Image Hallucination with Primal Sketch Priors

Jian Sun[†*] Nan-Ning Zheng[†] Hai Tao[‡] Heung-Yeung Shum[††]

The Institute of AI and Robotics[†]
Xi'an Jiaotong University
Xi'an China 710049
{*sj, nnzheng*}*@aiar.xjtu.edu.cn*

Department of C.E.[‡]
Univ. of California, SC.
Santa Cruz, CA 95064
tao@soe.ucsc.edu

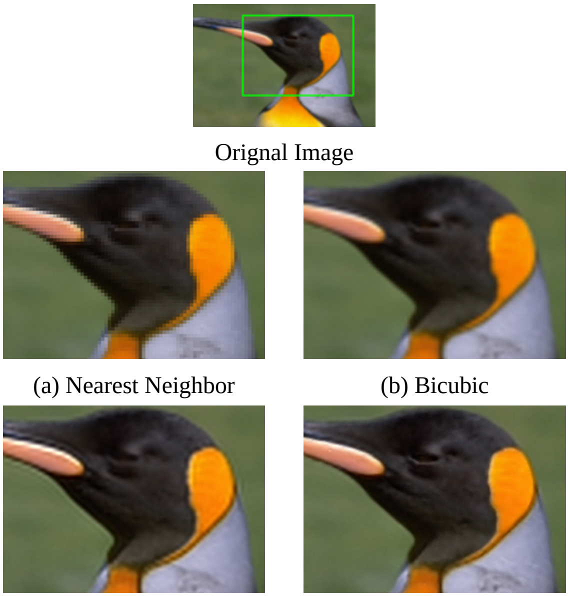
Microsoft Research Asia[††]
No.49, Zhichun Road
Beijing China 100080
hshum@microsoft.com

Abstract

In this paper, we propose a Bayesian approach to image hallucination. Given a generic low resolution image, we hallucinate a high resolution image using a set of training images. Our work is inspired by recent progress on natural image statistics that the priors of image primitives can be well represented by examples. Specifically, primal sketch priors (e.g., edges, ridges and corners) are constructed and used to enhance the quality of the hallucinated high resolution image. Moreover, a contour smoothness constraint enforces consistency of primitives in the hallucinated image by a Markov-chain based inference algorithm. A reconstruction constraint is also applied to further improve the quality of the hallucinated image. Experiments demonstrate that our approach can hallucinate high quality super-resolution images.

1. Introduction

Image super-resolution has become an active research topic in computer vision lately. Super-resolution techniques have many applications ranging from video quality enhancement to image compression. Most super-resolution techniques require multiple low resolution images to be aligned in sub-pixel accuracy. In this paper, however, we focus on image super-resolution from a single image.



Orignal Image

(a) Nearest Neighbor

(b) Bicubic

(c) Backprojection

(d) Our Approach

Figure 1. Comparison of different super-resolution techniques. Top: the original image. (a) Nearest Neighbor (simply copying pixels), (b) Bicubic (functional interpolation), (c) Backporjection (reconstruction-based) and (d) Image hallucination (learning-based approach).

Clearly, single image super-resolution is an under-constrained problem because many high resolution images can produce the same low resolution image. Previous work on single image super-resolution can be categorized into three classes: functional interpolation, reconstruction-based and learning-based. Functional interpolation methods often blur the discontinuities and do not satisfy the reconstruction constraint. Under the reconstruct constraint, the downsampled high resolution reconstructed image should be as close as possible to the original low resolution image. Figures 1(a) and (b) show the results of nearest neighbor interpolation and bicubic interpolation of a low resolution image respectively. Edge-based interpolation methods [2, 15] have also been proposed. Reconstruction-based methods [6, 9] satisfy the reconstruction constraint but cannot guarantee contour smoothness. Figure 1 (c) shows the result of a reconstruction-based approach using backprojection [6]. Some "jaggy" and "ringing" artifacts are clearly visible along the edges. In this paper, we propose a learning-

[*] This work was performed while Jian Sun was visiting Microsoft Research Asia.

1063-6919/03 $17.00 © 2003 IEEE

based approach. To construct the super-resolution image, we "steal" high frequency details that do not exist in the low resolution image from a number of training images. A good quality super-resolution image reconstructed using our approach is shown in Figure 1(d). This is, in spirit, similar to face hallucination [3] and other related low-level learning work [4, 5, 7].

This is why we call our approach "image hallucination". Unlike "face hallucination" [3], however, our approach works for generic images. Instead of assuming generic smoothness priors that are used in interpolation approaches, learning-based approaches choose a recognition-based prior based on a set of recognition decisions on the low resolution image I_L. For instance, the input I_L can be divided into a number of partitions where each partition is classified into a subclass and is associated with a subclass prior. If the integration of subclass priors is more powerful than a generic smoothness prior, the learning-based approach can outperform the other approaches. Impressive results have been obtained in domain-specific applications (e.g., face, text [3, 7]). However, to do "image hallucination" for generic images, what are the basic recognition elements in the generic image? How to learn the prior for each subclass?

