4$). Figure 5 shows the curve of predicted sufficient number of samples using the most updated variance and mean of HRIs, where p and ε are chosen as described at the end of Section 2.6. We see that the estimated

[7]When noise is considered, Eqn. (3) is changed to

$$\begin{aligned}\tilde{b}(N,m) = \; & \frac{1}{4}tr\left[(\mathbf{I}-\mathbf{UD})\mathbf{\Sigma}(\mathbf{I}-\mathbf{UD})^t\right] + \frac{1}{4}tr\left(\mathbf{U}\mathbf{\Sigma}_n\mathbf{U}^t\right) \\ & + \frac{1}{4}\left|\left|(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}} - \mathbf{U}\bar{\mathbf{n}}\right|\right|^2,\end{aligned} \tag{22}$$

where $\mathbf{\Sigma}_n$ and $\bar{\mathbf{n}}$ are the variance matrix and the mean of the noise, respectively. Details omitted.

$b(4,2)$ already becomes stable even the number of samples is still smaller than the predicted number. There is still small fluctuation in $b(4,2)$ when $M > \hat{M}(p,\varepsilon)$ because we allow the deviation from the true value to be within 0.25 at above 99% certainty. Therefore, this result is consistent with Theorem 2.2.

4 Conclusions and Future Work

This paper presents the first attempt to analyze the limits of learning-based SR algorithms. We have proven a closed form lower bound of the expected risk of SR algorithms. We also sample real images to estimate the lower bound. Finally, we prove the formula that gives the sufficient number of HRIs to be sampled in order to ensure the accuracy of the estimate.

We have also observed from experiments that the lower bound $b(N,m)$ may be dependent on m only and the growth rate of $b(N,m)$ may be $(m-1)^{1/2}$. These are important observations, implying that one may more conveniently compute with small sized images and at small magnification factors and then *predict* the limits. This would save much computation and memory. We hope to prove in the future what we have observed.

As no authoritative threshold T is currently available, our estimated limit (roughly 10 times) of learning-based SR algorithms for general natural images is not convincing enough. We are investigating how to propose an objective threshold and how to effectively sample the statistics of noise in (22) to produce a tighter limit.

Also, we will investigate the limits of learning-based SR algorithms under more specific scenarios, e.g., for face hallucination and text SR. We expect that more specific prior knowledge of the HRI distribution will be required.

References

[1] S. Baker and T. Kanade. Limits on Super-resolution and How to Break Them. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.24, No.9, pp.1167-1183, 2002.

[2] C.M. Bishop, A. Blake, and B. Marthi. Super-resolution Enhancement of Video. In C. M. Bishop and B. Frey (Eds.), Proc. Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics, 2003.

[3] S. Borman and R.L. Stevenson. Spatial Resolution Enhancement of Low-Resolution Image Sequences: A Comprehensive Review with Directions for Future Research, Technical Report, University of Notre Dame, 1998.

[4] D. Capel and A. Zisserman. Super-Resolution from Multiple Views Using Learnt Image Models. Proc. Computer Vision and Pattern Recognition, pp. II 627-634, 2001.

[5] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and Challenges in Super-Resolution. Int'l J. Imaging Systems and Technology, Vol. 14, No. 2, pp. 47-57, 2004.

[6] W.T. Freeman and E.C. Pasztor. Learning Low-Level Vision. In Proc. Seventh Int'l Conf. Computer Vision, Corfu, Greece, pp. 1182-1189, 1999.

[7] B.K. Gunturk, A.U. Batur, Y. Altunbasak, M.H. Hayes III, and R.M. Mersereau. Eigenface-domain super-resolution for face recognition. IEEE Trans. on Image Process., Vol. 12, No. 5, pp. 597-606, 2003.

[8] Z. Lin and H.-Y. Shum. Fundamental Limits of Reconstruction-Based Superresolution Algorithms under Local Translation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.26, No.1, pp.83-97, 2004.

[9] W. Liu, D. Lin, and X. Tang. Hallucinating Faces: TensorPatch Super-resolution and Coupled Residue Compensation. Proc. Computer Vision and Pattern Recognition, pp. II 478-484, 2005.

