Image Hallucination Using Neighbor Embedding over Visual Primitive Manifolds

Wei Fan & Dit-Yan Yeung
Department of Computer Science and Engineering,
Hong Kong University of Science and Technology
{fwkevin,dyyeung}@cse.ust.hk

Abstract

In this paper, we propose a novel learning-based method for image hallucination, with image super-resolution being a specific application that we focus on here. Given a low-resolution image, its underlying higher-resolution details are synthesized based on a set of training images. In order to build a compact yet descriptive training set, we investigate the characteristic local structures contained in large volumes of small image patches. Inspired by recent progress in manifold learning research, we take the assumption that small image patches in the low-resolution and high-resolution images form manifolds with similar local geometry in the corresponding image feature spaces. This assumption leads to a super-resolution approach which reconstructs the feature vector corresponding to an image patch by its neighbors in the feature space. In addition, the residual errors associated with the reconstructed image patches are also estimated to compensate for the information loss in the local averaging process. Experimental results show that our hallucination method can synthesize higher-quality images compared with other methods.

1. Introduction

Image super-resolution refers to the process by which a higher-resolution enhanced image is synthesized from one or more low-resolution images. It finds a number of real-world applications, which include restoring historic photographs, enlarging "thumbnail" images on web pages, and image-based rendering for high-quality display purposes. Practical super-resolution methods may make use of a single still image or a sequence of consecutive video frames with sub-pixel translation for synthesizing a higher-resolution image. In this paper, we focus on the problem of single-image "hallucination" with the goal of inferring some high-resolution details missing in the original image that cannot be achieved by simple sharpening.

In recent years, there has been a good deal of research into learning-based approaches for image hallucination as well as other related low-level vision problems [2, 4, 5, 6, 7, 11]. These learning-based methods share the common characteristic of using a training set of image (observation) and scene (state) pairs to build a co-occurrence model. With the learnt model, one can then predict the missing details in the observed input image by "borrowing" information from some similar examples in the training set.

Due to the contiguous nature of objects and surfaces in visual environments, images from natural scenes only constitute a minuscule fraction of the space of all possible images [9]. However, it is difficult, if not totally impossible, to precisely model the probability distribution of natural images for the generic super-resolution task. Instead, what we can do is to study the distribution of small image patches and see what kinds of local image structures (e.g., edges or corners) are likely to occur in the image. These local image patches, from either the low-resolution or high-resolution image, are the building blocks of our super-resolution or image hallucination approach. They are expected to lie along a continuous nonlinear manifold embedded in a high-dimensional image space. Inspired by a well-known manifold learning method called locally linear embedding (LLE) [10], we assume that small image patches in the low-resolution and high-resolution images form manifolds with similar local geometry in the corresponding image spaces. This assumption generally holds as long as the image patches are associated with image primitives and the feature descriptions for the two corresponding images are both isometric. Our contribution is to devise an effective method for generic image hallucination using locally linear fitting and a learnt image primitive model.

A flowchart of our image hallucination approach is shown in Figure 1. We highlight the major steps of our approach here. In the learning phase, large volumes of image primitive patches are extracted from both the low-resolution and high-resolution images used for training. A training set is constructed by analyzing the local neighborhood relation-

1-4244-1180-7/07/$25.00 ©2007 IEEE

ships between the low-resolution and high-resolution image patches and keeping only the isometric regions along the two manifolds. During the synthesis phase, a low-resolution image is presented to the system. Each element in the target high-resolution image comes from an optimal linear reconstruction by its nearest neighbors in the training set. Moreover, the residual errors associated with the reconstructed image patches are also estimated to compensate for the information loss in the local averaging process. Finally, we enforce the local compatibility and smoothness constraints between patches in the target high-resolution image through overlapping.