In this paper, we propose primal sketches [8] as the natural basic recognition elements to get a recognition-based prior for generic images. Firstly, the low resolution image is interpolated as the low frequency part of a high resolution image. This low frequency image is then decomposed into a low frequency primitive layer and a non-primitive layer. Each primitive in the primitive layer is recognized as part of a subclass, e.g. an edge, a ridge or a corner at different orientations and scales. For each subclass, its training data (i.e., high frequency and low frequency primitive pairs) are collected from a set of natural images. Secondly, for the input low resolution image, a set of candidate high frequency primitives are selected from the training data based on low frequency primitives. From this set of candidates, a consistent high frequency primitive layer is inferred using a Markov chain model. The super-resolution image is obtained by combining the high frequency primitive layer with the low frequency image, followed by a backprojection algorithm enforcing the reconstruction constraint.

The performance of the learning-based approach is dependent on the priors we use. Specifically, using training samples, the priors are represented by a set of examples in a non-parametric way. The generalization of training data is the key to do hallucination for the generic image. Whether or not sample in a generic image can find a good match in the training data determines how successful a learning based approach can be. However, it is hard to learn a good prior for an arbitrary image patch in natural images. It is demonstrated by the statistical analysis on an empirical data set in Section 3. Fortunately, the statistical analysis in Section 3 also shows primal sketch priors can be learned well from a number of examples that we can computationally afford today. Therefore, we propose to do image hallucination with primal sketch priors.

Our work on image hallucination is also motivated by the recent progress on natural image statistics [1, 14]. For example, it is shown in [1] that the intrinsic dimensionality of image primitives is very low. Low dimensionality makes it possible to represent well all the image primitives in natural images by a small number of examples. These inspire us to use the image primitive as the basis recognition element to take advantage of the strong structure information in generic images.

The rest of this paper is organized as follows. In Section 2, we give the overview of our image hallucination. The details of algorithm are described in Sections 3 and 4. The experimental results shown in Section 5 demonstrate that our model is effective and efficient. We conclude the paper in Section 6.

2. Overview

An overview of our approach is shown in Figure 2. The approach consists of three steps. In step 1, a low frequency image I_H^l is interpolated from the low resolution image I_L. In step 2, a high frequency primitive layer I_H^p is hallucinated or inferred from I_H^l based on the primal sketch priors. In step 3, we enforce the reconstruction constraint to get the final high resolution image I_H.

In our approach, we hallucinate the lost high frequency information of primitives (e.g., edges) in the image, but not the non-primitive parts of the image. The key observation in this paper is that we hallucinate only the primitive part of the image, because we can effectively learn the priors of primitives - "primal sketch priors", but not the priors of non-primitives. The MAP of primitive layer I_H^p is hallucinated from I_H^l and prior $p(I_H^p)$,

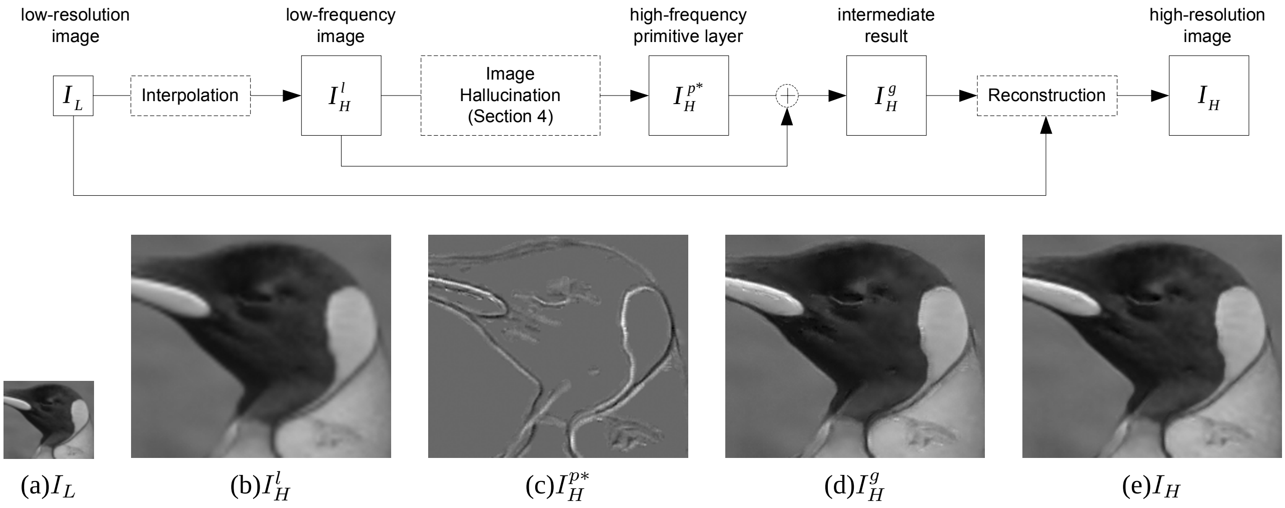
$$\begin{aligned} I_H^{p*} &= \arg\max p(I_H^p|I_H^l) \\ &= \arg\max p(I_H^l|I_H^p)p(I_H^p). \end{aligned} \tag{1}$$

Section 3 shows the details about how to learn the primal sketch priors $p(I_H^p)$. And how to hallucinate I_H^{p*} is presented in Section 4.

