To Appear in the Proceedings of the 2000 IEEE Conference On Computer Vision and Pattern Recognition.

Limits on Super-Resolution and How to Break Them

Simon Baker and Takeo Kanade

The Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213

Abstract

We analyze the super-resolution reconstruction constraints. In particular, we derive a sequence of results which all show that the constraints provide far less useful information as the magnification factor increases. It is well established that the use of a smoothness prior may help somewhat, however for large enough magnification factors any smoothness prior leads to overly smooth results. We therefore propose an algorithm that learns recognition-based priors for specific classes of scenes, the use of which gives far better super-resolution results for both faces and text.

1 Introduction

Super-resolution is the process of combining multiple low resolution images to form a higher resolution one. Numerous algorithms have been proposed for it, dating back to the frequency domain approach of Huang and Tsai [1984]. In practice, however, the results obtained are mixed. While the super-resolution images are usually a huge improvement over the inputs, for large magnification factors the high frequencies are generally not reconstructed very well.

Most super-resolution algorithms are based on the constraints that the super-resolution image, when appropriately warped and down-sampled to model the image formation process, should yield the low resolution input images. We refer to super-resolution algorithms that explicitly use these constraints as *reconstruction-based.*

These reconstruction constraints have been used by numerous authors since first studied by Peleg *et al.* [1987] [Irani and Peleg, 1991]. The constraints can easily be embedded in a Bayesian framework incorporating a prior on the high resolution image [Schultz and Stevenson, 1996] [Hardie *et al.*, 1997] [Elad and Feuer, 1997]. The solution can be estimated either in batch mode or recursively using a Kalman filter [Elad and Feuer, 1999] [Dellaert *et al.*, 1998]. Several refinements have been proposed, including simultaneously computing structure [Cheeseman *et al.*, 1994] [Shekarforoush *et al.*, 1996] and removing other degrading effects such as motion blur [Bascle *et al.*, 1996].

In this paper, we first derive a sequence of results which all show that super-resolution gets much harder as the magnification factor increases. For square point spread functions (and integer magnifications), we show that the reconstruction constraints are not invertible, and that the dimension of the null space grows as a quadratic function of the magnification. For more general point spread functions, we show that both the condition number and the volume of the set of solutions grow equally fast. (We emphasize that these results hold even when the algorithms can use as many low-resolution images as they wish. It is not just that higher magnification requires more images.) The rate of increase in the difficulty of the problem is so great that beyond a magnification of around 8–16 (in each direction), the reconstruction constraints barely provide any new information. Our analysis shows that two factors combine to cause these difficulties: (1) the discretiztion of the intensities into a finite set of grey-levels, and (2) the integration of the illumination over a finite photosensitive area.

A partial solution to these problems is to impose a prior on the super-resolution image. Beyond a point, however, the use of typical "smoothness" priors cannot compensate for the fact that the reconstruction equations do not provide any more useful information. High magnification super-resolution results using smoothness priors therefore tend to look overly smooth. See, for example, the results in Figure 1 for the algorithm of Hardie *et al.* [1997].

In the second half of this paper, we introduce the notion of a *recognition-based* prior as a prior that is a function of a collection of recognition decisions. We propose an algorithm to learn a recognition-based prior for specific classes of objects, scenes, or images. We apply this algorithm to super-resolution, both for faces and text, obtaining significantly better results than traditional reconstruction-based super-resolution using standard smoothness priors.

2 The Reconstruction Constraints

Denote the low resolution images by $\mathrm{Lo}_i(\mathbf{m})$ and the high (super) resolution image by $\mathrm{Hi}(\mathbf{p})$, where i is an index, and $\mathbf{m} = (m, n)$ and $\mathbf{p} = (p, q)$ are the pixel coordinates in $\mathbf{Z}^2$. We assume that the low resolution images have been registered with the coordinate frame of the high resolution image (which is typically defined by one of the low resolution images.) Suppose $\mathbf{r}_i(\mathbf{p})$ denotes the point (pixel) in image Lo_i that corresponds to the point (pixel) $\mathbf{p}$ in Hi. The reconstruction constraints then take the form:

$$\mathrm{Lo}_i(\mathbf{m}) \; = \; \sum_{\mathbf{p}} \mathrm{W}_{\mathbf{r}_i}(\mathbf{m}, \mathbf{p}) \cdot \mathrm{Hi}(\mathbf{p}). \tag{1}$$

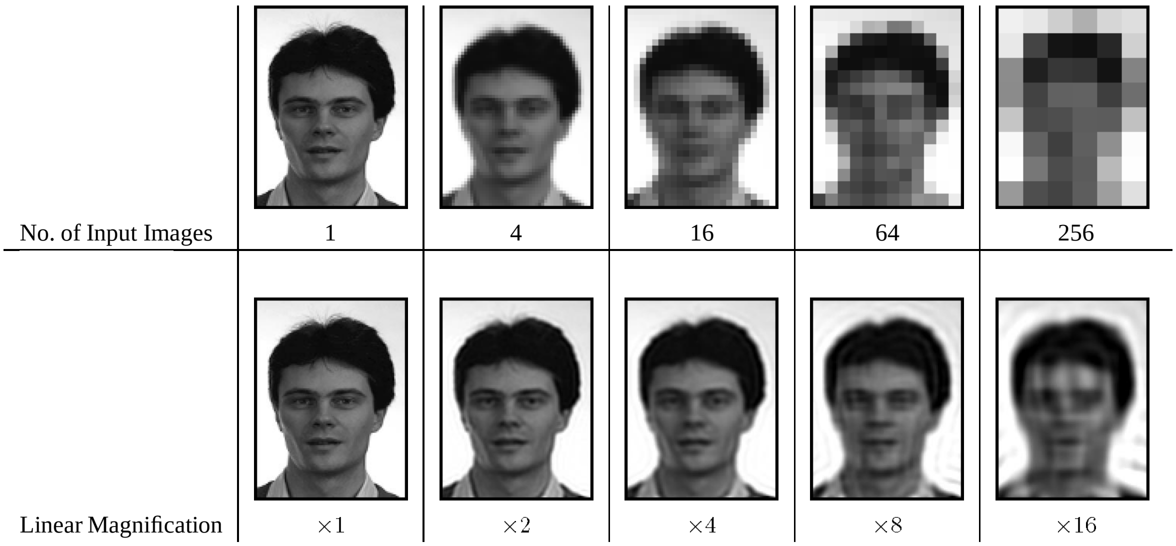


Figure 1: The results of the reconstruction-based algorithm [Hardie *et al.*, 1997] for various magnification factors. The original high-resolution image is translated multiple times, blurred with a Gaussian, and down-sampled. The algorithm is provided with knowledge of the point spread function and the translations. Comparing the images in the right-most column, we see that the algorithm does quite well given the resolution of the input. The degradation in performance as the magnification factor increases, however, is very dramatic.