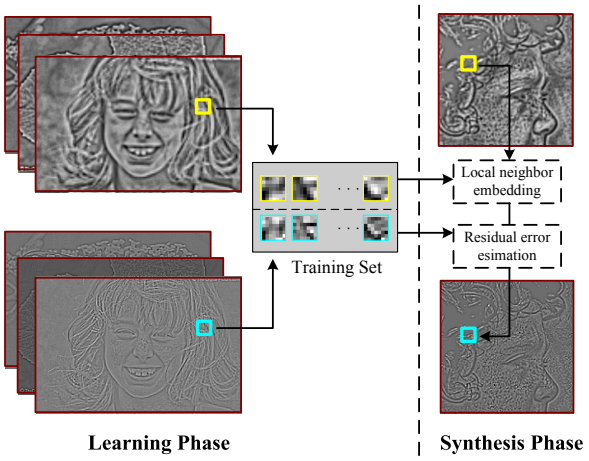

Figure 1. Flowchart of our image hallucination approach.

The rest of this paper is organized as follows. In Section 2, we give a brief overview of the background and some related work. The problem setting and detailed algorithm are described in Sections 3 and 4, respectively. Section 5 presents some experimental results, demonstrating the effectiveness and efficiency of our method. Finally, we conclude our paper in Section 6.

2. Related work

Image super-resolution is intrinsically an ill-posed problem since, theoretically, many high-resolution images can give rise to the same low-resolution image through some operations such as smoothing and subsampling. Traditional pixel interpolation methods with smoothness priors, such as pixel replication and cubic-spline interpolation, introduce artifacts and blurred edges. Reconstruction-based methods [8] aim to restore the lost image details by requiring the down-sampled version of the high-resolution reconstructed image to be as close to the original low-resolution image as possible. However, in the absence of additional information, these generic constraints are not every effective for synthesizing perceptually plausible images which are of higher quality than the original low-resolution images.

Over the past few years, learning-based approaches have produced compelling results for various low-level vision tasks, including image hallucination [2, 4, 5, 7, 11], image analogy [6], and texture synthesis [3]. Despite some implementation-level differences, these algorithms are all similar in spirit. In the learning stage, they learn the underlying scene details that correspond to different image regions observed in the input. During the inference stage, they use those learned relationships to predict missing details in another image, which is the target high-resolution image for super-resolution problems. Thus each element in the target image comes from only one "best" example in the training set. The neighbor embedding method proposed by [2] introduces a more general way of using the training examples. In their method, multiple training examples can contribute simultaneously to the generation of each image patch in the high-resolution image. The underlying assumption of the method is that small image patches in the low-resolution and high-resolution images form manifolds with similar local geometry in the two corresponding image spaces. However, they use uniformly sampled image patches from several training images to build a medium-sized training set, which may not satisfy the manifold assumption and hence may not lead to good generalization.

Motivated by the work of [11] in applying primal sketch priors for image hallucination, we conjecture that the image patches associated with the image primitives (e.g., edges, corners, and blobs, similar to the primal sketches in [11]) are of greater significance for building a good training set. This conjecture will be verified empirically in Section 4 through statistical analysis, showing that these image patches preserve well the local isometric relationships between the manifolds for the low-resolution and high-resolution image patches.

Our method also benefits from the face hallucination work of [7], in which the relationships between the low-resolution and high-resolution residues are learnt by coupled PCA to refine the final hallucinated face image. In our case, the residual error vectors of neighboring examples are consistent, since they are expected to lie on a subspace perpendicular to the linear tangent plane of the curved manifold. Adding a simple average of these residues is sufficient for error compensation.

In the next two sections, we will formulate the problem more precisely and then present details of different components of our method.

3. Problem setting

The single-image super-resolution problem that we want to solve can be formulated as follows. Given a low-resolution image L_t as input, we estimate the target high-resolution image H_t with the help of a training set of one or

more low-resolution images L_s and the corresponding high-resolution images H_s. We represent each low-resolution or high-resolution image as a set of small overlapping image patches.

Ideally, each patch generated for the high-resolution image H_t should not only be related appropriately to the corresponding patch in the low-resolution image L_t, but should also preserve some inter-patch relationships with adjacent patches in H_t. The former determines the accuracy while the latter determines the local compatibility and smoothness of the high-resolution image. To satisfy these requirements as much as possible, our method has the following properties: (a) Each patch in H_t is associated with multiple patch transformations learned from the training set. (b) Local relationships between patches in L_t should be preserved in H_t. (c) Neighboring patches in H_t are constrained through overlapping to enforce local compatibility and smoothness.