After getting hallucinated primitive layer I_H^p, we can obtain an intermediate result I_H^g that does not satisfy the reconstruction constraint in general. Backprojection [6] is an iterative gradient-based minimization method to minimize the reconstruction error:

$$I_H^{t+1} = I_H^t + (((I_H^t * h) \downarrow s - I_L) \uparrow s) * p \tag{2}$$

where p is a "backprojection" filter. In our case, the final solution is obtained simply by using I_H^g as the starting point.



(a)I_L (b)I_H^l (c)I_H^{p*} (d)I_H^g (e)I_H

Figure 2. The overview of our approach. I_L is the low resolution image. I_H^l is the bicubic interpolation of I_L. The key of our approach is that a high frequency primitive layer I_H^{p*} is hallucinated based on the primal sketch prior $p(I_H^p)$ provided by the primitives training data. The final high resolution image I_H is obtained from the intermediate result I_H^g by enforcing the reconstruction constraint.

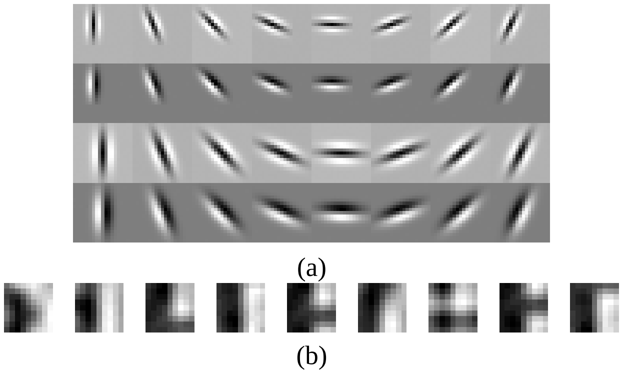


Figure 3. The filter bank used for primitives extraction (a) and typical primitives extracted (b).

The final high resolution image I_H is shown in Figure 2 (e). Noise and artifacts are significantly reduced with the reconstruction constraint.

3 Primal Sketch Priors

In this section, we describe how to learn the primal sketch priors. Furthermore, we study why the primitives can be effectively represented by samples but the non-primitives cannot. This statistical analysis sheds light on the difficulty of generic image super-resolution using learning-based approaches and sample images.

3.1 Primal Sketch

We take an example-based approach to learn two things from training data. One is the primal sketch prior $p(I_H^p)$. This prior is actually represented by a collection of examples in a non-parametric form. The other is the statistical relationship between low frequency primitives (interpolation of low resolution primitive) and high frequency primitive (difference between high resolution primitive and low frequency primitives). Each example consist of a pair of primitives. These pairs capture the statistical relationship in which we are interested.

We represent each image primitive by a 9x9 image patch. The primitives are extracted by orientation energy [13],

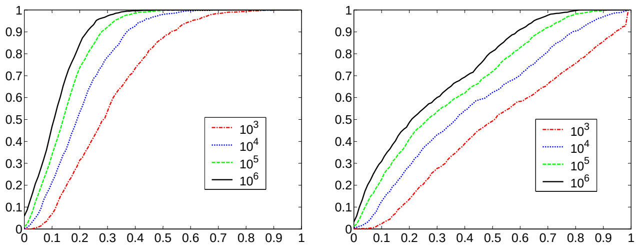
$$OE_{\sigma,\theta} \quad = \quad (I * f_{\sigma,\theta}^{odd})^2 + (I * f_{\sigma,\theta}^{even})^2$$

where $f_{\sigma,\theta}^{odd}$ and $f_{\sigma,\theta}^{even}$ are the first and second Gaussian derivative filters at scale σ and orientation θ. These filters consist of a filter bank shown in Figure 3 (a) (2 scales and 16 orientations). We extract the patches along the contours. The primitives such as step-edge, ridge, corners, T-junction and terminations are extracted. Typical patches in a subclass are shown in Figure 3 (b).

From Pattern theory [11], the observed image primitive $\mathbf{x}$ is generated by the *latent pattern* $\mathbf{z}$ underlying some global geometric and photometric transformations, such as translation, scaling, orientation and lighting. The generative model of image primitive B can be defined as,

$$B \quad = \quad c \cdot \mathbf{G}_t \mathbf{G}_o \mathbf{G}_s \mathbf{z} + d \tag{3}$$

where c is contrast, d is DC bias for lighting, and $\mathbf{G}_t$, $\mathbf{G}_s$



(a)

(b)

Figure 4. The ROC curves of primitive training data (a) and component training data (b) at different sizes. X-axis is match error and Y-axis is hit-rate.

and $\mathbf{G}_o$ are translation, scaling and orientation transformation matrices respectively. The local transformations such as subpixel location, curvature and local intensity variations, are absorbed into $\mathbf{z}$.

