Fast and Robust Multiframe Super Resolution

Sina Farsiu, M. Dirk Robinson, *Student Member, IEEE*, Michael Elad, and Peyman Milanfar, *Senior Member, IEEE*

***Abstract*—Super-resolution reconstruction produces one or a set of high-resolution images from a set of low-resolution images. In the last two decades, a variety of super-resolution methods have been proposed. These methods are usually very sensitive to their assumed model of data and noise, which limits their utility. This paper reviews some of these methods and addresses their shortcomings. We propose an alternate approach using L_1 norm minimization and robust regularization based on a bilateral prior to deal with different data and noise models. This computationally inexpensive method is robust to errors in motion and blur estimation and results in images with sharp edges. Simulation results confirm the effectiveness of our method and demonstrate its superiority to other super-resolution methods.**

***Index Terms*—Bilateral filter, deblurring, enhancement, image restoration, multiframe, regularization, robust estimation, super resolution, total variation (TV).**

I. Introduction

THEORETICAL and practical limitations usually constrain the achievable resolution of any imaging device. A dynamic scene with continuous intensity distribution $X(x,y)$ is seen to be warped at the camera lens because of the relative motion between the scene and camera. The images are blurred both by atmospheric turbulence and camera lens by continuous point spread functions $H_{atm}(x,y)$ and $H_{cam}(x,y)$. Then, they will be discretized at the CCD resulting in a digitized noisy frame $Y[m,n]$. We represent this forward model by the following:

$$Y[m,n] = [H_{cam}(x,y) ** F(H_{atm}(x,y) ** X(x,y))] \downarrow + V[m,n] \qquad (1)$$

in which $**$ is the two-dimensional convolution operator, F is the warping operator, $\downarrow$ is the discretizing operator, $V[m,n]$ is the system noise, and $Y[m,n]$ is the resulting discrete noisy and blurred image. Fig. 1 illustrates this equation.

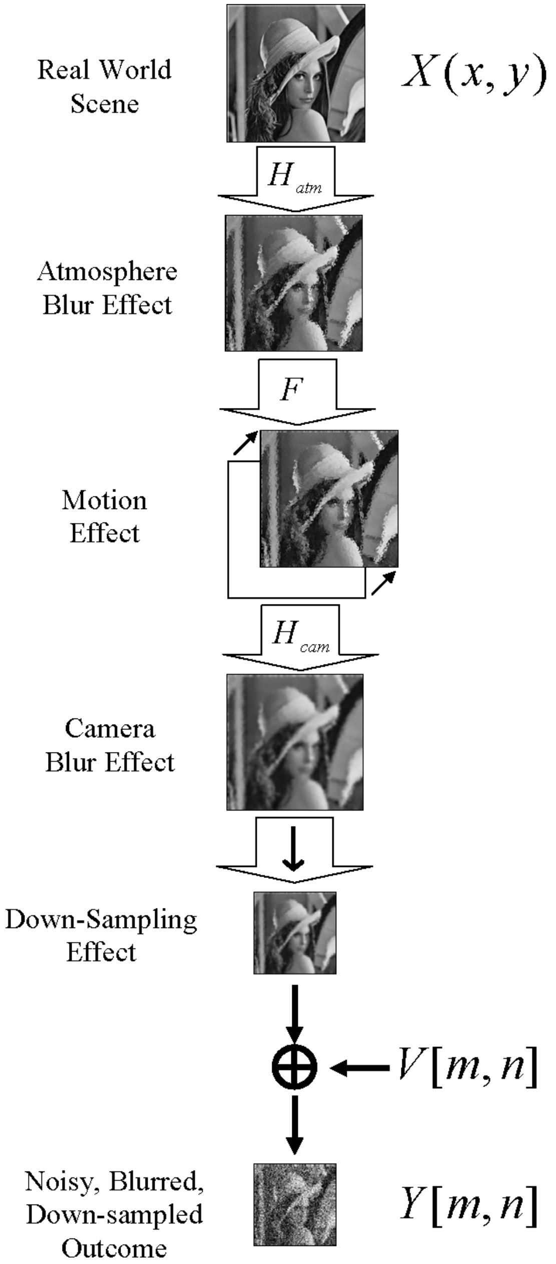


Fig. 1. Block diagram representation of (1), where $X(x,y)$ is the continuous intensity distribution of the scene, $V[m,n]$ is the additive noise, and $Y[m,n]$ is the resulting discrete low-quality image.