As will be shown in the remainder of this section, the matrix of coefficients in this linear system $\mathrm{W}_{\mathbf{r}_i}(\cdot,\cdot)$ is a function of both the registration $\mathbf{r}_i$ and the point spread function $\mathrm{PSF}_i(\cdot)$ of the i^{th} low resolution image.

2.1 Derivation from the Point Spread Function

The reconstruction constraints in Equation (1) are derived from the continuous image formation equation:

$$\mathrm{Lo}_i(\mathbf{m}) \;=\; \int_{\mathrm{Lo}_i} \mathrm{PSF}_i(\mathbf{x}-\mathbf{m}) \cdot E(\mathbf{x})\,\mathrm{d}\mathbf{x} \qquad (2)$$

where $E(\cdot)$ is the continuous irradiance light-field that would have reached the image plane of Lo_i under the pinhole model, $\mathrm{PSF}_i(\cdot)$ is the point spread function of the camera, and $\mathbf{x} = (x, y) \in \mathbf{R}^2$ are coordinates on the image plane. (The additional integrations over time and illumination wavelength that are performed by a real camera are omitted since they do not affect the spatial analysis.)

2.1.1 The Point Spread Function

The point spread function of a camera is usually decomposed into two components:

$$\mathrm{PSF}_i(\mathbf{x}) \;=\; (a_i * \omega_i)(\mathbf{x}) \qquad (3)$$

where $\omega_i(\mathbf{x})$ models the blurring caused by the lens, $a_i(\mathbf{x})$ models the spatial integration performed by the sensor, and $*$ is the 2D convolution operator. The blurring factor $\omega_i(\cdot)$ is typically further split into a defocus factor that is approximated by a pill-box function [Born and Wolf, 1965], and the diffraction-limited optical transfer function that is approximated by the square of the first-order Bessel function of the first kind [Born and Wolf, 1965]. If the photosensitive areas of the pixels are square [Barbe, 1980], the spatial integration function is:

$$a_i(\mathbf{x}) \;=\; \begin{cases} \frac{1}{S_i^2} & \text{if } |x| \le \frac{S_i}{2} \text{ and } |y| \le \frac{S_i}{2} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where $S_i \in [0, 1]$ is the width of the photosensitive area.

The point spread function of a camera is therefore a very complex function that depends upon a large number of parameters that describe the defocus effects, the diffraction effects, and the shape and size of the photosensitive areas of the pixels. In practice, it is easiest to assume a simple parametric form for $\mathrm{PSF}_i(\cdot)$, for example that it is a Gaussian, and then estimate the parameters empirically. Since the point spread function of a sensor describes "the image of an isolated point object located on a uniformly black background" [Nalwa, 1993], it can be estimated from the image of a point light source placed a large distance away.

2.1.2 What is Super-Resolution Anyway?

The integration in Equation (2) is performed over the low resolution image plane. Transforming to the high resolution image plane using the registration $\mathbf{x} = \mathbf{r}_i(\mathbf{z})$ gives:

$$\mathrm{Lo}_i(\mathbf{m}) \;=\; \int_{\mathrm{Hi}} \mathrm{PSF}_i(\mathbf{r}_i(\mathbf{z})-\mathbf{m}) \cdot E(\mathbf{r}_i(\mathbf{z})) \cdot \left|\frac{\partial \mathbf{r}_i}{\partial \mathbf{z}}\right| \,\mathrm{d}\mathbf{z} \quad (5)$$

where $\left|\frac{\partial \mathbf{r}_i}{\partial \mathbf{z}}\right|$ is the determinant of the Jacobian of the registration transformation $\mathbf{r}_i$. (Note that here we have assumed that $\mathbf{r}_i$ is invertible. A similar analysis, albeit approximate, can be conducted wherever $\mathbf{r}_i$ is locally invertible by truncating the point spread function.) There are then at least two interpretations of super-resolution:

Super-Resolution As Image Restoration

The goal here is to recover $E(\mathbf{r}_i(\mathbf{z}))$, the irradiance under the pinhole model transformed into the coordinate frame of Hi. Recovering $E(\mathbf{r}_i(\mathbf{z}))$ requires both increasing the resolution and "deblurring" the image; ie. removing the effects of the convolution with the point spread function.

Super-Resolution As "Smaller Pixels"

Here the goal is to estimate $(\omega_i * E)(\mathbf{r}_i(\mathbf{z}))$, the irradiance reaching the sensor plane after passing through the optics, again transformed into the coordinate frame of Hi. From $(\omega_i * E)(\mathbf{r}_i(\mathbf{z}))$, it is easy to determine what the low resolution images would have been had the sensor arrays contained a larger number of smaller pixels.

In the remainder of this document, we consider the first of these two possibilities. The analysis of the second case is the same as the first under the special case that $\omega_i(\mathbf{x})$ is set to be the 2D "unit-impulse" Dirac delta function $\delta(\mathbf{x})$.