To reduce the dimensionality of primitives, we follow the same assumption [4] that the statistical relationship between low frequency primitives and high frequency primitives is independent of some transformations including contrast, DC bias and translation. Let B^l be a low frequency primitive and B^h a corresponding high frequency primitive. We normalize B^l to get a normalized low frequency primitive $\widehat{B}^l$,

$$\widehat{B}^l = \frac{1}{c^l} \cdot \mathbf{G}_t^{-1}(B^l - d^l) = \mathbf{G}_o \mathbf{G}_s \mathbf{z}^l \qquad (4)$$

where $\mathbf{G}_t$ is approximated by $\mathbf{I}$ because the center of each primitive we extract is on the contour. DC bias d is estimated by the mean $E[B]$. The contrast c is estimated by $E[|B - E[B]|]$.

Each example consists of a normalized low frequency primitive $\widehat{B}^l$, its contrast c^l and a high frequency primitive B^h. The primal sketch priors are represented by all the examples in a non-parametric way.

3.2 Why Primal sketch?

Why do we choose only the primitive for hallucination? The answer lies in the low dimensionality of the primitive manifold. On the other hand, the dimensionality of the non-primitive manifold is too high to be represented well by the number of examples we can afford computationally. We demonstrate this key observation by statistical analysis on an empirical data set. Luckily, humans are more sensitive to the high contrast intensity changes [8] because strong stimuli are produced in the visual field by the structural elements, i.e., primitives in image.

To evaluate the generalization capability of training data for nearest neighbor matching, a Receiver Operating Characteristics (ROC) curve is used to demonstrate the tradeoff between hit rate and match error. For a given match error e, the hit rate h is the percentage of test data whose match errors are less than e. Each test sample $\mathbf{x}$'s match error $e(\mathbf{x})$ is defined by a metric between $\mathbf{x}$ and the nearest sample $\mathbf{x}'$ in the training data. We use the metric $e(\mathbf{x}) = \frac{\|\mathbf{x}-\mathbf{x}'\|}{\|\mathbf{x}\|}$. At a given match error, the higher hit rate represents the better generalization of the training data.

For convenience, each 9×9 patch extracted from an image is called *component*. We study two ROC curves from a primitive training data set D_p (where each example is a primitive) and a component training data D_i (where each example is not necessarily a primitive), as shown in Figure 4. An empirical data set (1000 Hateren natural images [14][1]) are divided equally into training images and test images. D_p and D_i are constructed (with uniformly sampling) from training images. Each component is normalized as well. The ROC characteristics of D_p and D_i are evaluated on test images. About 10,000 test samples are uniformly sampled from the test images. To reduce the sensitivity of sampling, each curve is the average result of 50 repeated experiments (the training data and test data in each experiment are re-sampled from images).

Two observations are found from the ROC curves in Figure 4. One is that the hit rate of D_p is higher than that of D_i (for $|D_i| = |D_p|$) at any match error (except for 0 and 1). When $|D_p| = 10^6$, the match error is less than 0.2 for 85% primitives in the test images. Furthermore, 97% od the test data can find good matching examples in D_p in error range 0.3. But the corresponding hit rates are 48% and 60% for D_i. That means about half of the components cannot find good examples in the training data if we use components for image hallucination. The other one is that the slope of D_p's ROC curve increases significantly as $|D_p|$ increases. A better ROC of D_p can be expected when $N = 10^7$ (3GB byte memory storage required for 9x9 patches!). However, the the slope of D_i's ROC curve is close to a constant at different $|D_i|s$. If we extrapolate Figure 4 (b), reaching a 80% hit rate at match error 0.2 is hopeless with current storage and computing capabilities. Therefore, the primitive manifold can be represented well by a small number of examples, but the component manifold cannot. This is why we only focus on the primitive layer in image hallucination.

4 Image hallucination

The task now is to hallucinate the high frequency primitive layer I_H^{p*} given I_H^l according to MAP (1). Figure 5 shows the training phase and synthesis phase of our image hallucination.

[1] http://hlab.phys.rug.nl/archive.html

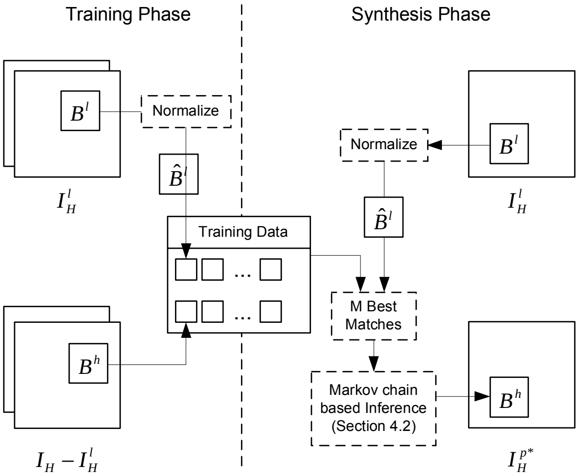


Figure 5. Image Hallucination. In the training phase, pairs of normalized low frequency primitive $\widehat{B}^l$ and high frequency primitive B^h are collected into the training data. In the synthesis phase, the M best matched examples are selected from the training data for each normalized low frequency primitive $\widehat{B}^l$ in the test image. The final high frequency primitive B^h is obtained by a Markov chain based inference algorithm.