[10] C. Liu, H.Y. Shum, and C.S. Zhang. A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model. Proc. Computer Vision and Pattern Recognition, pp. 192-198, 2001.

[11] S.C. Park, M.K. Park, and M.G. Kang. Super-Resolution Image Reconstruction: A Technical Overview. IEEE Signal Processing Magazine, Vol. 20, Pt. 3, pp. 21-36, 2003.

[12] L.C. Pickup, S.J. Roberts, and A. Zisserman. A Sample Texture Prior for Image Super-resolution. Advances in Neural Information Processing Systems, pp. 1587-1594, 2003.

[13] A.N. Shiryaev. Probability. Spinger-Verlag, 1995.

[14] A. Srivastava, A.B. Lee, E.P. Simoncelli, and S.-C. Zhu. On Advances in Statistical Modeling of Natural Images, J. Mathematical Imaging and Vision, 18: 17-33, 2003.

[15] J. Sun, H. Tao, and H.-Y. Shum. Image Hallucination with Primal Sketch Priors, Proc. Computer Vision and Pattern Recognition, pp. II 729-736, 2003.

[16] Y.L. Tong. Probability Inequalities in Multivariate Distributions. Academic Press, 1980.

[17] V.N. Vapnik. Statistical Learning Theory. John Wiley & Sons, Inc., 1998.

[18] X. Wang and X. Tang. Hallucinating Face by Eigentransformation. IEEE Trans. Systems, Man, and Cybernetics, Part C, vol. 35, no. 3, pp. 425-434, 2005.

[19] R. Wilson. MGMM: Multiresolution Gaussian Mixture Models for Computer Vision. Proc. Int'l Conf. Pattern Recognition, pp. I 212-215, 2000.

Appendix

Proposition 4.1 $\mathbf{Q}$ *exists.*

Proof: Suppose the SVD of $\mathbf{U}$ is: $\mathbf{U} = \mathbf{O}_1 \begin{pmatrix} \mathbf{\Lambda} \\ \mathbf{0} \end{pmatrix} \mathbf{O}_2^t$, where $\mathbf{\Lambda}$ is a non-degenerate square matrix. Then all the solutions to $\mathbf{XU} = \mathbf{0}$ can be written as: $\mathbf{X} = \mathbf{O}_2 \begin{pmatrix} \mathbf{0} & \mathbf{Y} \end{pmatrix} \mathbf{O}_1^t$, where $\mathbf{Y}$ is any matrix of proper size. On the other hand, from $\mathbf{DU} = \mathbf{I}$ we know that there exists some $\mathbf{Y}_0$ such that $\mathbf{D} = \mathbf{O}_2 \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{Y}_0 \end{pmatrix} \mathbf{O}_1^t$. Therefore, $\begin{pmatrix} \mathbf{D} \\ \mathbf{Q} \end{pmatrix} = \mathbf{O}_2 \begin{pmatrix} \mathbf{\Lambda}^{-1} & \mathbf{Y}_0 \\ \mathbf{0} & \mathbf{Y} \end{pmatrix} \mathbf{O}_1^t$. When $\mathbf{Y}$ is of full-rank, $\begin{pmatrix} \mathbf{D} \\ \mathbf{Q} \end{pmatrix}$ is a non-degenerate square matrix.

Proposition 4.2 *(10) is true.*

Proof: The optimal HF function given in (8) is inconvenient for estimating a lower bound for $\tilde{g}(N, m)$, because we do not know $\mathbf{V}$ and $\tilde{p}_y(\mathbf{y}|\,\mathbf{x})$ therein. To overcome this, we assume that the density of HRIs is provided by the mixture of Gaussians (MoGs):

