MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Markov Networks for Super-Resolution

William T. Freeman and Egon C. Pasztor*
MERL, Mitsubishi Electric Research Labs.
201 Broadway
Cambridge, MA 02139

TR-2000-08 March 2000

Abstract

We address the super-resolution problem: how to estimate missing high spatial frequency components of a static image. From a training set of full- and low-resolution images, we build a database of patches of corrsponding high- and low-frequency image information. Given a new low-resolution image to enhance, we select from the training data a set of 10 candidate high-frequency patches for each patch of the low-resolution image. We use compatibility relationships between neighboring candidates in Bayesian belief propagation to select the most probable candidate high-frequency interpretation at each image patch. The resulting estimates of the high-frequency image are good. The algorithm maintains sharp edges, and makes visually plausible guesses in regions of texture.

Published in Proceedings of 34th Annual Conference on Information Sciences and Systems (CISS 2000), Dept. Electrical Engineering, Princeton University, Princeton, NJ 08544-5263, March, 2000

1. First printing, TR2000-08, March, 2000

* Egon Pasztor's present address:
MIT Media Lab
20 Ames St.
Cambridge, MA 02139

Markov Networks for Super-Resolution

W. T. Freeman and E. C. Pasztor[1]
MERL (Mitsubishi Electric Research Lab)
201 Broadway, Cambridge, MA 02139
e-mail: freeman@merl.com, pasztor@media.mit.edu

Abstract — **We address the super-resolution problem: how to estimate missing high spatial frequency components of a static image. From a training set of full- and low- resolution images, we build a database of patches of corrsponding high- and low-frequency image information. Given a new low-resolution image to enhance, we select from the training data a set of 10 candidate high-frequency patches for each patch of the low-resolution image. We use compatibility relationships between neighboring candidates in Bayesian belief propagation to select the most probable candidate high-frequency interpretation at each image patch. The resulting estimates of the high-frequency image are good. The algorithm maintains sharp edges, and makes visually plausible guesses in regions of texture.**

I. Introduction

One goal of computer vision is to infer the underlying "scene" which renders to the observed "image". The scene might be a description of lighting, shape and reflectance. The image might be a line drawing or a photograph or a video sequence. One can also treat a signal processing problem in this estimation theory framework: what are the missing high frequency details (scene) implied by a given low-resolution picture (image)?

It is typically too complex to estimate the scene corresponding to an entire image all at once. A common approach is to form interpretations of *local* image regions, and then propagate those interpretations across space (eg, [25, 1]).

We follow that approach in a probabilistic framework. We store a large training set of candidate local interpretations for patches of image data. We also determine the compatibility between neighboring scene interpretations, forming a Markov network model of the image and the underlying scene (Fig. 1). Given new image data, we find an approximation to the most probable scene interpretation using Bayesian belief propagation [20, 26].

We call this approach to vision problems VISTA, Vision by Image-Scene TrAining. Journal [9] and conference reports [7, 8] supplement this workshop manuscript. Here, we focus on one application, that of estimating high resolution detail from low resolution images.

II. Super-resolution

For the super-resolution problem, the input *image* is a low-resolution image. The *scene* to be estimated is the high resolution version of the same image. (Note this is different than another problem sometimes called super-resolution, that of estimating a single high resolution image from multiple low-resolution ones). A good solution to the super-resolution problem would allow pixel-based images to be handled in an almost resolution-independent manner. Applications could include enlargment of digital or film photographs, upconversion of video from NTSC format to HDTV, or image compression.

At first, the task may seem impossible—the high resolution data is missing. However, we can visually identify edges in the low-resolution image that we know should remain sharp at the next resolution level. Furthermore, the successes of recent texture synthesis methods [14, 6, 29, 24], gives us hope that we might handle textured areas well, too.

Others [23] have used a Bayesian method for super-resolution, making-up the prior probability. In contrast, the VISTA approach learns the relationship between sharp and blurred images from training examples, and, we believe, achieves better results. Among non-Bayesian methods for super-resolution, fractal image representation [22] (Fig. 8c) in effect gathers training data from only one image, which may not be adequate. Selecting the nearest neighbor from training data [21] (Fig. 6a) omits important spatial consistency constraints.

In our approach, training data provides candidate high-resolution explanations for the low-resolution image data. Modelling the image as a Markov network, Bayesian belief propagation quickly finds estimates for the most probable corresponding high-resolution explanation.