2.1.3 Representing Continuous Images

In order to proceed, we need to define which continuous function $E(\mathbf{r}_i(\mathbf{z}))$ is represented by the discrete image $\mathrm{Hi}(\mathbf{p})$ that we are trying to reconstruct. The simplest case is that $\mathrm{Hi}(\mathbf{p})$ represents the piecewise constant function:

$$E(\mathbf{r}_i(\mathbf{z})) \;=\; \mathrm{Hi}(\mathbf{p}) \qquad (6)$$

for all $\mathbf{z} \in (p-0.5, p+0.5] \times (q-0.5, q+0.5]$ and where $\mathbf{p} = (p,q) \in \mathbf{Z}^2$ are the coordinates of a pixel in Hi. Then, Equation (5) can be rearranged to give:

$$\mathrm{Lo}_i(\mathbf{m}) \;=\; \sum_{\mathbf{p}} \mathrm{Hi}(\mathbf{p}) \cdot \int_{\mathbf{p}} \mathrm{PSF}_i(\mathbf{r}_i(\mathbf{z}) - \mathbf{m}) \cdot \left| \frac{\partial \mathbf{r}_i}{\partial \mathbf{z}} \right| \, \mathrm{d}\mathbf{z} \quad (7)$$

where the integration is performed over the pixel $\mathbf{p}$; ie. over $(p-0.5, p+0.5] \times (q-0.5, q+0.5]$. Comparing this equation with Equation (1) gives:

$$\mathrm{W}_{\mathbf{r}_i}(\mathbf{m}, \mathbf{p}) \;=\; \int_{\mathbf{p}} \mathrm{PSF}_i(\mathbf{r}_i(\mathbf{z}) - \mathbf{m}) \cdot \left| \frac{\partial \mathbf{r}_i}{\partial \mathbf{z}} \right| \mathrm{d}\mathbf{z}. \qquad (8)$$

(Similar derivations can be performed for other representations of $E(\mathbf{r}_i(\mathbf{z}))$, such as piecewise linear ones.)

3 Analysis of the Reconstruction Constraints

The reconstruction constraints are therefore defined by Equation (7) (where $i = 1, 2, \ldots$). We now analyze these constraints under the following three ideal conditions:

The Point Spread Function is Constant and Known

We assume that $\mathrm{PSF}_i(\cdot)$ is the same for all of the images Lo_i (in particular the width of the photosensitive area S_i is constant) and that full knowledge of it is available.

The Registration is Known and is a Translation

We assume that the registration $\mathbf{r}_i(\cdot)$ is fully known by the super-resolution algorithm and that it takes the form:

$$\mathbf{r}_i(\mathbf{z}) \;=\; \frac{1}{M}\mathbf{z} + \mathbf{c}_i \qquad (9)$$

where $\mathbf{c}_i = (c_i, d_i) \in \mathbf{R}^2$ is a known constant and $M > 0$ is the *linear magnification* of the super-resolution task.

Arbitrary Number of Images with Chosen Translation

The super-resolution algorithm can use as many images as it wishes, and these images are captured with translations $\mathbf{r}_i(\cdot)$ chosen by the super-resolution algorithm.

All of these conditions make the super-resolution algorithm more powerful than in practice except: (1) assuming the PSF is a constant, and (2) assuming that the registration is a translation with constant magnification. If the PSF is not constant, exact super-resolution could be obtained simply by changing the size of the pixels so that they match those in the high resolution image. Similarly, if the registration were arbitrary, super-resolution could be obtained by setting $\mathbf{r}_i(\cdot)$ to be the identity. Both of these assumptions are also needed to give precise meanings to S_i and M which will appear in our analysis. One thing is clear, however. If we can derive limits on reconstruction-based super-resolution under these ideal conditions, performing super-resolution in practice will only be more difficult. Note that similar assumptions were used by Elad and Feuer [1997] [1999] to analyze super-resolution from varying defocus.

3.1 Real Valued Analysis, Square PSFs

First, we assume that all the quantities are real-valued; ie. we neglect the discretiztion performed by the CCD and the fact that the set of pixel grey-levels is bounded above and below. Secondly, we assume that the point spread function is square; ie. either the blurring caused by the optics can be ignored and so $\omega_i(\mathbf{x}) = \delta(\mathbf{x})$ or we interpret super-resolution as "smaller pixels". These assumptions will be removed in the following sections.

Under these assumptions, and using knowledge that the registration is a translation, Equation (7) simplifies to:

$$\mathrm{Lo}_i(\mathbf{m}) \;=\; \sum_{\mathbf{p}} \frac{\mathrm{Hi}(\mathbf{p})}{M^2} \cdot \int_{\mathbf{p}} a_i \left(\frac{1}{M}\mathbf{z} + \mathbf{c}_i - \mathbf{m} \right) \mathrm{d}\mathbf{z}. \qquad (10)$$

To interpret this equation, remember that $a_i(\mathbf{z})$ is $\frac{1}{S_i^2}$ iff:

$$\mathbf{z} \;\in\; (-0.5 \cdot S_i, 0.5 \cdot S_i] \times (-0.5 \cdot S_i, 0.5 \cdot S_i] \qquad (11)$$

The integral in Equation (10) is $1/S_i^2$ times the area of the intersection of the two squares in Figure 2. We then have:

Theorem 1 *If $M \cdot S_i$ is an integer greater than 1, then for all choices of $\mathbf{c}_i$ the set of Equations (10) is not invertible. Moreover, the minimum achievable dimension of the null space is $(M \cdot S_i - 1)^2$. If $M \cdot S_i$ is not an integer, $\mathbf{c}_i$'s can be chosen such that the equations are invertible.*

Proof: We provide a proof only for 1D images since the extension to 2D is straight-forward, but messy.

The null space of Equations (10) is defined by the constraints $\sum_{\mathbf{p}} W'(\mathbf{m}, \mathbf{p}) \cdot \mathrm{Hi}(\mathbf{p}) \;=\; 0$ where $W'(\cdot, \cdot)$ is the

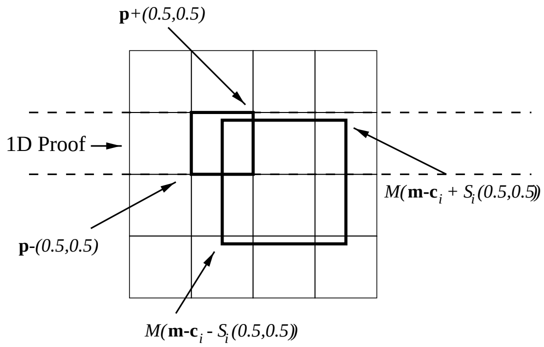


Figure 2: The integral in Equation (10) equals $1/S_i^2$ times the area of the intersection of the two highlighted squares.