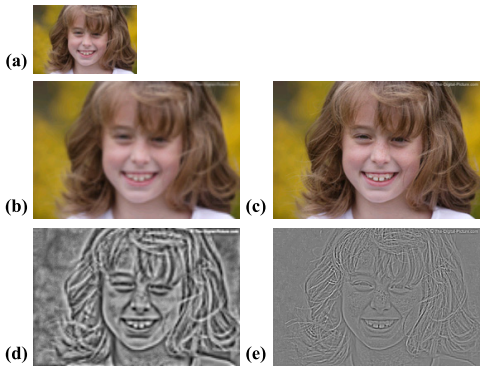

Figure 2. Image preprocessing steps. (a) low-resolution image; (b) initial interpolation of (a) to a higher resolution; (c) original high-resolution image; (d) band-pass filtered and contrast normalized version of (b); (e) high-pass filtered and contrast normalized version of (c).

4. Details of our method

4.1. Preprocessing

In the preprocessing step, a high-resolution natural image H (Figure 2(c)) is blurred and sub-sampled to generate a corresponding low-resolution image L (Figure 2(a)). Applying an initial enhancement through bilinear interpolation to L, we obtain an image H^l (Figure 2(b)) which has the same size as H but lacks the high-resolution details. In the training set, we only need to store the differences between H and H^l (Figure 2(e)), which correspond to the missing high-frequency components caused by the image degradation process. Through band-pass filtering, we further decompose each interpolated image H^l into the sum of two images containing the medium and low spatial frequencies, respectively. Following the assumption in [5] that the highest spatial-frequency components of the low-resolution image are most important for predicting the extra details of H, we only store the example patches from the medium frequency layer (Figure 2(d)). Finally, to achieve good generalization, the high- and medium-frequency image pairs are contrast normalized by a local measure of energy in the image. We undo this normalization step later when we reconstruct the high-resolution image. The final output is the sum of the interpolated low-resolution image and the high-frequency predictions.

4.2. Image primitives for training

An essential factor attributing to the success of learning-based approaches is how to construct a good training set, which is descriptive enough in giving useful information about the image-scene relationships and is also compact enough for computational efficiency and good generalization. The patches extracted from the preprocessed images can be regarded as points in a vector space with each dimension corresponding to one pixel in the patch. As natural images contain characteristic statistical regularities, we expect these feature points to lie on a continuous nonlinear manifold. Different regions of the manifold may correspond to characteristic image primitives such as edges, corners, blobs, etc. When building our training set, we put emphasis on the patches associated with these image primitives for two main reasons. First, the missing high-frequency details to be estimated are densely distributed over the regions of image primitives, as shown in Figure 3(b) and 3(c). Focusing on these regions can lead to significant speedup as fewer patches need to be transformed. Second, we believe the local neighborhood relationships between low-resolution and high-resolution primitive patches in the two feature spaces are more consistent than those between general image patches. This is supported by our experimental investigation to be reported in Section 4.

The primitive patches are extracted by convolving the interpolated low-resolution image H^l with a bank of maximum response (MR) filters (Figure 3(a)) [13]. The filter bank consists of a Gaussian kernel and a Laplacian of Gaussian kernel, which are arranged in three scales and six orientations each. To achieve scale invariance, the outputs are "collapsed" by recording only the maximum filter response across all scales. This reduces the number of responses for each pixel from 36 (six orientations at three scales for each of two oriented filters) to 12 (six orientations for each of two filters). Figure 3(c) depicts the magnitude map of the filtered responses for a low-resolution image. Compared with the high-frequency difference image shown in Figure 3(b), we can clearly see the consistent relationships between them.