III. Belief Propagation

For given image data, y, we seek to estimate the underlying scene, x (we omit the vector symbols for notational simplicity). We first calculate the posterior probability, $P(x|y) = cP(x, y)$ (for this analysis, we ignore the normalization, $c = \frac{1}{P(y)}$, a constant over x). Under two common loss functions [2], the best scene estimate, $\hat{x}$, is the mean (minimum mean squared error, MMSE) or the mode (maximum a posteriori, MAP) of the posterior probability.

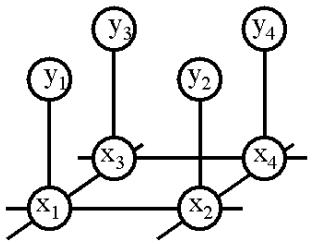


Fig. 1: Markov network for vision problems. Observations, y, have underlying scene explanations, x. Lines in the graph indicate statistical dependencies between nodes.

[1]Present address: MIT Media Lab, 20 Ames St., Cambridge, MA 02139

For networks without loops, the Markov assumption leads to simple "message-passing" rules for computing the MAP and MMSE estimates [20, 27, 16]. To derive these rules, we first write the MAP and MMSE estimates for x_j at node j by marginalizing (MMSE) or taking the maximum (MAP) over the other variables in the posterior probability:

$$\hat{x_j}_{MMSE} = \int_{x_j} x_j dx_j \int_{\text{all } x_i,\ i \neq j} P(x,y)dx \quad (1)$$

$$\hat{x_j}_{MAP} = \underset{[x_j]}{\text{argmax}} \ \underset{[\text{ all } x_i,\ i \neq j\ ]}{\text{max}} \ P(x,y). \quad (2)$$

For a Markov random field, the joint probability over the scenes x and images y can be written as [3, 13, 12]:

$$P(x,y) = \prod_{\text{neighboring } i,j} \Psi(x_i, x_j) \prod_k \Phi(x_k, y_k), \quad (3)$$

where we have introduced pairwise compatibility functions, Ψ and Φ, described below. The factorized structure of Eq. (3) allows the marginalization and maximization operators of Eqs. (1) and (2) to pass through compatibility function factors with unrelated arguments.

Consider the example network of three nodes, x_1, x_2, and x_3, connected in a chain, with a corresponding y_i node attached to each x_i node. We have

$$\begin{aligned} x_1\text{MAP} &= \text{argmax}_{x1}\text{max}_{x2}\text{max}_{x3} \\ &\quad P(x_1, x_2, x_3, y_1, y_2, y_3) \quad (4) \\ &= \text{argmax}_{x1}\text{max}_{x2}\text{max}_{x3} \\ &\quad \Phi(x_1, y_1)\Phi(x_2, y_2)\Phi(x_3, y_3)\Psi(x_1, x_2)\Psi(x_2, x_3) \quad (5) \\ &= \text{argmax}_{x1}\Phi(x_1, y_1) \\ &\quad \text{max}_{x2}\Psi(x_1, x_2)\Phi(x_2, y_2) \\ &\quad \text{max}_{x3}\Phi(x_3, y_3)\Psi(x_2, x_3). \quad (6) \end{aligned}$$

Each line of Eq. (6) is a local computation involving only one node and its neighbors. The analogous expressions for x_2MAP and x_3MAP use the same local calculations. Iterating those calculations lets each node j compute x_jMAP from the messages passed between nodes.

This works for any network without loops; Eqs. (2) and (1) can be computed by iterating the following steps [20, 27, 16]. The MAP estimate at node j is

$$\hat{x_j}_{MAP} = \underset{[x_j]}{\text{argmax}} \ \Phi(x_j, y_j) \prod_k L_{kj}, \quad (7)$$

where k runs over all scene node neighbors of node j. We calculate L_{kj} from:

$$L_{kj} = \underset{[x_k]}{\text{max}} \ \Psi(x_k, x_j)\Phi(x_k, y_k) \prod_{l \neq j} \tilde{L}_{lk} dx_k, \quad (8)$$

where $\tilde{L}_{lk}$ is L_{lk} from the previous iteration. The initial $\tilde{L}_{lk}$'s are 1. After at most one iteration of Eq. (8) per scene node variable, Eq. (7) gives the desired optimal estimate, $\hat{x_j}_{MAP}$. The MMSE estimate, Eq. (2), has analogous formulae, with the max_{x_k} of Eq. (8) replaced by $\int_{x_k}$, and argmax_{x_j} of Eq. (7) replaced by $\int_{x_j} x_j$. For linear topologies, these propagation rules are equivalent to standard Bayesian inference methods, such as the Kalman filter and the forward-backward algorithm for Hidden Markov Models [20, 18, 26, 16, 11].