$$p_h(\mathbf{h}) = \sum_{k=1}^{K} \alpha_k G_{h;k}(\mathbf{h}), \tag{23}$$

where $\alpha_k > 0$, $\sum_{k=1}^{K} \alpha_k = 1$, $G_{h;k}(\mathbf{h}) = G(\mathbf{h}; \mathbf{h}_k, \mathbf{\Sigma}_k)$ is the Gaussian with mean $\mathbf{h}_k$ and variance $\mathbf{\Sigma}_k$. Note that the above MoGs approximation may not give an exact $p_h(\mathbf{h})$. However, as every L_2 function can be approximated by MoGs at an arbitrary accuracy (in the sense of L_2 norm) [19], and $\mathbf{h} - s(\mathbf{Dh})$ must be bounded (e.g., every dimension is between -255 and 255), when the MoGs approximation is sufficiently accurate, we will give a sufficiently accurate estimate of $\tilde{g}(N, m)$. Therefore, in order not to introduce new notations, we simply write $p_h(\mathbf{h})$ as MoGs in our proof. More importantly, as we will see, MoGs actually serve as a bridge to pave our proving process. Our final results do *not* involve any parameters from MoGs, as shown in Theorem 2.1.

Writing in MoGs, we have

$$\begin{aligned} p_{x,y}\left(\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}\right) &= \sum_{k=1}^{K} \alpha_k G_{x,y;k}\left(\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}\right), \\ p_x(\mathbf{x}) &= \sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}), \\ \tilde{p}_y(\mathbf{y}|\,\mathbf{x}) &= \frac{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \tilde{G}_{y;k}(\mathbf{y}|\,\mathbf{x})}{\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})}, \end{aligned} \tag{24}$$

where $G_{x,y;k}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)$ is the Gaussian corresponding to $G_{h;k}(\mathbf{h})$ after the variable transform, $G_{x;k}(\mathbf{x})$ is the marginal distribution of $G_{x,y;k}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)$, and

$$\tilde{G}_{y;k}\left(\mathbf{y}|\,\mathbf{x}\right)=\frac{G_{x,y;k}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)}{G_{x;k}(\mathbf{x})} \tag{25}$$

is the conditional distribution. As we will not use the exact formulation of $G_{x,y;k}\left(\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}\right)$, $G_{x;k}(\mathbf{x})$ and $\tilde{G}_{y;k}\left(\mathbf{y}|\,\mathbf{x}\right)$, we omit their details.

Now $\phi_{opt}(\mathbf{x};\tilde{p}_y)$ can be written as

$$\begin{aligned}\phi_{opt}(\mathbf{x};\tilde{p}_y) &= \frac{\sum_{k=1}^{K}\alpha_k G_{x;k}(\mathbf{x})\mathbf{V}\int_{\mathbf{y}}\mathbf{y}\tilde{G}_{y;k}\left(\mathbf{y}|\,\mathbf{x}\right)\mathrm{d}\mathbf{y}}{\sum_{k=1}^{K}\alpha_k G_{x;k}(\mathbf{x})}\\ &= \frac{\sum_{k=1}^{K}\alpha_k G_{x;k}(\mathbf{x})\phi_{opt}(\mathbf{x};\tilde{G}_{y;k})}{\sum_{k=1}^{K}\alpha_k G_{x;k}(\mathbf{x})},\end{aligned} \tag{26}$$

where

$$\phi_{opt}(\mathbf{x};\tilde{G}_{y;k}) = \mathbf{V}\int_{\mathbf{y}}\mathbf{y}\tilde{G}_{y;k}\left(\mathbf{y}|\,\mathbf{x}\right)\mathrm{d}\mathbf{y}. \tag{27}$$

Next, we highlight two properties that general natural images have, and we will use them for our argument:

1. The prior distribution $p_h(\mathbf{h})$ is not concentrated around several HRIs and the marginal distribution $p_x(\mathbf{x})$ is not concentrated around several LRIs either. Noticing that general natural images cannot be classified into a small number of categories will testify to this. This property implies that the number K of Gaussians to approximate $p_h(\mathbf{h})$ is not too small, and for every $\mathbf{x}$, $\phi_{opt}\left(\mathbf{x};\tilde{G}_{y;k}\right)$, $k=1,\cdots,K$, are most likely quite different from each other.

2. Smoother LRIs have higher probability. This property is actually called the "smoothness prior" that is widely used for regularization, e.g., when doing reconstruction based SR. An ideal mathematical formulation of this property is [14]: $p_x(\mathbf{x})\sim\exp\left(-\frac{1}{2}\beta\left|\left|\nabla\mathbf{x}\right|\right|^2\right)$.