To sum up, each example in the training set is in the form

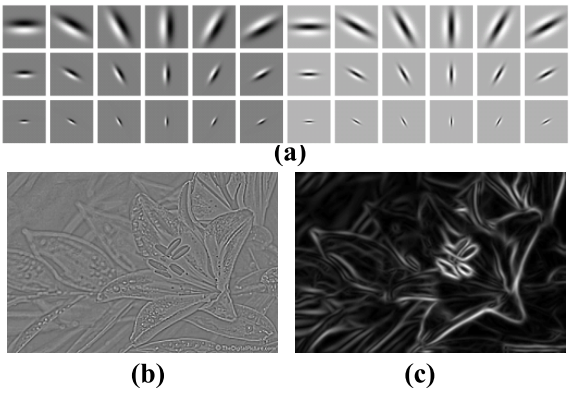

Figure 3. (a) Filter bank used for primitive extraction; (b) high-frequency difference image to be estimated; (c) magnitude map of the filtered responses for the low-resolution image.

of a pair of primitive patches. These pairs capture the statistical relationships that we are interested in. We represent each image primitive by a 7×7 image patch, which is selected from the region in Figure 3(c) with high magnitude value or energy level.

4.3. Local neighbor embedding

The manifold structure of image patches characterizes the smooth variation corresponding to some regular transformations in natural images. For instance, we can expect that the manifold coordinates of the edge structure correspond to its orientation, translation, and blurring variations. In super-resolution problems, we further assume that manifolds of the small patches in the low-resolution and high-resolution images bear similar local geometry in the two spaces. This assumption holds as long as the patches are associated with image primitives and the two feature descriptions are isometric.

In recent years, manifold learning (or nonlinear dimensionality reduction) methods have emerged as powerful tools for discovering a "faithful" low-dimensional representation of the original data embedded in some high-dimensional observation space [1, 10, 12, 14]. The main computations for these methods are based on tractable, polynomial-time optimizations, such as shortest path problems, least squares fits, semidefinite programming, and matrix diagonalization. Our super-resolution method to be described below has been inspired by the LLE algorithm [10]. Its key idea is that the local geometry in the neighborhood of each data point can be characterized by linear coefficients that reconstruct the data point from its neighbors.

For convenience, we use l_s^p , h_s^p , l_t^q and h_t^q to denote the feature vectors as well as the corresponding low- and high-resolution image patches, and L_s, H_s, L_t and H_t to denote the sets of feature vectors as well as the corresponding images. The neighbor embedding algorithm of our method can be summarized as follows:

Algorithm - Local neighbor embedding

Input: low-resolution test patches $L_t = \{l_t^1, l_t^2 \cdots, l_t^{n_t}\}$
Output: high-resolution patches $H_t = \{h_t^1, h_t^2 \cdots, h_t^{n_t}\}$
Begin

1. For each patch l_t^q in image L_t:
   (a) Find the set N_q of K nearest neighbors in L_s.
   (b) Compute the reconstruction weights of the neighbors that minimize the error of reconstructing l_t^q.
   (c) Compute the initial high-resolution embedding h_t^q using the appropriate high-resolution features of the K nearest neighbors and the reconstruction weights.
   (d) Estimate the high-resolution residual error vector e_t^q using the average residual error vector of the K nearest neighbors in N_q, and update h_t^q with $h_t^q + e_t^q$.
2. Construct the target high-resolution image H_t by enforcing the local compatibility and smoothness constraints between adjacent patches obtained in step 1(d).

End

We implement step 1(a) by using Euclidean distance to define neighborhood. Based on the K nearest neighbors identified, step 1(b) seeks to find the best reconstruction weights for each patch l_t^q in L_t. Optimality is achieved by minimizing the local reconstruction error for l_t^q

$$\varepsilon^q = \|l_t^q - \sum_{l_s^p \in N_q} \omega_{qp} l_s^p\|^2 \tag{1}$$

which is the squared distance between l_t^q and its reconstruction, subject to the constraints $\sum_{l_s^p \in N_q} \omega_{qp} = 1$ and $\omega_{qp} = 0$ for any $l_s^p \notin N_q$. Minimizing ε^q subject to the constraints is a constrained least squares problem. We define a local Gram matrix G_q for l_t^q as

$$G_q = (l_t^q 1^T - L)^T (l_t^q 1^T - L) \tag{2}$$

where 1 is a column vector of ones and L is a $D \times K$ matrix with its columns being the neighbors of l_t^q . Moreover, we group the weights of the neighbors to form a K-dimensional weight vector w_q by reordering the subscript p of each weight ω_{qp}. The constrained least squares problem has the following closed-form solution:

$$w_q = \frac{G_q^{-1} 1}{1^T G_q^{-1} 1} \tag{3}$$

Instead of inverting G_q, we apply a more efficient method by solving the linear system of equations $G_q w_q = 1$ and

then normalizing the weights so that $\sum_{l_s^p \in N_q} \omega_{qp} = 1$. After repeating steps 1(a) and 1(b) for all N_t patches in L_t, the reconstruction weights obtained form a weight matrix $W = [\omega_{qp}]_{N_t \times N_s}$.