For a network with loops, the factorization of Eqs. (1) and (2) into local calculations doesn't hold. Finding the posterior probability distribution for a Markov network with loops is computationally expensive and researchers have proposed a variety of approximations [13, 12, 16]. Strong empirical results in "Turbo codes" [17, 19] and recent theoretical work [27, 28] provide support for a very simple approximation: applying the propagation rules derived above *even in the network with loops*. Table 1 summarizes the results from [28, 10]: (1) for Gaussian processes, the MMSE propagation scheme will converge only to the true posterior means. (2) Even for non-Gaussian processes, if the MAP propagation scheme converges, it finds at least a local maximum of the true posterior probability. Furthermore, this condition of local optimality for the converged solution of the MAP algorithm is a strong one. For every subset of nodes of the network which form a tree, if the remaining network nodes are constrained to their converged values, the values of the sub-tree's nodes found by the MAP algorithm are the *global* maximum over that tree's nodes [28]. These experimental and theoretical results motivate and justify applying the belief propagation rules even in a Markov network with loops.

IV. Implementation

By blurring and downsampling sharp images, we construct a training set of sharp and blurred image pairs. We linearly interpolate each blurred image back up to the original sampling resolution, to form an input *image*. The *scene* to be estimated is the high frequency detail missing from the blurred image, Fig. 4a, b.

A key to good performance with real images is to reduce the modelling burden, which we do in two ways. (1) We believe the lowest frequencies of the blurred image don't predict the high frequencies of the scene, and we don't want to have to learn the image/scene relationship for all possible values of the low frequencies. So we bandpass filter the blurred image. (2) We believe that the relationship between the highest and lower frequencies in an image is the same for different image contrasts, just multiplicatively scaled in image intensity. We don't want to have to memorize that relationship for all possible values of local contrast, so we normalize both the bandpass and highpassed images by the local contrast [15] of the bandpassed image, Fig. 4c, d. We undo this normalization after estimating the scene.

We extracted center-aligned 7x7 and 3x3 pixel patches, Fig. 5, from the training images and scenes. Applying Principal Components Analysis (PCA) [4] to the training set, we summarized each 3-color patch of image or scene by a 9-d vector. We collected approximately 40,000 image/scene pair samples for the training data. For efficiency, we pruned frequently occurring image/scene pairs from the training set, based on a squared error similarity criterion. Figure 5 shows some typical training samples.

Given a new image, not in the training set, from which to infer the high frequency scene, we found the 10 training samples closest to the image data at each node (patch). The 10 corresponding scenes are the candidates for that node. We evaluated $\Psi(x_j, x_k)$ at 100 values (10 x_j by 10 x_k points) to form a compatibility matrix for messages from neighbor nodes j to k. [9] describes the method used to calculate $\Psi(x_i, x_j)$ and $\Phi(x_k, y_k)$ for the processed images shown here; Fig. 3 shows a preferred, simpler approach. Given $\Psi(x_i, x_j)$ and $\Phi(x_k, y_k)$,

Belief propagation algorithm	*Network topology*	
	no loops	*arbitrary topology*
MMSE rules	MMSE, correct posterior marginal probs.	For Gaussians, correct means, wrong covs.
MAP rules	MAP	Strong local max. of posterior, even for non-Gaussians.

Tab. 1: Summary of results from [28] regarding belief propagation after convergence.

we propagated the probabilities by Eq. (8), and read-out the maximum probability solution by Eq. (7).

To process Fig. 7a, we used a training set of 80 images from two Corel database categories: African grazing animals, and urban skylines. For reference, Fig. 6a shows the nearest neighbor solution, at each node using the scene corresponding to the closest image sample in the training set. Many different scene patches can explain each image patch, and the nearest neighbor solution is very choppy. Figures 6b, c, d show the first 3 iterations of MAP belief propagation. The spatial consistency imposed by the belief propagation finds plausible and consistent high frequencies for the tiger image from the candidate scenes.