Now we utilize the above two properties to argue for (10). We aim at estimating a reasonable constant μ, such that we are sure that the following inequality holds:

$$\int \left(\sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \right) ||\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right)||^2 \, \mathrm{d}\mathbf{x} \leq \mu \cdot \sum_{k=1}^{K} \alpha_k \int G_{x;k}(\mathbf{x}) \left|\left|\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)\right|\right|^2 \mathrm{d}\mathbf{x}. \quad (28)$$

We first infer a reasonable distribution for μ', such that most likely the following inequality holds:

$$||\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right)||^2 \leq \mu' \cdot \frac{\sum\limits_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \left|\left|\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)\right|\right|^2}{\sum\limits_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})}, \quad \forall \mathbf{x} \text{ that } \sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \neq 0. \quad (29)$$

Eqn. (26) shows that $\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right)$ is a convex combination of $\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)$, $k = 1, \cdots, K$. Due to the convexity of the squared vector norm, by Jensen's inequality [13], we have that $\mu' \leq 1$ is always true, where $\mu' = 1$ holds only when $\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)$, $k = 1, \cdots, K$, are identical. This will not happen due to the first property of general natural images. Another extreme case is $\mu' = 0$. This happens only when $\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right) = \mathbf{0}$. This will not happen either as this implies that the simple interpolation $s_{opt}(\mathbf{x}) = \mathbf{U}\mathbf{x}$ produces the optimal HRI.

Therefore, for general natural images μ' cannot be close to either 0 or 1. We also notice that the strong convexity of the squared norm (thinking in 1D, there is large vertical gap between the curve $y = x^2$ and the line segment linking (x_1, x_1^2) and (x_2, x_2^2) when x_1 and x_2 is not close to each other) implies that the scattering of $\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)$, $k = 1, \cdots, K$, will make $||\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right)||^2$ far below the weighted squared norms of $\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)$, $k = 1, \cdots, K$. This implies that although μ' could be a random number between 0 and 1, it should nevertheless *strongly* bias towards 0, i.e., the probability of $0 < \mu' \leq 0.5$ should be *much larger* than that of $0.5 < \mu' < 1$.

For those $\mathbf{x}$ whose μ' is closer to 1, their corresponding $\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)$, $k = 1, \cdots, K$, should be quite cluttered, implying that there is not much choice of adding different high frequency to recover different HRIs. This more likely happens when $\mathbf{x}$ itself is highly textured so that the high frequency is already constrained by the context of the image. Then by the second property of general natural images, such LRIs $\mathbf{x}$ have smaller probability $p_x(\mathbf{x})$ than those requiring smaller μ'.

Due to the bias of μ' and $p_x(\mathbf{x})$, and observing that (28) is actually the average of (29) over $\mathbf{x}$ weighted by $p_x(\mathbf{x}) = \sum\limits_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x})$, we deem that the value $3/4$ is sufficient for μ.[8]

[8] Actually, we believe that 1/2 is already enough due to the strong bias resulting from the convexity of the squared norm. We choose a larger 3/4 just for safety.

To further safeguard the upper bound for the left hand side of (28) and also obtain a concise mathematical formulation in Theorem 2.1, we add an extra nonnegative term to the right hand side of (28), i.e.,

$$\begin{aligned}
& \int_{\mathbf{x}} p_x(\mathbf{x}) \left|\left|\phi_{opt}\left(\mathbf{x}; \tilde{p}_y\right)\right|\right|^2 \mathbf{dx} \\
\leq\ & \frac{3}{4} \int_{\mathbf{x}} \sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \left(\left|\left|\phi_{opt}\left(\mathbf{x}; \tilde{G}_{y;k}\right)\right|\right|^2 + \int_{\mathbf{y}} \left|\left|\mathbf{Vy} - \phi_{opt}(\mathbf{x}; \tilde{G}_{y;k})\right|\right|^2 \tilde{G}_{y;k}(\mathbf{y}|\mathbf{x}) \mathbf{dy} \right) \mathbf{dx} \\
=\ & \frac{3}{4} \int_{\mathbf{x}} \sum_{k=1}^{K} \alpha_k G_{x;k}(\mathbf{x}) \int_{\mathbf{y}} \left|\left|\mathbf{Vy}\right|\right|^2 \tilde{G}_{y;k}(\mathbf{y}|\mathbf{x}) \mathbf{dydx} \\
=\ & \frac{3}{4} \int_{\mathbf{x}} \int_{\mathbf{y}} \left|\left|\mathbf{Vy}\right|\right|^2 p_{x,y}\left(\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \right) \mathbf{dydx}.
\end{aligned} \tag{30}$$

This proves (10).