Step 1(c) computes the initial value of h_t^q based on W:

$$h_t^q = \sum_{l_s^p \in N_q} \omega_{qp} h_s^p \tag{4}$$

Step 1(d) estimates the high-resolution residual error vector e_t^q of the linear reconstruction

$$\begin{aligned} e_t^q &= \frac{1}{K} \sum_{l_s^p \in N_q} e_s^p \\ &= \frac{1}{K} \sum_{l_s^p \in N_q} (h_s^p - \sum_{h_s^r \in N_p} \omega_{rp} h_s^r) \end{aligned} \tag{5}$$

where e_s^p is the residual error of the neighboring patch $l_s^p \in N_q$. It is calculated in the learning phase and stored together with its associated l_s^p. We use the average of e_s^p to estimate e_t^q, with the assumption that these neighboring residual error vectors are in approximately the same direction which is perpendicular to the tangent plane at h_t^q.

In step 2, we use a simple method to enforce inter-patch relationships by averaging the feature values in regions where adjacent patches overlap. Other more sophisticated methods may also be used.

4.4. Training set revisited

An essential assumption of our local neighbor embedding method is that the linear reconstruction weight of l_s^p and that of its corresponding h_s^p should be approximately the same in the two corresponding image spaces. In this subsection, we evaluate this assumption by computing the standard linear correlation coefficient $R(w_l, w_h)$ between two groups of weight vectors, w_l and w_h. Evaluation is performed based on two settings, either using a randomly sampled patch set or using an image primitive patch set. Each data set contains around 25,000 pairs of low-resolution and high-resolution patches. Figure 4 shows the histograms of the correlation coefficient under the two settings, with its value ranging from -1 to 1. From this comparison, we can clearly see that the local neighborhood relationships between low-resolution and high-resolution primitive patches are more consistent than those between general image patches.

However, since the image primitives are extracted using a hard threshold on the filtered low-resolution image, it is inevitable that some "noisy" patches will be included in the training set, resulting in the low correlation part of Figure 4(b). In our experiment, we only keep those primitive patches and their K nearest neighbors if their correlation coefficients are greater than 0.7. These patches form a compact training set which can characterize well the manifold structure of natural images.

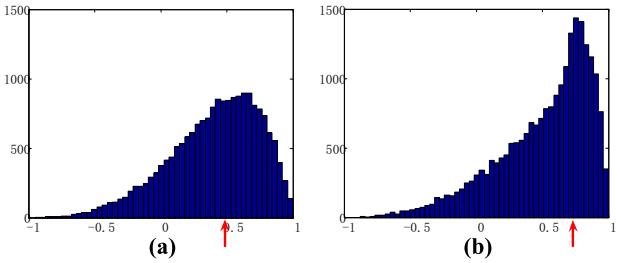

Figure 4. Histograms of correlation coefficient $R(w_l, w_h)$ between reconstruction weight vectors for l_s^p and h_s^p in the two image spaces under two settings: (a) randomly sampled image patches; (b) image primitive patches. The red arrows indicate the median values.

5. Experiments

We build training sets for the super-resolution algorithm from band-pass and high-pass pairs taken from a set of training images (see Figure 5). All the eight representative natural images were downloaded from a public web site.[1] They were taken with a Canon EOS D60 digital camera with a resolution of 500×433 pixels. About 400,000 primitive examples have been extracted from these training images.

Figure 5. Eight training images (downloaded from http://www.the-digital-picture.com/Gallery/) used in our experiments.