Super resolution is the process of combining a sequence of low-resolution (LR) noisy blurred images to produce a higher resolution image or sequence. The multiframe super-resolution problem was first addressed in [1], where they proposed a frequency domain approach, extended by others, such as [2]. Although the frequency domain methods are intuitively simple and computationally cheap, they are extremely sensitive to model errors [3], limiting their use. Also, by definition, only pure translational motion can be treated with such tools and even small

Manuscript received July 21, 2003; revised January 13, 2004. This work was supported in part by the National Science Foundation under Grant CCR-9984246, in part by the U.S. Air Force under Grant F49620–03-1-0387, and in part by the National Science Foundation Science and Technology Center for Adaptive Optics, managed by the University of California, Santa Cruz, under Cooperative Agreement AST-9876783. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Robert D. Nowak.

S. Farsiu, M. D. Robinson, and P. Milanfar are with the Electrical Engineering Department, University of California, Santa Cruz, CA 95064 USA (e-mail: farsiu@ee.ucsc.edu; dirkr@ee.ucsc.edu; milanfar@ee.ucsc.edu).

M. Elad is with the Computer Science Department, The Technion–Israel Institute of Technology, Haifa, Israel (e-mail: elad@cs.technion.ac.il).

Digital Object Identifier 10.1109/TIP.2004.834669

deviations from translational motion significantly degrade performance.

Another popular class of methods solves the problem of resolution enhancement in the spatial domain. Non-iterative spatial domain data fusion approaches were proposed in [4]–[6]. The iterative back-projection method was developed in papers such as [7] and [8]. In [9], the authors suggested a method based on the multichannel sampling theorem. In [10], a hybrid method, combining the simplicity of ML with proper prior information was suggested.

The spatial domain methods discussed so far are generally computationally expensive. The authors in [11] introduced a block circulant preconditioner for solving the Tikhonov regularized super-resolution problem formulated in [10] and addressed the calculation of regularization factor for the under-determined case by generalized cross validation in [12]. Later, a very fast super-resolution algorithm for pure translational motion and common space invariant blur was developed in [5]. Another fast spatial domain method was recently suggested in [13], where LR images are registered with respect to a reference frame defining a nonuniformly spaced high-resolution (HR) grid. Then, an interpolation method called Delaunay triangulation is used for creating a noisy and blurred HR image, which is subsequently deblurred. All of the above methods assumed the additive Gaussian noise model. Furthermore, regularization was either not implemented or it was limited to Tikhonov regularization. Considering outliers, [14] describes a very successful robust super-resolution method, but lacks the proper mathematical justification (limitations of this robust method and its relation to our proposed method are discussed in Appendix B). Finally, [15] and [16] have considered quantization noise resulting from video compression and proposed iterative methods to reduce compression noise effects in the super-resolved outcome.

The two most common matrix notations used to formulate the general super-resolution model of (1) represent the problem in the pixel domain. The more popular notation used in [5], [11], and [14] considers only camera lens blur and is defined as

$$\underline{Y}_k = D_k H_k^{cam} F_k \underline{X} + \underline{V}_k \quad k = 1, \dots, N \tag{2}$$

where the $[r^2M^2 \times r^2M^2]$ matrix F_k is the geometric motion operator between the HR frame $\underline{X}$ (of size $[r^2M^2 \times 1]$) and the kth LR frame $\underline{Y}_k$ (of size $[M^2 \times 1]$) which are rearranged in lexicographic order and r is the resolution enhancement factor. The camera's point spread function (PSF) is modeled by the $[r^2M^2 \times r^2M^2]$ blur matrix H_k^{cam}, and $[M^2 \times r^2M^2]$ matrix D_k represents the decimation operator. The $[M^2 \times 1]$ vector $\underline{V}_k$ is the system noise and N is the number of available LR frames.