4.1 Training

The training images are derived from 16 high resolution natural images in Figure 8. The low resolution images I_L are simulated from the high resolution images by blurring and down-sampling. Then, the low frequency image I_H^l is interpolated (bicubic) from I_L and the high frequency image is obtained from I_H by subtracting I_H^l. The low frequency primitive B^l and corresponding high frequency primitive B^h are extracted from these training images. We normalize B^l to get $\widehat{B}^l$ by (4). Each example in the training data consists of $\widehat{B}^l$, its contrast c^l and B^h.

4.2 Synthesis: Markov chain based Inference

For any low resolution test image I_L, a low frequency image I_H^l is interpolated from I_L at first. We assume that the primitive layer I_H^p to be inferred is a linear sum of a number of N high frequency primitives $\{B_n^h, n = 1, \ldots, N\}$. The underlying low frequency primitives $\{B_n^l, n = 1, \ldots, N\}$ in the I_H^l are shown in Figure 6 (b). Note that the center of each image patch is on the contours extracted in I_H^l and the neighboring patches are overlapped.

A straightforward nearest neighbor algorithm can be used for this task. For each low frequency primitive B_n^l, we get its normalized $\widehat{B}_n^l$, then we find the best matched normalized low frequency primitive to $\widehat{B}_n^l$ in the training data and paste the corresponding high frequency primitive. However, this simple method cannot preserve contour smoothness because the consistency of neighboring primitives is ignored, as shown in Figure 6 (c). Therefore, we present a Markov chain based inference algorithm to enforce the contour smoothness constraint (Figure 6 (d)).

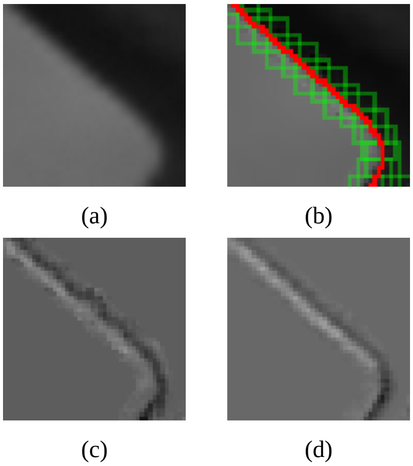

(a) (b)
(c) (d)

Figure 6. Comparison. (a) The low-frequency image. (b) The patches extracted along a contour. (c) Nearest neighbor algorithm. (d) Markov chain based algorithm.

To ensure the high frequency primitives to be consistent along the contour, the primitives are grouped into a number of K contours $C = \{C_k, k = 1, \ldots, K\}$ by a greedy 8-neighbors algorithm. We approximate the joint posterior $p(I_H^p|I_H^l)$ in (1) by the products of the posterior of each contour,

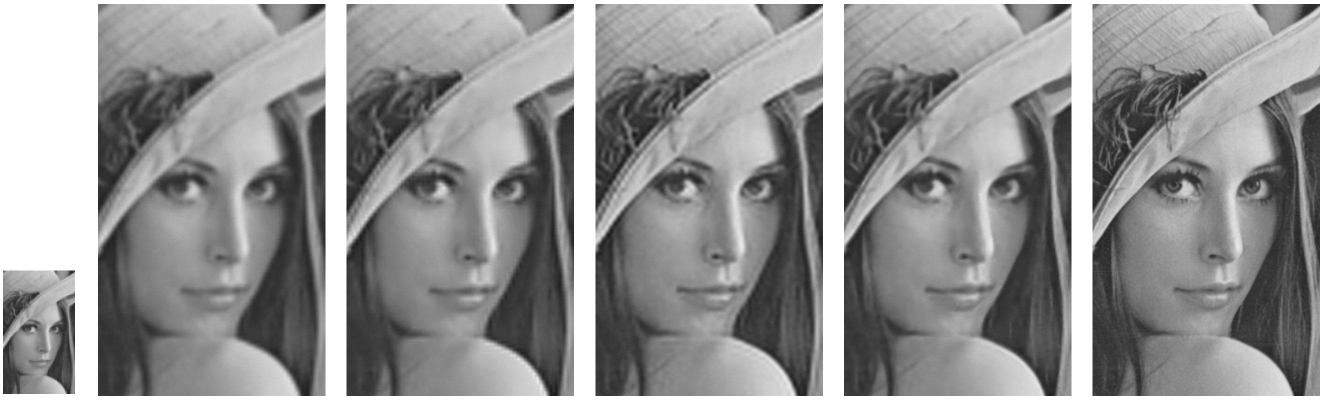
$$p(I_H^p|I_H^l) = p(C|I_H^l) \approx \prod_k p(C_k|I_H^l). \qquad (5)$$

Each contour C_k is a first order Markov chain model,

$$p(C_k|I_H^l) \propto \prod_i^{n_k-1} \Psi(B_i^h, B_{i+1}^h) \prod_i^{n_k} \Phi(B_i^l, B_i^h) \qquad (6)$$

where B_i^l is the *ith* low frequency primitive on contour C_k in I_H^l, B_i^h is the corresponding high frequency primitive to be inferred, n_k is the number of patches on C_k. $\Psi(B_i^h, B_{i+1}^h)$ is the compatibility function between two adjacent patches. $\Phi(B_i^l, B_i^h)$ is the local evidence function between B_i^l and B_i^h.