Figure 7 shows the result of applying this super-resolution method recursively to zoom two octaves. The algorithm keeps edges sharp and invents plausible textures. Standard cubic spline interpolation, blurrier, is shown for comparison. Figure 8 shows other results of the algorithm, using different training sets.

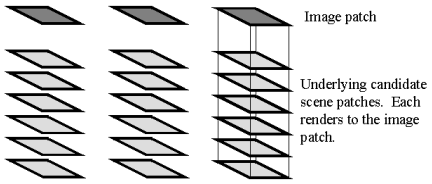


Fig. 2: Showing sampled version of the problem to be solved. From the training database, we gather a collection of candidate scene patches. Each candidate scene can explain the observed image data, possibly some better than others. Neighboring image patches have their own sets of scene candidates. The compatibilities between the candidates of neighboring scenes drives the optimal image interpretation.

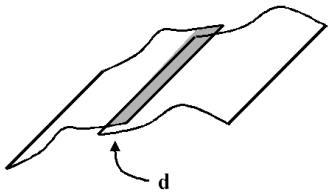


Fig. 3: Preferred method to compute compatibility function, $\Psi(x_k, x_j)$. Where neighboring patches overlap (grey region), we require that pixels in each patch have the same value, imposing in $\Psi(x_k, x_j)$ a mean zero Gaussian probability penalty for those pixel differences. This is not a smoothness constraint per se; the patch pixel values can be very "rough", provided they agree with the values of the neighboring patch in the overlap region.

V. Summary

We treat the image processing problem of resolution enhancement in the framework of low-level computer vision problems. Training data provides multiple candidate high resolution explanations at each local region of the image. Bayesian belief propagation selects a spatially consistent set of those candidate high frequency explanations. The result is a plausible high resolution extension of the original image.

References

[1] H. G. Barrow and J. M. Tenenbaum. Computational vision. *Proc. IEEE*, 69(5):572–595, 1981.

[2] J. O. Berger. *Statistical decision theory and Bayesian analysis.* Springer, 1985.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Royal Statist. Soc. B*, 36:192–326, 1974.

[4] C. M. Bishop. *Neural networks for pattern recognition.* Oxford, 1995.

[5] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans. Comm.*, 31(4):532–540, 1983.

[6] J. S. DeBonet and P. Viola. Texture recognition using a non-parametric multi-scale statistical model. In *Proc. IEEE Computer Vision and Pattern Recognition*, 1998.

[7] W. T. Freeman and E. Pasztor. Learning low-level vision. In *7th International Conference on Computer Vision*, pages 1182–1189, 1999. See also http://www.merl.com/reports/TR99-12/.

[8] W. T. Freeman and E. C. Pasztor. Learning to estimate scenes from images. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Adv. Neural Information Processing Systems*, volume 11, Cambridge, MA, 1999. MIT Press. See also http://www.merl.com/reports/TR99-05/.

[9] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 2000. Accepted pending revisions. See also http://www.merl.com/reports/TR2000-05/.

[10] W. T. Freeman and Y. Weiss. On the fixed points of the max-product algorithm. Technical Report 99–39, MERL, 201 Broadway, Cambridge, MA 02139, 1999.

[11] B. J. Frey. *Graphical Models for Machine Learning and Digital Communication.* MIT Press, 1998.

[12] D. Geiger and F. Girosi. Parallel and deterministic algorithms from MRF's: surface reconstruction. *IEEE Pattern Analysis and Machine Intelligence*, 13(5):401–412, May 1991.