area of intersection of the two squares in Figure 2. For 1D we just consider one row of the figure. By changing $\mathbf{c}_i$ to slide the large square along the row by some small amount, we immediately see that $\mathrm{Hi}(\mathbf{p})$ must equal both $\mathrm{Hi}(\mathbf{p} + (\lceil M \cdot S_i \rceil, 0)$ and $\mathrm{Hi}(\mathbf{p} + (\lfloor M \cdot S_i \rfloor, 0)$. If $M \cdot S_i$ is not an integer (or is 1), this proves that neighboring values of $\mathrm{Hi}(\mathbf{p})$ must be equal, and hence 0. If $M \cdot S_i$ is an integer this constraint places an upper bound of $M \cdot S_i - 1$ on the dimension of the null space (since the null space is contained in the set assignments to Hi that are periodic with period $M \cdot S_i$.) This value can also be shown to be a lower bound on the dimension of the null space by the space of assignments for which $\sum_{i=0}^{M \cdot S_i - 1} \mathrm{Hi}(\mathbf{p} + (i, 0)) = 0$. □

To validate this theorem, we solved the reconstruction constraints using gradient descent for the two cases $M = 2.0$ and $M = 1.5$, where $S_i = 1.0$. The results are presented in Figure 3. The input in both cases consisted of multiple down-sampled images similar to the one at the top of column 2 in Figure 1. As can be seen, for $M = 2.0$ the additive error is an approximately periodic image with period 2 pixels. For $M = 1.5$ the equations are invertible.

3.2 Real Valued Analysis, Arbitrary PSFs

Any linear system that is close to being not invertible is usually ill-conditioned. It is no surprise then that changing from a square point spread function to an arbitrary function $\mathrm{PSF}_i = a_i * \omega_i$ results in an ill-conditioned system:

Theorem 2 *Suppose $\omega_i(\mathbf{x})$ is an optical blurring function for which $\omega_i(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ and $\int \omega_i(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$. Then, the condition number of the linear system defined by replacing a_i with $a_i * \omega_i$ in Equation (10) is $\geq (M \cdot S_i)^2$.*

Proof: We first prove the result for a square point spread function and then generalize. The condition number of a linear operator A can be written as:

$$\mathrm{Cond}(A) \;=\; \frac{\sup_{\|\mathbf{x}\|_\infty = 1} \|A\mathbf{x}\|_\infty}{\inf_{\|\mathbf{x}\|_\infty = 1} \|A\mathbf{x}\|_\infty}. \tag{12}$$

From Equation (10), it follows that if $\mathrm{Hi}(\mathbf{p}) = 1$ for all $\mathbf{p}$, then $\mathrm{Lo}_i(\mathbf{m}) = 1$ for all $\mathbf{m}$. Hence, the numerator in Equation (12) is at least 1. Setting $\mathrm{Hi}(\mathbf{p})$ to be a checkerboard pattern (1 if $p+q$ is even, -1 if odd), we find that $|\mathrm{Lo}_i(\mathbf{m})| \leq 1/(M \cdot S_i)^2$, since the integration of the checkerboard over any square is $\in [-1, 1]$. (Proof omitted for brevity.) Hence, the denominator is at most $1/(M \cdot S_i)^2$.

(a) $M = 2.0$ (b) $M = 1.5$

Figure 3: Validation of Theorem 1: The results of solving the reconstruction constraints using gradient descent for a square point spread function with $S_i = 1.0$. (a) When $M \cdot S_i$ is an integer, the equations are not invertible and so a random periodic image in the null space is added to the original image. (b) When M is not an integer, the reconstruction constraints are invertible.

For arbitrary point spread functions, note that Equation (10) can be rewritten as:

$$\begin{aligned} \mathrm{Lo}_i(\mathbf{m}) &= \int_{\mathrm{Hi}} \frac{\mathrm{Hi}(\mathbf{z})}{M^2} \cdot a_i \left(\frac{1}{M}\mathbf{z} + \mathbf{c}_i - \mathbf{m} \right) \mathrm{d}\mathbf{z} \\ &= \left(a_i * \overline{\mathrm{Hi}} \right) (\mathbf{m} - \mathbf{c}_i) \end{aligned} \tag{13}$$

where we have changed variables $\mathbf{x} = \frac{1}{M}\mathbf{z}$, used the fact that a_i is even, and set $\overline{\mathrm{Hi}}(\mathrm{x}) = \mathrm{Hi}(M \cdot \mathrm{x})$. Both of the properties that we used for square point spread functions therefore also hold with a_i replaced by $a_i * \omega_i$ using standard properties of the convolution operator. □

If we could work with noiseless, real-valued quantities and could perform arbitrary precision arithmetic, then the fact that the reconstruction constraints are ill-conditioned would not be a problem. In reality, however, the low resolution images will be (intensity) discretized. There is therefore always noise in the measurements, even if it is only plus-or-minus half a grey-level. Before we present empirical results to validate Theorem 2, we prove a stronger version of it for quantized values.

3.3 Quantized Analysis, Arbitrary PSFs

Suppose that $\mathrm{int}[\cdot]$ denotes the quantization operator which takes a real-valued irradiance measurement and returns an integer-valued intensity in grey-levels. If we incorporate this quantization, Equation (13) becomes:

$$\mathrm{Lo}_i(\mathbf{m}) = \mathrm{int}\left[\int_{\mathrm{Hi}} \frac{\mathrm{Hi}(\mathbf{z})}{M^2} \cdot \mathrm{PSF}_i \left(\frac{\mathbf{z}}{M} + \mathbf{c}_i - \mathbf{m} \right) \mathrm{d}\mathbf{z} \right] \tag{14}$$

Suppose also that Hi is a finite size image with n pixels. We then have:

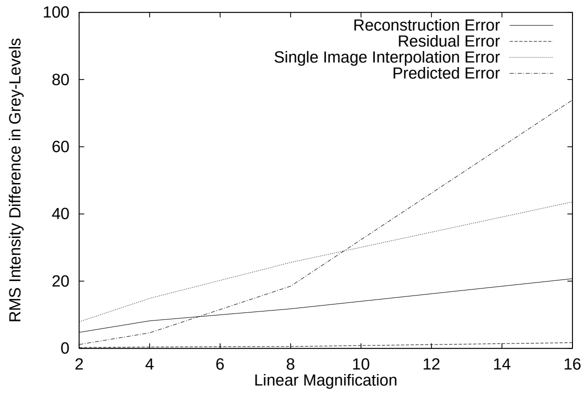


Figure 4: An illustration of Theorems 2 and 3 using the same inputs as in Figure 1. The reconstruction error is much higher than the residual, as would be expected for an ill-conditioned system. For low magnifications, the prior is unnecessary and so the results are worse than predicted. For high magnifications, the prior does help, but at the price of overly smooth results. (See Figure 1.)