For each B_i^l, we compute its normalized primitive $\widehat{B}_i^l$ and the contrast c_i^l by equation (4). Its scaling and orientation parameters are estimated during primitive extraction. Because the relationship between $\widehat{B}_i^l$ and B_i^h is one-to-multiple mapping, M (8 in our experiments) best matching normalized low frequency primitives $\{\widehat{B}^l(m), m =$



a) Input b) Bicubic c) Sharpen Bicubic d) Backprojection e) Our approach f) Original

Figure 7. Comparison of the "Lena" image at 3X magnification.

$1, \cdots, M\}$ to $\widehat{B}_i^l$ are selected from the same subclass as B_i^l in the training data. Let $B^h(m)$ and c_m^l be the corresponding high frequency primitive and the contrast of $\widehat{B}^l(m)$ in the training data. The number of m high frequency patches are $\{B_i^h(m) = \frac{c_i^l}{c_m^l} B^h(m), m = 1, \cdots, M\}$. The scale factor $\frac{c_i^l}{c_m^l}$ compensates $B^h(m)$ for the different contrasts between $\widehat{B}_i^l$ and $\widehat{B}^l(m)$.

Each candidate $B_i^h(m)$ is treated equally by setting $\Phi(B_i^l, B_i^h) = \frac{1}{M}$. The compatibility function $\Psi(B_i^h, B_{i+1}^h)$ is defined by the compatibility of neighboring patches, $\Psi(B_i^h, B_{i+1}^h) = exp(-(d(B_i^h, B_{i+1}^h)/\sigma_d^2)$, where $d(B_i^h, B_{i+1}^h)$ is the Sum Squared Difference (SSD) of the overlapping region between B_i^h and B_{i+1}^h and σ_d is a tuning variance.

The optimal MAP solution of (6) for each contour C_k is obtained by running the Belief Propagation (BP) [12] algorithm. The details of the BP algorithm are not presented due to space limitations.

5. Experimental Results

We tested our approach on a set of generic images. The input low resolution image is produced by blurring and downsampling the high resolution image. Our experimental results are shown in Figures 7, 9 - 12, all with a magnification factor of 3. The PSF is a Gaussian function with a standard variance of 1.4. The "backprojection" filter p is also a Gaussian kernel with a standard variance of 2.2. Note that we do hallucination on the image intensity only because the humans are more sensitive to the brightness information. The color channels are simply interpolated by a bicubic function.

About 1,400,000 primitive examples are extracted from 16 representative natural images (see Figure 8) on a Kodak website [2]. All primitives are divided into 36 subclasses (2 scales x 16 orientations) by the scale and orientation information estimated using the orientation energy. Thus, the training data is organized hierarchically. The top level captures the primitive's global structure. The bottom level is a non-parametric representation that captures the local variations of the primitives. This two-level structure can speed up the AAN tree searching algorithm [10] in the training data. The run time of this algorithm is 20-100 seconds on a Pentium IV 1.7G PC for all the images in our experiments.

Figure 8. Training images. All training examples in this paper are extracted from these images (1536x1024 pixels).

We compare our approach with bicubic interpolation, sharpened bicubic interploation (using the "unsharp mask" in Adobe Photoshop with the default parameters onto the bicubic interpolation) and backprojection algorithm in Figure 7, 9-12. Bicubic is the smoothest one. Sharpened bicubic and backprojection methods introduce strong "ringing effect", especially along the contours in images. On the other hand, sharper and smoother contours are hallucinated by our approach (see the edges of the hat in Figure 7 (e), hairs in Figure 11, etc.). Figure 12 shows more results. (We

[2] http://www.kodak.com/digitalImaging/samples/imageIntro.shtml

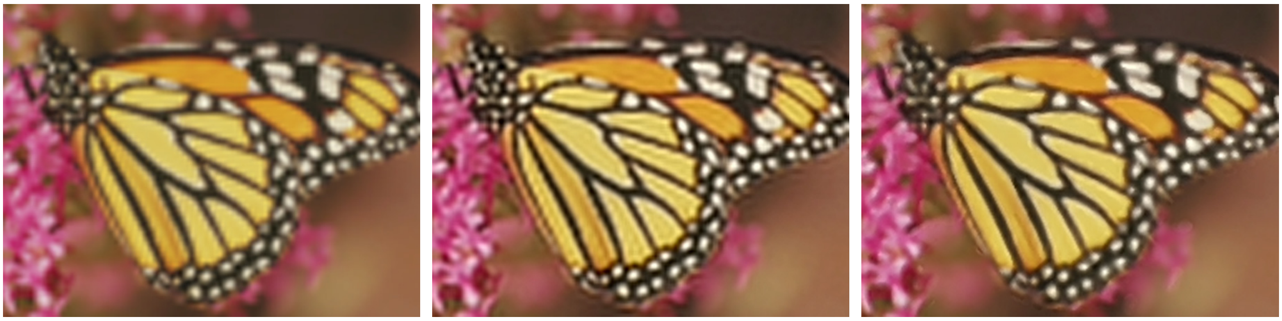


Figure 9. The “Monarch” image magnified 3X using sharpen bicubic (left), backprojection (middle) and our approach (right).