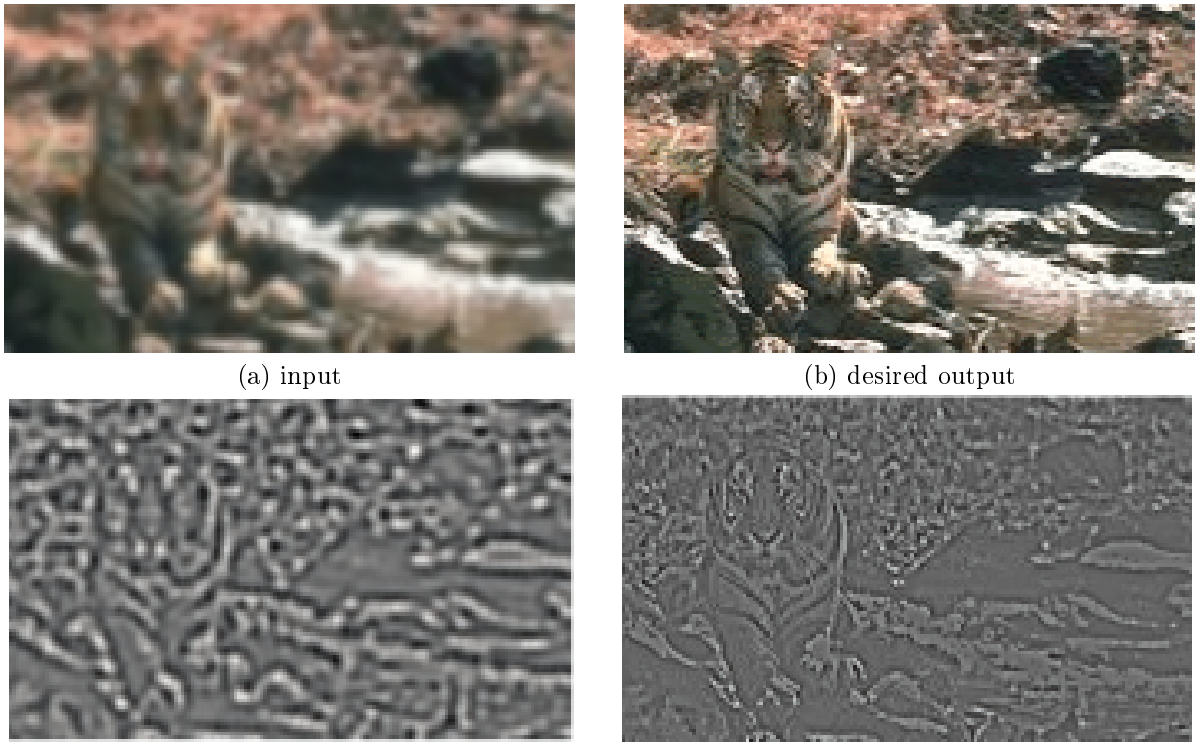
[13] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[14] D. J. Heeger and J. R. Bergen. Pyramid-based texture analysis/synthesis. In *ACM SIGGRAPH*, pages 229–236, 1995. In *Computer Graphics* Proceedings, Annual Conference Series.

[15] B. Jahne. *Digital Image Processing.* Springer-Verlag, 1991.

[16] M. I. Jordan, editor. *Learning in graphical models.* MIT Press, 1998.

[17] F. R. Kschischang and B. J. Frey. Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Journal on Selected Areas in Communication*, 16(2):219–230, 1998.



(a) input (b) desired output

(c) input image (d) scene to be estimated

Fig. 4: We want to estimate (b) from (a). The original image, (b) is blurred, subsampled, then interpolated back up to the original sampling rate to form (a). The missing high frequency detail, (b) minus (a), is the "scene" to be estimated, (d) (this is the first level of a Laplacian pyramid [5]). Two image processing steps are taken for efficiency: the low frequencies of (a) are removed to form the input bandpassed "image". We contrast normalize the image and scene by the local contrast of the input bandpassed image, yielding (c) and (d).

Fig. 5: Some training data samples for super-resolution problem. The large squares are the *image* data (mid-frequency data). The small squares below them are the corresponding *scene* data (high-frequency data).

[18] M. R. Luettgen, W. C. Karl, and A. S. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Trans. Image Processing*, 3(1):41–64, 1994.

[19] R. McEliece, D. MackKay, and J. Cheng. Turbo decoding as as an instance of Pearl's 'belief propagation' algorithm. *IEEE Journal on Selected Areas in Communication*, 16(2):140–152, 1998.

[20] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[21] A. Pentland and B. Horowitz. A practical approach to fractal-based image compression. In A. B. Watson, editor, *Digital images and human vision*. MIT Press, 1993.

[22] M. Polvere. Mars v. 1.0, a quadtree based fractal image coder/decoder, 1998. http://inls.ucsd.edu/y/Fractals/.

[23] R. R. Schultz and R. L. Stevenson. A Bayesian approach to image expansion for improved definition. *IEEE Trans. Image Processing*, 3(3):233–242, 1994.

[24] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conf. on Sig., Sys. and Computers*, Pacific Grove, CA, 1997.