Theorem 3 *If* $\mathrm{int}[\cdot]$ *is the standard rounding operator which replaces a real number with the nearest integer, then the volume of the set of solutions of Equation (14) asymptotically grows at least as fast as* $(M \cdot S_i)^{2 \cdot n}$ *(treating* n *as a constant and* M *and* S_i *as variables.)*

Proof: First note that the space of solutions is convex since the operator is linear. Next note that one solution of Equation (14) is the solution to:

$$\mathrm{Lo}_i(\mathbf{m}) - 0.5 = \int_{\mathrm{Hi}} \frac{\mathrm{Hi}(\mathbf{z})}{M^2} \cdot \mathrm{PSF}_i \left(\frac{\mathbf{z}}{M} + \mathbf{c}_i - \mathbf{m} \right) \mathrm{d}\mathbf{z} \tag{15}$$

The properties of the convolution give $0 \leq \mathrm{PSF}_i \leq 1/S_i^2$. Therefore, adding $(M \cdot S_i)^2$ to any pixel in Hi is still a solution since the right hand side of Equation (15) increases by at most 1. The volume of solutions therefore contains an n-dimensional simplex, where the angles at one vertex are all right-angles, and the sides are all $(M \cdot S_i)^2$ units long. The volume of such a simplex grows like $(M \cdot S_i)^{2n}$ (treating n as a constant and M and S_i as variables). □

In Figure 4 we present quantitative results to illustrate Theorems 2 and 3. We again used the reconstruction-based algorithm [Hardie *et al.*, 1997]. We verified our implementation in two ways: (1) we checked that for small magnification factors and no prior, our implementation does yield perfect reconstructions, and (2) for magnifications of 4, we checked that our numerical results agree with those in [Hardie *et al.*, 1997]. We also tried the related algorithm of [Schultz and Stevenson, 1996] and obtained similar results.

Using the same inputs as Figure 1, we plot the reconstruction error against the magnification; ie. the difference between the reconstructed high resolution image and the original. We compare this error with the residual error; ie. the difference between the low resolution inputs and their predictions from the reconstructed high resolution image. As expected for an ill-conditioned system, the reconstruction error is much higher than the residual. We also compare with a prediction of the reconstruction error obtained by multiplying the lower bound on the condition number ($M \cdot S_i^2$) by an estimate of the residual, assuming the grey-levels are discretized from a uniform distribution. For low magnification factors, this estimate is an underestimate because the the prior is unnecessary for noise free data; ie. better results would be obtained without the prior. On the other hand, for high magnifications the prediction is an over-estimate because the assumption of local smoothness does help the reconstruction. This assumption is at the expense of the overly smooth results in Figure 1.

We also plot interpolation results in Figure 4; ie. using just a single image reconstruction constraint. The difference between this curve and the reconstruction error curve is a measure of how much information the reconstruction constraints provide. Similarly, the difference between the reconstruction error and the predicted error is a measure of how much information the smoothness prior provides. For a magnification of 16, we see that the prior provides more information than the reconstruction constraints. This is the reason the results in Figure 1 are so smooth.

4 Class-Specific Recognition-Based Priors

Suppose it is possible to recognize an object (or part of an object) in the low resolution images. This additional information could then be incorporated into a super-resolution algorithm, and perhaps better results obtained. For example, if the image contains text data, OCR (optical character recognition) would provide strong constraints on the reconstructed image. In this section, we propose an algorithm to learn a *recognition-based* prior which can be used to improve the performance of super-resolution. (For lack of space, many of the details and results are omitted, but can be found in [Baker and Kanade, 1999].)

Our approach is closely related to that of [Freeman and Pasztor, 1999] who recently, and independently, proposed a learning framework for low-level vision, one application of which is image interpolation. Besides being applicable to an arbitrary number of images, the other major advantage of our approach is that it uses a prior that is both specific to the type (class) of object (in the "class-based" sense of [Riklin-Raviv and Shashua, 1999]) and a set of (local) recognition decisions. Our algorithm is also closely related to [Edwards *et al.*, 1998], in which the parameters of an "active-appearance" model are used for super-resolution.

4.1 Bayesian MAP Formulation

One way of incorporating a prior into super-resolution is to estimate the maximum *a posteriori* (MAP) solution: $\arg\max_{\mathrm{Hi}} \Pr[\mathrm{Hi} \,|\, \mathrm{Lo}_i]$. (See [Schultz and Stevenson, 1996], [Hardie *et al.*, 1997], and [Elad and Feuer, 1997].) Bayes law for this estimation problem is:

$$\Pr[\mathrm{Hi} \mid \mathrm{Lo}_i] \;=\; \frac{\Pr[\mathrm{Lo}_i \mid \mathrm{Hi}] \cdot \Pr[\mathrm{Hi}]}{\Pr[\mathrm{Lo}_i]}. \tag{16}$$

Since $\Pr[\mathrm{Lo}_i]$ is a constant because Lo_i is an input, and since the logarithm function is a monotonically increasing function, we have: $\arg\max_{\mathrm{Hi}} \Pr[\mathrm{Hi} \mid \mathrm{Lo}_i] =$

$$\arg\min_{\mathrm{Hi}} \left(-\ln \Pr[\mathrm{Lo}_i \mid \mathrm{Hi}] - \ln \Pr[\mathrm{Hi}]\right). \tag{17}$$

The first term in this expression $-\ln \Pr[\mathrm{Lo}_i \mid \mathrm{Hi}]$ is the (negative log) probability of reconstructing the low resolution images Lo_i, given that the high resolution image is Hi. It is therefore normally set to be a quadratic (energy) function of the reconstruction error in Equation (1).