Proposition 4.3 *(13) is true.*

Proof:

$$\begin{aligned}
& |\hat{\tilde{b}}(N,m) - \tilde{b}(N,m)| \\
=\ & \frac{1}{4} \left| tr\left[(\mathbf{I}-\mathbf{UD}) \left(\hat{\mathbf{\Sigma}}_M - \mathbf{\Sigma} \right) (\mathbf{I}-\mathbf{UD})^t \right] + \left|\left| (\mathbf{I}-\mathbf{UD}) \hat{\bar{\mathbf{h}}}_M \right|\right|^2 - \left|\left| (\mathbf{I}-\mathbf{UD}) \bar{\mathbf{h}} \right|\right|^2 \right| \\
=\ & \frac{1}{4} \left| tr\left[\mathbf{B} \left(\hat{\mathbf{\Sigma}}_M - \mathbf{\Sigma} \right) \right] + \left|\left| (\mathbf{I}-\mathbf{UD})[(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) + \bar{\mathbf{h}}] \right|\right|^2 - \left|\left| (\mathbf{I}-\mathbf{UD}) \bar{\mathbf{h}} \right|\right|^2 \right| \\
=\ & \frac{1}{4} \left| tr\left[\mathbf{B} \left(\hat{\tilde{\mathbf{\Sigma}}}_M - \mathbf{\Sigma} \right) \right] + tr\left[\mathbf{B} \left(\hat{\mathbf{\Sigma}}_M - \hat{\tilde{\mathbf{\Sigma}}}_M \right) \right] \right. \\
& \left. + \left|\left| (\mathbf{I}-\mathbf{UD})(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) \right|\right|^2 + 2[(\mathbf{I}-\mathbf{UD})\bar{\mathbf{h}}]^t (\mathbf{I}-\mathbf{UD})(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) \right| \\
=\ & \frac{1}{4} \left| \sum_{i,j=1}^{mN} B_{ij} \left(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij} \right) - tr\left[\mathbf{B} (\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}})(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}})^t \right] \right. \\
& \left. + \left|\left| (\mathbf{I}-\mathbf{UD})(\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) \right|\right|^2 + 2\left[(\mathbf{I}-\mathbf{UD})^t (\mathbf{I}-\mathbf{UD}) \bar{\mathbf{h}} \right]^t (\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) \right| \\
=\ & \frac{1}{4} \left| \sum_{i,j=1}^{mN} B_{ij} \left(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij} \right) + 2\bar{\mathbf{b}}^t (\hat{\bar{\mathbf{h}}}_M - \bar{\mathbf{h}}) \right| \\
=\ & \frac{1}{4} \left| \sum_{i,j=1}^{mN} B_{ij} \left(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij} \right) + 2 \sum_{i=1}^{mN} \bar{b}_i (\hat{\bar{h}}_{M;i} - \bar{h}_i) \right| \\
\leq\ & \frac{1}{4} \left| \sum_{i,j=1}^{mN} B_{ij} \left(\hat{\tilde{\Sigma}}_{M;ij} - \Sigma_{ij} \right) \right| + \frac{1}{2} \left| \sum_{i=1}^{mN} \bar{b}_i (\hat{\bar{h}}_{M;i} - \bar{h}_i) \right|.
\end{aligned} \tag{31}$$