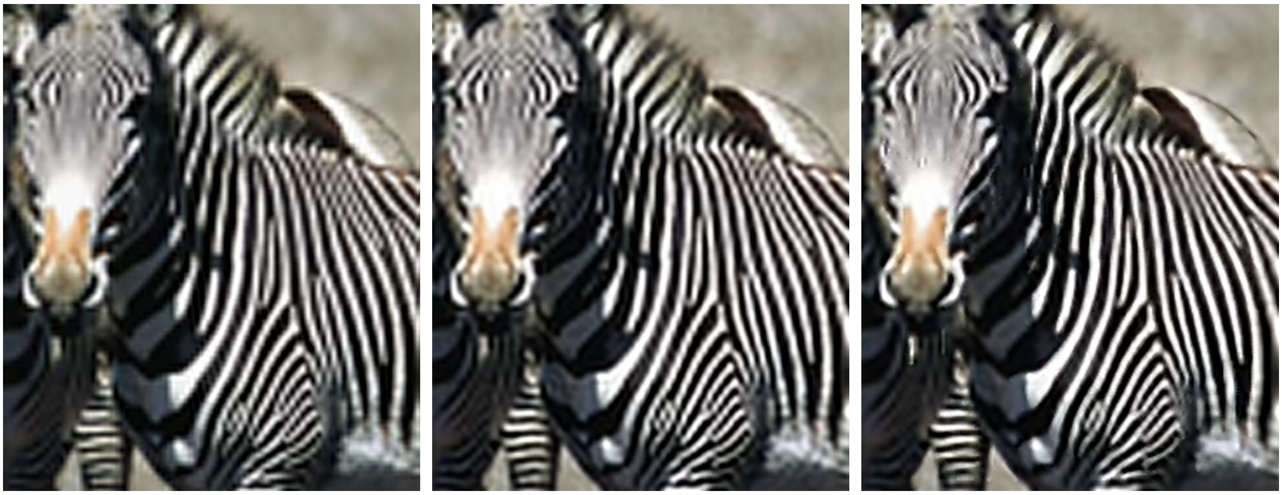


Figure 10. The “Zebra” image magnified 3X using sharpen bicubic (left), backprojection (middle) and our approach (right).

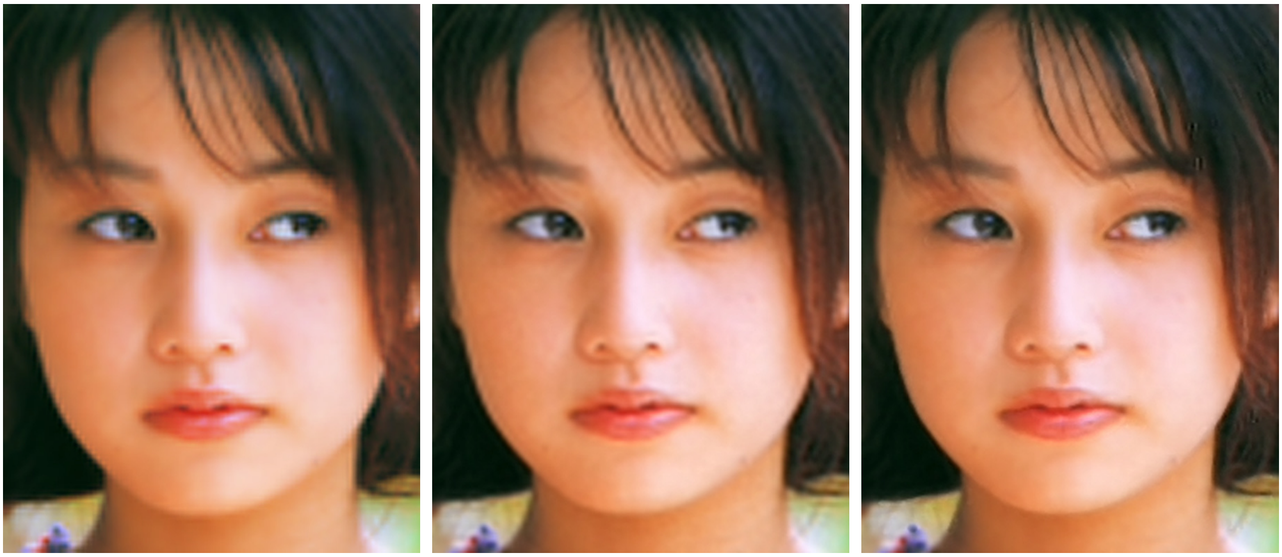


Figure 11. The “Girl” image magnified 3X using sharpen bicubic (left), backprojection (middle) and our approach (right).