Our method has only three parameters to determine. The first parameter is the number of nearest neighbors K for neighbor embedding. Our experiments show that the super-resolution result is not very sensitive to the choice of K. We set K to 5 for all our experiments. The second and third parameters are the patch size and the degree of overlap between adjacent patches. For both the low-resolution and high-resolution images, we use 7×7 pixel patches and let the overlap between adjacent patches be 4 pixels. The corresponding low-resolution and high-resolution image patches are properly aligned by their geometrical centers in the image plane. Principal component analysis (PCA) is performed on the low-resolution patch set to reduce its dimensionality to 15, which covers more than 98% of the total

[1] http://www.the-digital-picture.com/Gallery/

variance. The high-resolution feature vector is represented by concatenating all 7×7 pixels in the patch, since we cannot find an 'elbow' at which the eigenvalue curve ceases to decrease significantly with added dimensions. Note that we perform hallucination on the image intensity only because humans are more sensitive to the brightness information. The color channels are simply interpolated by a bilinear function.

We compare our approach with bicubic interpolation and the neighbor embedding method of Chang *et al.* [2] on different super-resolution examples, all with a magnification factor of 3 (Figure 6). When implementing the method in [2], we use uniformly sampled patches in the training images to build a training set with the same size as ours. It is clear to see that bicubic interpolation gives the smoothest result. Chang *et al.*'s method gives better result for some details in the images. On the other hand, sharper and smoother contours are hallucinated by our approach (e.g., see the edges of the leaf in the first example).[2]

We also calculate the RMS error between the super-resolution image generated and the ground-truth image as the number of nearest neighbors K varies over a range. Figure 7 shows the results for the four examples discussed above. As we can see, the RMS error attains its lowest value when K is between 4 and 6, showing that using multiple nearest neighbors (as opposed to only one nearest neighbor as in the existing methods) does give improved results.

6. Conclusion

In this paper, we have proposed a learning-based method for image hallucination based on local neighbor embedding of the image primitive manifold constructed from an input image. In particular, we study image hallucination in the context of the single-image super-resolution problem. A compact yet descriptive training set is constructed from characteristic regions in images where the manifold assumption holds well. Compared with other generic single-image super-resolution methods, our method can synthesize higher-quality images. In our future work, we will apply a similar approach to other image hallucination problems.

Acknowledgments

This research has been supported by research grants HKUST621305 and N-HKUST602/05 from the Research Grants Council (RGC) of Hong Kong and the National Natural Science Foundation of China (NSFC).

[2] Readers are recommended to see the enlarged electronic version of the figure.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003. 4

[2] H. Chang, D. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 275–282, 2004. 1, 2, 6

[3] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038, 1999. 2

[4] W. Freeman, T. Jones, and E. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):55–65, 2002. 1, 2

[5] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000. 1, 2, 3

[6] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. Salesin. Image analogies. In *SIGGRAPH*, 2001. 1, 2

[7] W. Liu, D. Lin, and X. Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 478–484, 2005. 1, 2

[8] B. S. Morse and D. Schwartzwald. Image magnification using level-set reconstruction. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 275–282, 2001. 2

[9] B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339, 1996. 1

[10] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000. 1, 4

[11] J. Sun, N. Zheng, H. Tao, and H. Shum. Image hallucination with primal sketch priors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 729–736, 2003. 1, 2

[12] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000. 4

[13] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 67(1-2):61–81, 2005. 3

[14] K. Weinberger and L. Saul. Unsupervised learning of image manifolds by semidefinite programming. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 275–282, 2004. 4