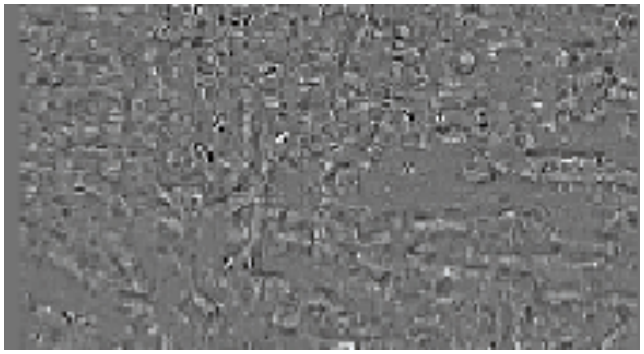
[25] D. Waltz. Generating semantic descriptions from drawings of scenes with shadows. In P. Winston, editor, *The psychology of computer vision*, pages 19–92. McGraw-Hill, New York, 1975.

[26] Y. Weiss. Interpreting images by propagating Bayesian beliefs. In *Adv. in Neural Information Processing Systems*, volume 9, pages 908–915, 1997.

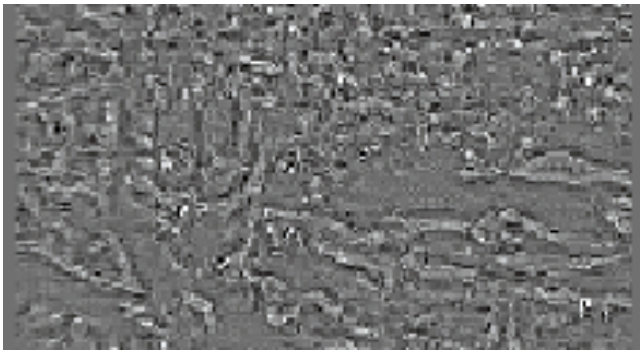
[27] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.

[28] Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. Technical Report UCB.CSD-99-1046, Berkeley Computer Science Dept., 1999. www.cs.berkeley.edu/~ yweiss/gaussTR.ps.gz.

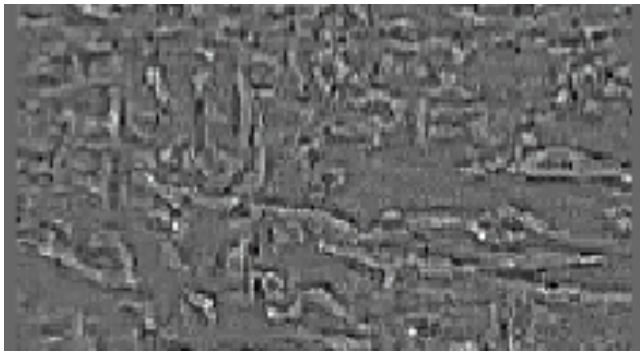
[29] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Pattern Analysis and Machine Intelligence*, 19(11), 1997.