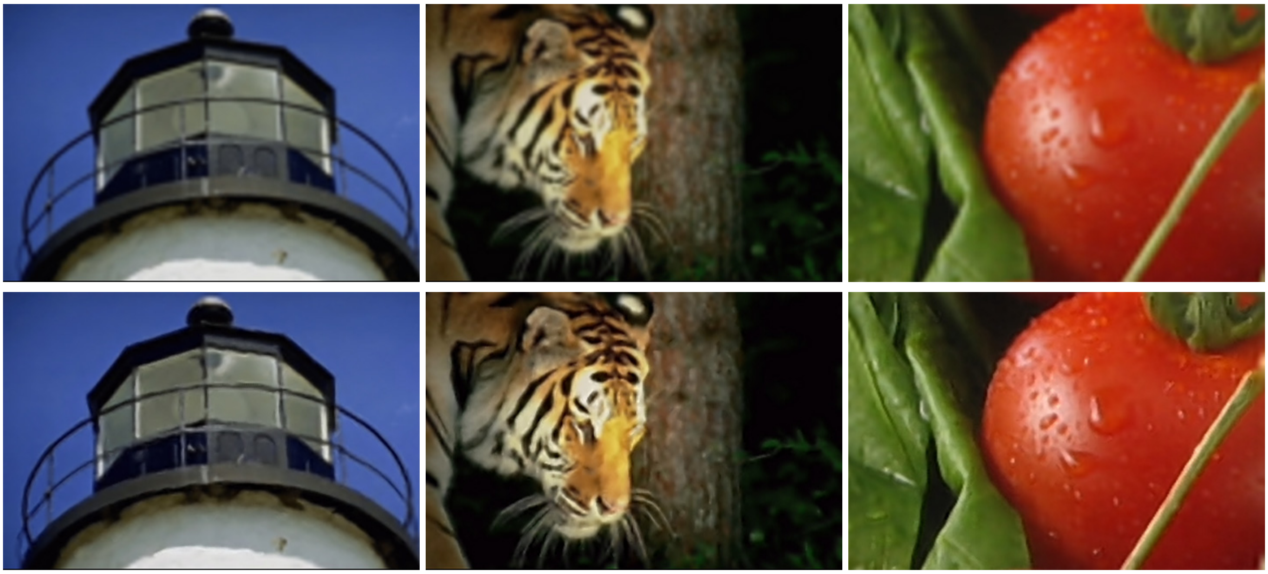


Figure 12. Image Hallucination results magnified 3X. The bottom row is hallucinated from the top row.

	Bicubic		Backprojection		Our Method	
Image	*RMS*	*ERMS*	*RMS*	*ERMS*	*RMS*	*ERMS*
Lena	14.0	18.1	7.4	9.3	6.0	7.4
Monarch	32.1	38.4	22.2	26.6	20.4	21.6
Zebra	59.1	64.7	42.0	45.4	38.0	40.9
Girl	11.3	14.5	7.8	10.2	7.1	9.1

Table 1. *RMS* pixel error and Edge Squared Mean Error (*ERMS*) pixel error for different approaches.

recommend the audience to see the electronic version.)

To compare the results quantitatively, we compute the RMS pixel error on the whole image and the edge regions respectively. Table 1 shows the results of applying on four images. Our approach outperforms the other approaches, especially around the edge regions where human perception cares most.

6. Conclusions

In this paper, an image hallucination approach has been presented based on the primal sketch priors. For single image super-resolution, encouraging results are obtained for generic images. For practical applications, the robustness of our approach with respect to an inaccurate PSF needs to be studied in further work.

References

[1] Lee. A.B., K.S. Pedersen, and D. Mumford. The complex statistics of high-contrast patches in natural images. *SCTV*, 2001.

[2] J. Allebach and P. W. Wong. Edge-directed interpolation. *ICIP*, 1996.

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *CVPR*, 2000.

[4] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *IJCV*, 2000.

[5] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D.H. Salesin. Image analogies. *SIGGRAPH*, 2001.

[6] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion and transparency. *JVCI*, 1993.

[7] C. Liu, H.Y. Shum, and C.S. Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. *CVPR*, 2001.

[8] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1982.

[9] B. S. Morse and D. Schwartzwald. Image magnification using level-set reconstruction. *CVPR*, 2001.

[10] D. Mount and S. Arya. Ann: Library for approximate nearest neighbor searching. *http://www.cs.umd.edu/ mount/ANN/*.

[11] D. Mumford. Pattern theory: A unifying perspective. *Perception as Bayesian Inference*, 1996.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 1988.

[13] P. Perona and J. Malik. Detecting and localizing edges composed of steps, peaks and roofs. *ICCV*, 1990.

[14] J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society ser B.*, 1998.

[15] L. Xin and M.T. Orchard. New edge-directed interpolation. *ICIP*, 2000.