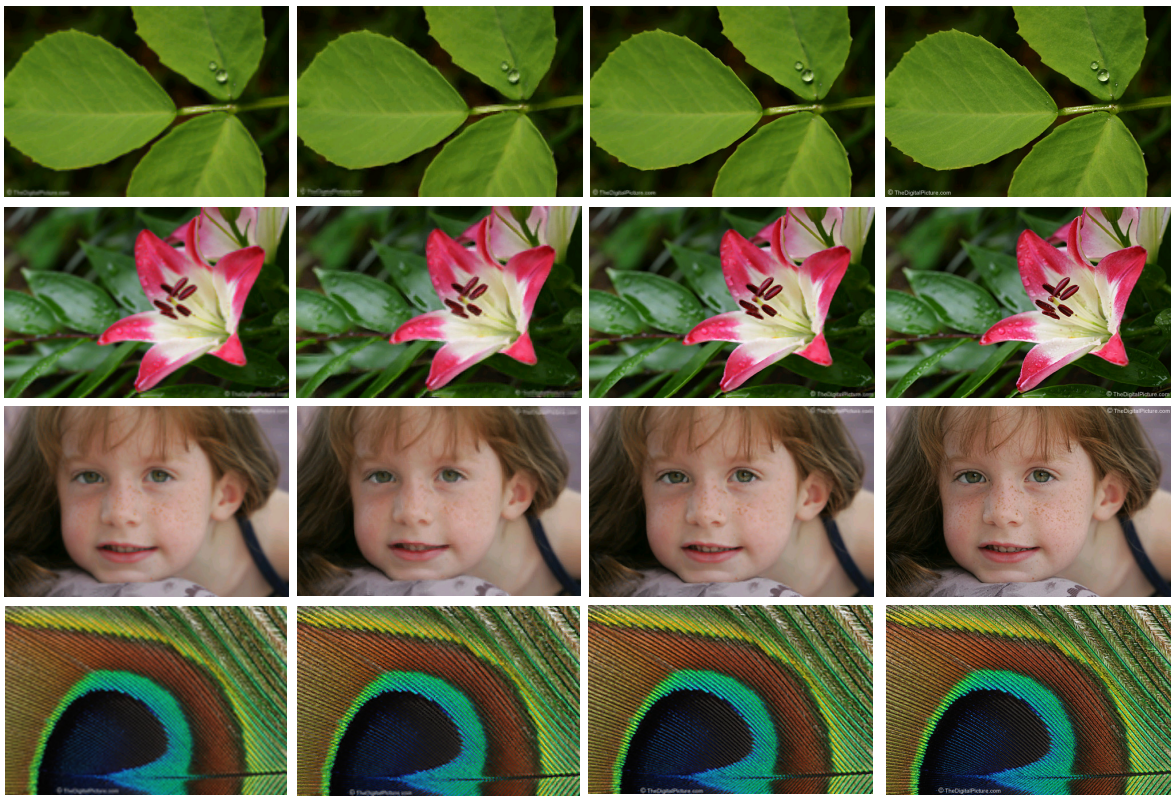

Figure 6. Test images magnified by three times using (a) bicubic interpolation, (b) Chang *et al.*'s method, and (c) our approach; (d) original high-resolution image.

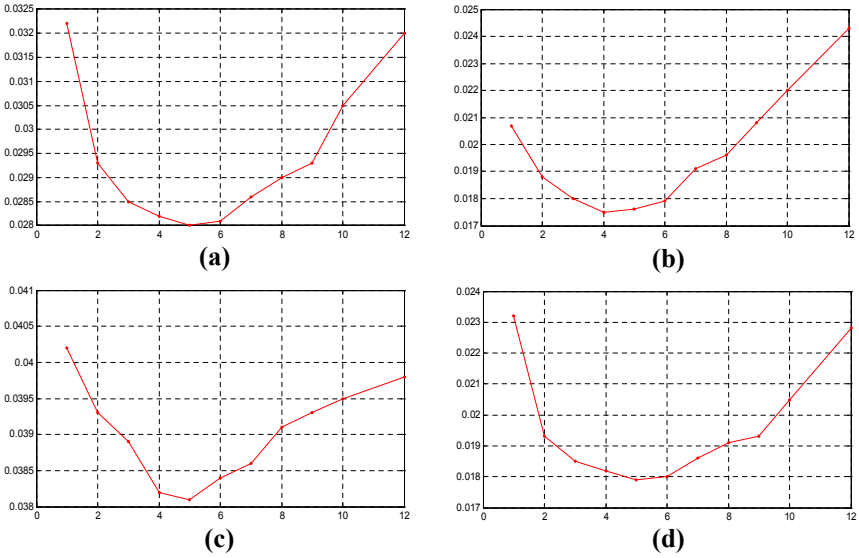

Figure 7. RMS error between the super-resolution image generated and the ground-truth image as a function of the number of nearest neighbors used: (a) leaf image; (b) flower image; (c) face image; (d) pattern image.