Considering only atmosphere and motion blur, [13] recently presented an alternate matrix formulation of (1) as

$$\underline{Y}_k = D_k F_k H_k^{atm} \underline{X} + \underline{V}_k \quad k = 1, \dots, N. \tag{3}$$

In conventional imaging systems (such as video cameras), camera lens blur has a more important effect than the atmospheric blur (which is very important for astronomical images). In this paper, we use the model (2). Note that, under some assumptions which will be discussed in Section II-B, blur and motion matrices commute and the general matrix super-resolution formulation from (1) can be rewritten as

$$\begin{aligned}\underline{Y}_k &= D_k H_k^{cam} F_k H_k^{atm} \underline{X} + \underline{V}_k \\ &= D_k H_k^{cam} H_k^{atm} F_k \underline{X} + \underline{V}_k \quad k = 1, \dots, N.\end{aligned} \tag{4}$$

Defining $H_k = H_k^{cam} H_k^{atm}$ merges both models into a form similar to (2).

In this paper, we propose a fast and robust super-resolution algorithm using the L_1 norm, both for the regularization and the data fusion terms. Whereas the former is responsible for edge preservation, the latter seeks robustness with respect to motion error, blur, outliers, and other kinds of errors not explicitly modeled in the fused images. We show that our method's performance is superior to what was proposed earlier in [5], [11], [14], etc., and has fast convergence. We also mathematically justify a noniterative data fusion algorithm using a median operation and explain its superior performance.

This paper is organized as follows. Section II explains the main concepts of robust super resolution. Section II-B justifies using the L_1 norm to minimize the data error term; Section II-C justifies using our proposed regularization term. Section II-D combines the results of the two previous sections and explains our method and Section II-E proposes a faster implementation method. Simulations on both real and synthetic data sequences are presented in Section III, and Section IV concludes this paper.

II. Robust Super Resolution

A. Robust Estimation

Estimation of an unknown HR image is not exclusively based on the LR measurements. It is also based on many assumptions such as noise or motion models. These models are not supposed to be exactly true, as they are merely mathematically convenient formulations of some general prior information.

From many available estimators, which estimate a HR image from a set of noisy LR images, one may choose an estimation method which promises the optimal estimation of the HR frame, based on certain assumptions on data and noise models. When the fundamental assumptions of data and noise models do not faithfully describe the measured data, the estimator performance degrades. Furthermore, existence of outliers, which are defined as data points with different distributional characteristics than the assumed model, will produce erroneous estimates. A method which promises optimality for a limited class of data and noise models may not be the most effective overall approach. Often, suboptimal estimation methods which are not as sensitive to modeling and data errors may produce better and more stable results (robustness).

To study the effect of outliers, the concept of a breakdown point has been used to measure the robustness of an algorithm. The breakdown point is the smallest percentage of outlier contamination that may force the value of the estimate outside some range [17]. For instance, the breakdown point of the simple mean estimator is zero, meaning that one single outlier is sufficient to move the estimate outside any predicted bound. A robust estimator, such as the median estimator, may achieve a breakdown equal to 0.5, which is the highest value for breakdown points. This suggests that median estimation may not be affected by data sets in which outlier contaminated measurements form less that 50% of all data points.

A popular family of estimators are the ML-type estimators (M estimators) [18]. We rewrite the definition of these estimators in the super resolution context as the following minimization problem:

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \rho(\underline{Y}_k, D_k H_k F_k \underline{X}) \right] \tag{5}$$

or by an implicit equation

$$\sum_{k} \Psi(\underline{Y}_k, D_k H_k F_k \underline{X}) = 0 \tag{6}$$

where ρ is measuring the "distance" between the model and measurements and $\Psi(\underline{Y}_k, D_k H_k F_k \underline{X}) = (\partial/\partial \underline{X}) \rho(\underline{Y}_k, D_k H_k F_k \underline{X})$. The ML estimate of $\underline{X}$ for an assumed underlying family of exponential densities $f(\underline{Y}_k, D_k H_k F_k \underline{X})$ can be achieved when $\Psi(\underline{Y}_k, D_k H_k F_k \underline{X}) = -\log f(\underline{Y}_k, D_k H_k F_k \underline{X})$.