Proposition 4.4 *(15) is true.*

Proof:

$$
\begin{aligned}
& E\left(\xi^2\right) \\
= \; & E\left(\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\hat{\bar{\Sigma}}_{M;ij}\hat{\bar{\Sigma}}_{M;i'j'}\right) \\
= \; & \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}E\left(\left[\sum_{k=1}^{M}(\hat{h}_{k;i}-\bar{h}_i)(\hat{h}_{k;j}-\bar{h}_j)\right]\left[\sum_{r=1}^{M}(\hat{h}_{r;i'}-\bar{h}_{i'})(\hat{h}_{r;j'}-\bar{h}_{j'})\right]\right) \\
= \; & \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\left\{\sum_{k=1}^{M}E\left[(\hat{h}_{k;i}-\bar{h}_i)(\hat{h}_{k;j}-\bar{h}_j)(\hat{h}_{k;i'}-\bar{h}_{i'})(\hat{h}_{k;j'}-\bar{h}_{j'})\right]\right. \\
& \left.+2E\left[\sum_{1\le k<r\le M}(\hat{h}_{k;i}-\bar{h}_i)(\hat{h}_{k;j}-\bar{h}_j)(\hat{h}_{r;i'}-\bar{h}_{i'})(\hat{h}_{r;j'}-\bar{h}_{j'})\right]\right\} \\
= \; & \frac{1}{M^2}\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\left\{ME\left[(h_i-\bar{h}_i)(h_j-\bar{h}_j)(h_{i'}-\bar{h}_{i'})(h_{j'}-\bar{h}_{j'})\right]\right. \\
& \left.+M(M-1)E\left[(h_i-\bar{h}_i)(h_j-\bar{h}_j)\right]E\left[(h_{i'}-\bar{h}_{i'})(h_{j'}-\bar{h}_{j'})\right]\right\} \\
= \; & \frac{1}{M}\left\{E\left[\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}(h_i-\bar{h}_i)(h_j-\bar{h}_j)(h_{i'}-\bar{h}_{i'})(h_{j'}-\bar{h}_{j'})\right]\right. \\
& \left.+(M-1)\sum_{i,j,i',j'=1}^{mN} B_{ij}B_{i'j'}\Sigma_{ij}\Sigma_{i'j'}\right\} \\
= \; & \frac{1}{M}\left\{E\left[tr^2\left(\mathbf{B}(\mathbf{h}-\bar{\mathbf{h}})(\mathbf{h}-\bar{\mathbf{h}})^t\right)\right]+(M-1)\left[tr(\mathbf{B\Sigma})\right]^2\right\}
\end{aligned}
\tag{32}
$$

Therefore,

$$
\begin{aligned}
& var(\xi) \\
= \; & E\left(\xi^2\right)-\left[E(\xi)\right]^2 \\
= \; & \frac{1}{M}\left\{E\left[\left|\left|(\mathbf{I}-\mathbf{UD})(\mathbf{h}-\bar{\mathbf{h}})\right|\right|^4\right]-\left[tr(\mathbf{B\Sigma})\right]^2\right\}.
\end{aligned}
\tag{33}
$$

Proposition 4.5 *(16) is true.*

Proof:

$$
\begin{aligned}
& var(\eta) \\
= \; & E\left(\left(\sum_{i=1}^{mN}\bar{b}_i(\hat{\bar{h}}_{M;i}-\bar{h}_i)\right)^2\right) \\
= \; & E\left(\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_j(\hat{\bar{h}}_{M;i}-\bar{h}_i)(\hat{\bar{h}}_{M;j}-\bar{h}_j)\right) \\
= \; & \frac{1}{M^2}\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_jE\left(\sum_{k=1}^{M}(\hat{h}_{k;i}-\bar{h}_i)\sum_{r=1}^{M}(\hat{h}_{r;j}-\bar{h}_j)\right) \\
= \; & \frac{1}{M^2}\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_j\left[\sum_{k=1}^{M}E\left((\hat{h}_{k;i}-\bar{h}_i)(\hat{h}_{k;j}-\bar{h}_j)\right)+\sum_{1\le k<r\le M}E\left((\hat{h}_{k;i}-\bar{h}_i)(\hat{h}_{r;j}-\bar{h}_j)\right)\right] \\
= \; & \frac{1}{M^2}\sum_{i,j=1}^{mN}\bar{b}_i\bar{b}_j(M\Sigma_{ij}) \\
= \; & \frac{1}{M}\bar{\mathbf{b}}^t\mathbf{\Sigma}\bar{\mathbf{b}}.
\end{aligned}
\tag{34}
$$