4.2 Recognition-Based Priors

The second term $-\ln \Pr[\mathrm{Hi}]$ is the prior on the high resolution image. Usually $-\ln \Pr[\mathrm{Hi}]$ is chosen to be a smoothness prior. We would like to choose it to be a function of a set of recognition decisions. Suppose that the outputs of the recognition decisions partition the set of inputs Lo_i into a set of subclasses $\{C_k \mid k = 1, 2, \ldots\}$. We then define a *recognition-based* prior as follows:

$$\Pr[\mathrm{Hi}] \;=\; \sum_k \Pr[\mathrm{Hi} \mid \mathrm{Lo}_i \in C_k] \cdot \Pr[\mathrm{Lo}_i \in C_k]. \tag{18}$$

Once the low resolution inputs Lo_i are available, the recognition algorithm(s) can be applied, and it can be determined which subclass C_k the inputs lies in. The prior $\Pr[\mathrm{Hi}]$ then reduces to the more powerful prior $\Pr[\mathrm{Hi} \mid \mathrm{Lo}_i \in C_k]$.

4.3 Learning a Recognition-Based Prior

Suppose we have a set of high resolution training images T_i. We can compute their Gaussian $G_0(T_i), \ldots, G_N(T_i)$ and Laplacian $L_0(T_i), \ldots, L_N(T_i)$ pyramids, the horizontal $H_0(T_i), \ldots, H_N(T_i)$ and vertical $V_0(T_i), \ldots, V_N(T_i)$ first derivatives of the Gaussian pyramids, and the horizontal $H_0^2(T_i), \ldots, H_N^2(T_i)$ and vertical $V_0^2(T_i), \ldots, V_N^2(T_i)$ second derivatives of the Gaussian pyramids [Baker and Kanade, 1999]. We can then form a pyramid of feature vectors:

$$\mathbf{F}_j(T_i) = \left(L_j(T_i), H_j(T_i), V_j(T_i), H_j^2(T_i), V_j^2(T_i)\right) \tag{19}$$

for $j = 0, \ldots, N$.

Given a low resolution image Lo_i that is $M = 2^k$ times smaller than the training samples, we can compute the Gaussian pyramid from level k and upwards $G_k(\mathrm{Lo}_i), \ldots, G_N(\mathrm{Lo}_i)$. Similarly, we can compute the feature pyramids for those levels $\mathbf{F}_k(\mathrm{Lo}_i), \ldots, \mathbf{F}_N(\mathrm{Lo}_i)$. We know nothing at all, however, about the lower levels $\mathbf{F}_0(\mathrm{Lo}_i), \ldots, \mathbf{F}_{k-1}(\mathrm{Lo}_i)$, and in particular $\mathbf{F}_0(\mathrm{Lo}_i)$.

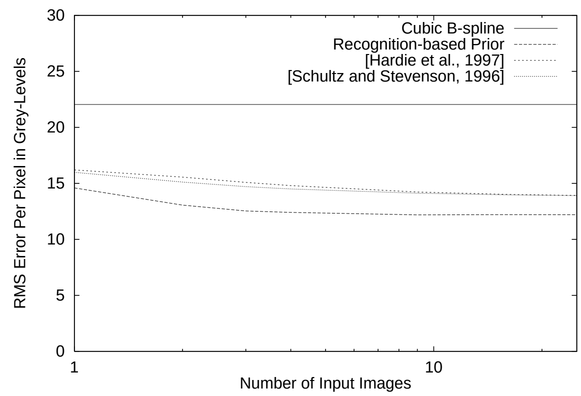


Figure 5: A comparison of our recognition-based algorithm with those of [Schultz and Stevenson, 1996] and [Hardie *et al.*, 1997]. Our algorithm out-performs these standard reconstruction-based algorithms across the entire range of number of input images.

Our recognition-based prior is based on an algorithm for predicting $\mathbf{F}_0(\mathrm{Lo}_i)$ from the training pyramids $\mathbf{F}_j(T_i)$. To describe the algorithm, we need one further piece of notation. If (m, n) is a pixel in the l^{th} level of a pyramid, its parent at the $l + 1^{\mathrm{th}}$ level is $(\lfloor \frac{m}{2} \rfloor, \lfloor \frac{n}{2} \rfloor)$. We therefore define the Parent Structure vector of a pixel (m, n) in the l^{th} level to be: $\mathbf{PS}_l(\mathrm{Lo}_i)(m, n) =$

$$\left(\mathbf{F}_l(\mathrm{Lo}_i)(m, n), \ldots, \mathbf{F}_N(\mathrm{Lo}_i)(\left\lfloor \frac{m}{2^{N-1}} \right\rfloor, \left\lfloor \frac{n}{2^{N-1}} \right\rfloor)\right). \tag{20}$$

We then use the following algorithm to predict $\overline{\mathbf{F}}_0(\mathrm{Lo}_i)$. (We use over-line to denote predicted values.) The algorithm operates by recognizing the closest matching training sample in the higher levels of the pyramid, and then copying the values for the lowest level from it.

Gradient Prediction Algorithm

For each pixel (m, n) in the bottom level of the feature pyramid to be predicted $\overline{\mathbf{F}}_0(\mathrm{Lo}_i)(m, n)$, do:

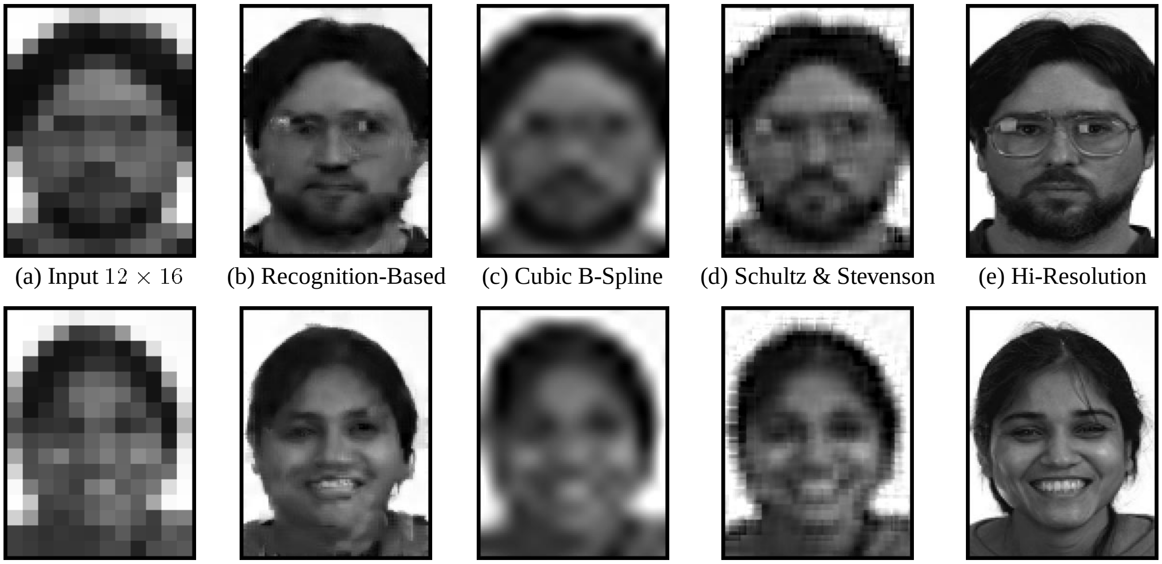
1. Find $j = \arg\min_l$

$$\left\| \mathbf{PS}_k(\mathrm{Lo}_i)(\left\lfloor \frac{m}{2^k} \right\rfloor, \left\lfloor \frac{n}{2^k} \right\rfloor) - \mathbf{PS}_k(T_l)(\left\lfloor \frac{m}{2^k} \right\rfloor, \left\lfloor \frac{n}{2^k} \right\rfloor) \right\|$$

2. Copy $\mathbf{F}_0(T_j)(m, n)$ into $\overline{\mathbf{F}}_0(\mathrm{Lo}_i)(m, n)$.

In this algorithm, Step 1. recognizes the closest matching training sample, and Step 2. copies the information about it into the lowest level of the feature pyramid for Lo_i.

Once $\overline{\mathbf{F}}_0(\mathrm{Lo}_i)$ has been estimated, the horizontal and vertical derivatives of the high resolution image ($\overline{H}_0(\mathrm{Hi})$ and $\overline{V}_0(\mathrm{Hi})$) can be predicted from it by extracting the derivatives of Lo_i using Equation (19) and accounting for the translation. The derivatives of Hi should equal these values. Parametric expressions for $H_0(\mathrm{Hi})$ and $V_0(\mathrm{Hi})$ can be derived in terms of the unknown pixels in the high resolution image Hi. We assume that the errors are



(a) Input 12×16 (b) Recognition-Based (c) Cubic B-Spline (d) Schultz & Stevenson (e) Hi-Resolution

(f) Input 12×16 (g) Recognition-Based (h) Cubic B-Spline (i) Schultz & Stevenson (j) Hi-Resolution

Figure 6: Selected results where the input consists of only three 12 x 16 pixel images. Note how high frequency features such as the eye-brows, lips, and nose are reconstructed even though there is almost no evidence for them in the input.

i.i.d. and Gaussian with covariance σ^2_{∇}. Therefore we set: $-\ln \Pr[\mathrm{Hi}] = K + \frac{1}{2\sigma^2_{\nabla}} \sum_{i,m,n}$

$$\begin{aligned} & \left[H_0(\mathrm{Hi})(m,n) - \overline{H}_0(\mathrm{Lo}_i)(m+c_i, n+d_i)\right]^2 \\ + \quad & \left[V_0(\mathrm{Hi})(m,n) - \overline{V}_0(\mathrm{Lo}_i)(m+c_i, n+d_i)\right]^2 \quad (21) \end{aligned}$$

where K is a constant that only depends upon σ^2_{∇} (and therefore can be ignored.) This prior is a recognition-based prior because it is a function of the recognition decisions made in Step 1. of the gradient prediction algorithm.

This algorithm is similar to the random texture synthesis algorithm of [De Bonet, 1997]. One major difference is that our algorithm is deterministic. It chooses the most likely values for the Parent Structure vector, rather than randomly sampling from a set of values. Another difference is that we take the class-based approach of [Riklin-Raviv and Shashua, 1999]. For each pixel (m, n), the algorithm therefore only looks at the corresponding pixels in the training samples since the image statistics will be a function of space. (For text data, where the image statistics are largely independent of the spatial location, we do actually consider all of the pixels in the training data.)

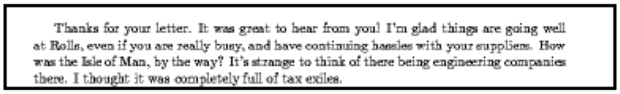
4.4 Experimental Results for Faces

Our experiments for faces were conducted with a subset of the FERET data set [Philips *et al.*, 1997] consisting of 596 images of 278 individuals (92 women and 186 men). The images (which are all frontal) must be aligned in the class-based approach so we can assume that the same part of the face appears in roughly the same part of the image [Riklin-Raviv and Shashua, 1999]. This alignment was performed by hand marking the location of 3 points; the centers of the eyes and the lower tip of the nose. We used a "leave-one-out" methodology to test our algorithm; for each image in the test set, we removed all images of that individual from the training set and then re-trained. Since this step is quite time consuming, we used a test set with 100 images chosen randomly from the FERET data.

We compared our algorithm with those of [Schultz and Stevenson, 1996] and [Hardie *et al.*, 1997]. In Figure 5 we plot the RMS pixel error of the algorithms against the number of images used. We also plot results for cubic B-spline interpolation for comparison. (Since cubic B-spline is an interpolation algorithm, only one image is used and so the performance is independent of the number of images.) In Figure 5 we see that our recognition-based algorithm does out-perform both of the other super-resolution algorithms. We present some example images in Figure 6, where the input consists of only three 12×16 pixel images. The recognition-based results are a huge improvement over both cubic B-spline interpolation and the Schultz and Stevenson algorithm [Schultz and Stevenson, 1996]. (The results for [Hardie *et al.*, 1997] are similar.) In particular, note how high resolution features such as the eyebrows and lips are recovered, even though there is little evidence for them in the input. Also try squinting at the images, a standard test of enhancement quality. Unlike the results in [Freeman and Pasztor, 1999], a marked difference can be seen between the input and the output.

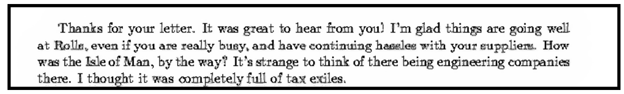
4.5 Experimental Results for Text Data

We also tried our algorithm on text data. We grabbed an image of an X-window displaying one page of a letter and down-sampled it. The image was split into disjoint train-



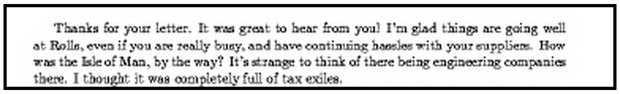
Thanks for your letter. It was great to hear from you! I'm glad things are going well at Rolls, even if you are really busy, and have continuing hassles with your suppliers. How was the Isle of Man, by the way? It's strange to think of there being engineering companies there. I thought it was completely full of tax exiles.