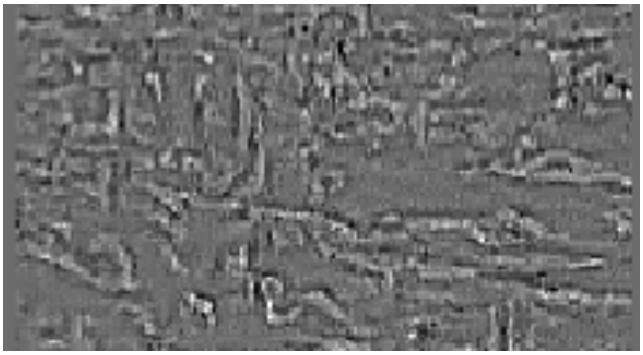
(a) Nearest neighbor



(b) belief prop., iter. 0



(c) belief prop., iter. 1



(d) belief prop., iter. 3

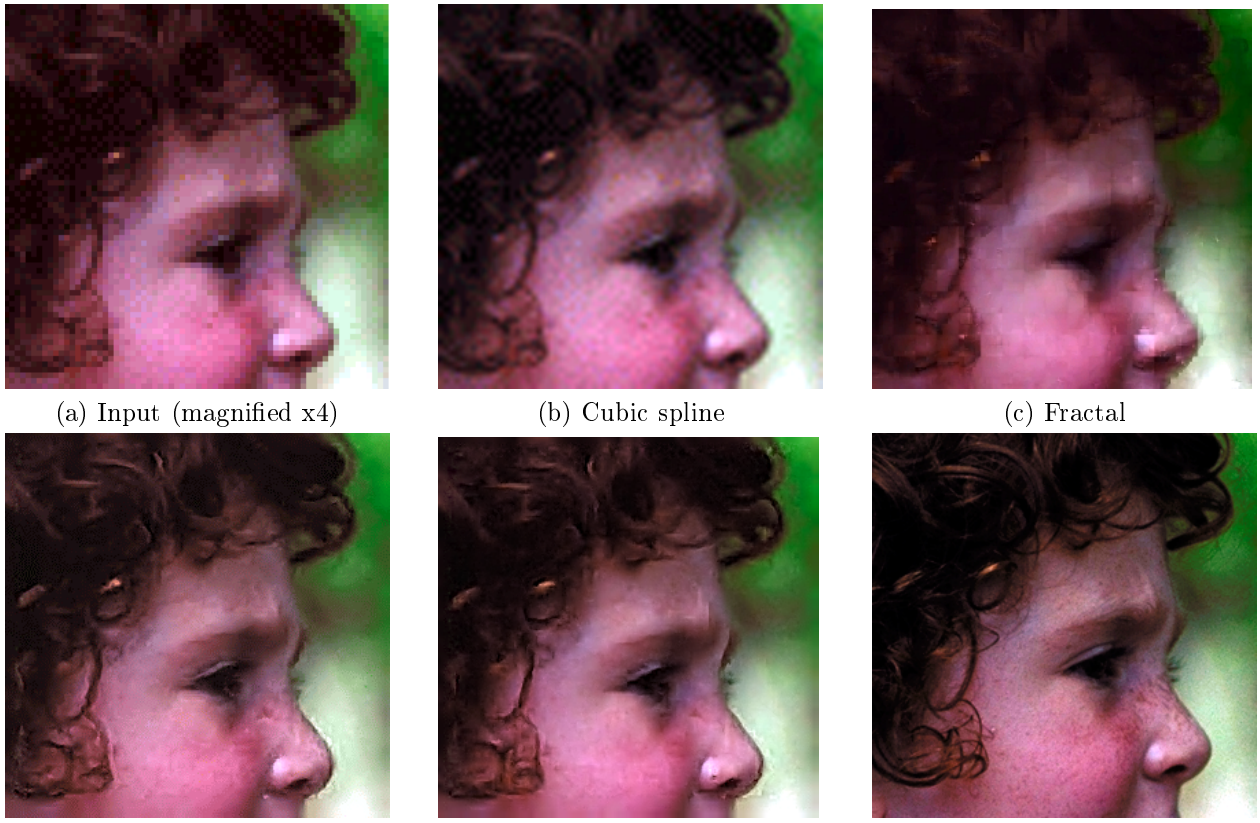
Fig. 6: (a) Nearest neighbor solution. The choppiness indicates that many feasible high resolution scenes correspond to a given low resolution image patch. (b), (c), (d): iterations 0, 1, and 3 of Bayesian belief propagation. The initial guess is not the same as the nearest neighbor solution because of mixture model fitting to $P(y|x)$. Underlying the most probable guess shown are 9 other scene candidates at each node. 3 iterations of Bayesian belief propagation yields a probable guess for the high resolution scene, consistent with the observed low resolution data, and spatially consistent across scene nodes.

(a) 85 x 51 input

(b) cubic spline

(c) belief propagation

Fig. 7: (a) 85 x 51 resolution input. (b) cubic spline interpolation in Adobe Photoshop to 340x204. (c) belief propagation zoom to 340x204, zooming up one octave twice.



(a) Input (magnified x4) (b) Cubic spline (c) Fractal

(d) "Picnic" training set (e) "Generic" training set (f) Actual full-resolution

Fig. 8: (a) Low-resolution input image. (b) Cubic spline 400% zoom in Adobe Photoshop. (c) Zooming luminance by public domain fractal image compression routine [22], set for maximum image fidelity (chrominance components were zoomed by cubic spline, to avoid color artifacts). Both (c) and (d) are blurry, or have serious artifacts. (d) Markov network reconstruction using a training set of 10 images taken at the same picnic, none of this person. This is the best possible fair training set for this image. (e) Markov network reconstrution using a training set of *generic* photographs, none at this picnic or of this person, and fewer than 50% of people. The two Markov network results show good synthesis of hair and eye details, with few artifacts, but (d) looks slightly better (see brow furrow). Edges and textures seem sharp and plausible. (f) is the true full-resolution image.