(a) Input Image. (Just one image is used.)



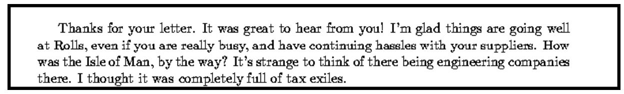
Thanks for your letter. It was great to hear from you! I'm glad things are going well at Rolls, even if you are really busy, and have continuing hassles with your suppliers. How was the Isle of Man, by the way? It's strange to think of there being engineering companies there. I thought it was completely full of tax exiles.

(b) Recognition-based, RMS Error 24.5



Thanks for your letter. It was great to hear from you! I'm glad things are going well at Rolls, even if you are really busy, and have continuing hassles with your suppliers. How was the Isle of Man, by the way? It's strange to think of there being engineering companies there. I thought it was completely full of tax exiles.

(c) Schultz and Stevenson, RMS Error 48.4



Thanks for your letter. It was great to hear from you! I'm glad things are going well at Rolls, even if you are really busy, and have continuing hassles with your suppliers. How was the Isle of Man, by the way? It's strange to think of there being engineering companies there. I thought it was completely full of tax exiles.

(d) Original High Resolution Image

Figure 7: The results of enhancing the resolution of some text by a factor of two using a single input. Our algorithm produces a sharp result using no explicit knowledge that the input is text (unlike the results in [Chiang and Boult, 1997] which assume step discontinuities.)

ing and test samples. (The training and test data therefore contain the same font, are at exactly the same scale, and the data is noiseless.) Our super-resolution results in Figure 7 are for a single input image and the magnification factor is 2. The recognition-based result in Figure 7(b) is by far the best reconstruction, both visually and in terms of the RMS grey-level pixel error (only 24.5 grey levels compared to over 48.4 for [Schultz and Stevenson, 1996].) The results for cubic B-spline and [Hardie *et al.*, 1997] are similar.

5 Discussion

We have shown that the super-resolution reconstruction constraints get weaker very rapidly as the magnification factor increases. The major cause is the averaging over the photosensitive area; ie. the fact that S_i is non-zero. This result means that there are fundamental limits on traditional reconstruction-based super-resolution algorithms. We have also shown that recognition algorithms can be embedded into the reconstruction process to enhance the performance of super-resolution algorithms; if the "scene" can be recognized it can be reconstructed far more accurately. Similar approaches may aid other (ie. 3D) reconstruction tasks.

References

[Baker and Kanade, 1999] S. Baker and T. Kanade. Hallucinating faces. Technical Report CMU-RI-TR-99-32, Robotics Institute, Carnegie Mellon University, 1999.

[Barbe, 1980] D.F. Barbe. *Charge-Coupled Devices*. Springer-Verlag, 1980.

[Bascle *et al.*, 1996] B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *4th ECCV*, pages 573–581, 1996.

[Born and Wolf, 1965] M. Born and E. Wolf. *Principles of Optics*. Permagon Press, 1965.

[Cheeseman *et al.*, 1994] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson. Super-resolved surface reconstruction from multiple images. Technical Report FIA-94-12, NASA Ames Research Center, 1994.

[Chiang and Boult, 1997] M.-C. Chiang and T.E. Boult. Local blur estimation and super-resolution. In *CVPR '97*, pages 821–826, 1997.

[De Bonet, 1997] J.S. De Bonet. Multiresolution sampling procedure for analysis and synthesis of texture images. In *SIGGRAPH '97*, pages 361–368, 1997.

[Dellaert *et al.*, 1998] F. Dellaert, S. Thrun, and C.E. Thorpe. Jacobian images of super-resolved texture maps for model-based motion estimation and tracking. In *4th Wkshp on Appl. of Computer Vision*, pages 2–7, 1998.

[Edwards *et al.*, 1998] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Learning to identify and track faces in image sequences. In *Third ICAFGR*, pages 260–265, 1998.

[Elad and Feuer, 1997] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy and undersampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–58, 1997.

[Elad and Feuer, 1999] M. Elad and A. Feuer. Super-resolution reconstruction of image sequences. *IEEE Trans. Pattern Anal. and Machine Intell.*, 21(9), 1999.

[Freeman and Pasztor, 1999] W. Freeman and E. Pasztor. Learning low-level vision. In *ICCV '99*, 1999.

[Hardie *et al.*, 1997] R.C. Hardie, K.J. Barnard, and E.E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Trans. on Im. Proc.*, 6(12):1621–1633, 1997.

[Huang and Tsai, 1984] T.S. Huang and R. Tsai. Multi-frame image restoration and registration. *Advances in Computer Vision and Image Proc.*, 1:317–339, 1984.

[Irani and Peleg, 1991] M. Irani and S. Peleg. Improving resolution by image restoration. *Computer Vision, Graphics, and Image Processing*, 53:231–239, 1991.

[Nalwa, 1993] V.S. Nalwa. *A Guided Tour of Computer Vision*. Addison-Wesley, 1993.

[Peleg *et al.*, 1987] S. Peleg, D. Keren, and L. Schweitzer. Improve image resolution using subpixel motion. *Pattern Recognition Letters*, pages 223–226, 1987.

[Philips *et al.*, 1997] P.J. Philips, H. Moon, P. Rauss, and S.A. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *CVPR '97*, 1997.

[Riklin-Raviv and Shashua, 1999] T. Riklin-Raviv and A. Shashua. The Quotient image: Class based recognition and synthesis under varying illumination. In *CVPR '99*, pages 566–571, 1999.

[Schultz and Stevenson, 1996] R. Schultz and R. Stevenson. Extraction of high-resolution frames from video sequences. *Trans. on Im. Proc.*, 5(6):996–1011, 1996.

[Shekarforoush *et al.*, 1996] H. Shekarforoush, M. Berthod, J. Zerubia, and M. Werman. Sub-pixel bayesian estimation of albedo and height. *IJCV*, 19(3), 1996.