To find the ML estimate of the HR image, many papers such as [2], [5], and [11] adopt a data model such as (2) and model $\underline{V}_k$(additive noise) as white Gaussian noise. With this noise model, least-squares approach will result in the ML estimate [19]. The least-squares formulation is achieved when ρ is the L_2 norm of residual

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \|D_k H_k F_k \underline{X} - \underline{Y}_k\|_2^2 \right]. \tag{7}$$

For the special case of super resolution, based on [5], we will show in the next section, that least-squares estimation has the interpretation of being a nonrobust mean estimation. As a result, least squares-based estimation of a HR image, from a data set contaminated with non-Gaussian outliers, produces an image with visually apparent errors.

To appreciate this claim and study the visual effects of different sources of outliers in a video sequence, we set up the following experiments. In these experiments, four LR images were used to reconstruct a higher resolution image with two times more pixels in vertical and horizontal directions [a resolution enhancement factor of two using the least-squares approach (7)]. Fig. 2(a) shows the original HR image and Fig. 2(b) shows one of these LR images which has been acquired by shifting Fig. 2(a) in vertical and horizontal directions and subsampling it by factor of two (pixel replication is used to match its size with other pictures).

In the first experiment one of the four LR images contained affine motion with respect to the other LR images. If the model assumes translational motion, this results in a very common source of error when super resolution is applied to real data sequences, as the respective motion of camera and the scene are seldom pure translational. Fig. 2(c) shows this outlier image. Fig. 2(d) shows the effect of this error in the motion model (shadows around Lena's hat) when the non robust least-squares approach [5] is used for reconstruction.

To study the effect of non-Gaussian noise models, in the second experiment all four LR images were contaminated with salt and pepper noise. Fig. 2(e) shows one of these LR images and Fig. 2(f) is the outcome of the least-squares approach for reconstruction.

As the outlier effects are visible in the output results of least-squares-based super-resolution methods, it seems essential to find an alternative estimator. This new estimator should have the essential properties of robustness to outliers and fast implementation.

B. Robust Data Fusion

In Section II-A, we discussed the shortcomings of least squares-based HR image reconstruction. In this subsection, we study the family of L_p, $1 \leq p \leq 2$ norm estimators. We choose the most robust estimator of this family and show how implementation of this estimator requires minimum memory usage and is very fast.

The following expression formulates the L_p minimization criterion:

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \|D_k H_k F_k \underline{X} - \underline{Y}_k\|_p^p \right]. \tag{8}$$

Note that if $p = 2$, then (8) will be equal to (7).

Considering translational motion and with reasonable assumptions such as common space-invariant PSF, and similar decimation factor for all LR frames (i.e., $\forall k$ $H_k = H$ and $D_k = D$ which is true when all images are acquired with a unique camera), we calculate the gradient of the L_p cost. We will show that L_p norm minimization is equivalent to pixelwise weighted averaging of the registered frames. We calculate these weights for the special case of L_1 norm minimization and show that L_1 norm converges to median estimation which has the highest breakpoint value.

Since H and F_k are block circulant matrices, they commute ($F_k H = H F_k$ and $F_k^T H^T = H^T F_k^T$). Therefore, (8) may be rewritten as

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \|D F_k H \underline{X} - \underline{Y}_k\|_p^p \right]. \tag{9}$$

We define $\underline{Z} = H\underline{X}$. So, $\underline{Z}$ is the blurred version of the ideal HR image $\underline{X}$. Thus, we break our minimization problem in two separate steps:

1) finding a blurred HR image from the LR measurements (we call this result $\underline{\widehat{Z}}$);
2) estimating the deblurred image $\underline{\widehat{X}}$ from $\underline{\widehat{Z}}$.

Note that anything in the null space of H will not converge by the proposed scheme. However, if we choose an initialization that has no gradient energy in the null space, this will not pose a problem (see [5] for more details). As it turns out, the null space of H corresponds to very high frequencies, which are not part of our desired solution. Note that addition of an appropriate regularization term (Section II-C) will result in a well-posed problem with an empty null space. To find $\underline{\widehat{Z}}$, we substitute $H\underline{X}$ with $\underline{Z}$

$$\underline{\widehat{Z}} = \underset{\underline{Z}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \|D F_k \underline{Z} - \underline{Y}_k\|_p^p \right]. \tag{10}$$

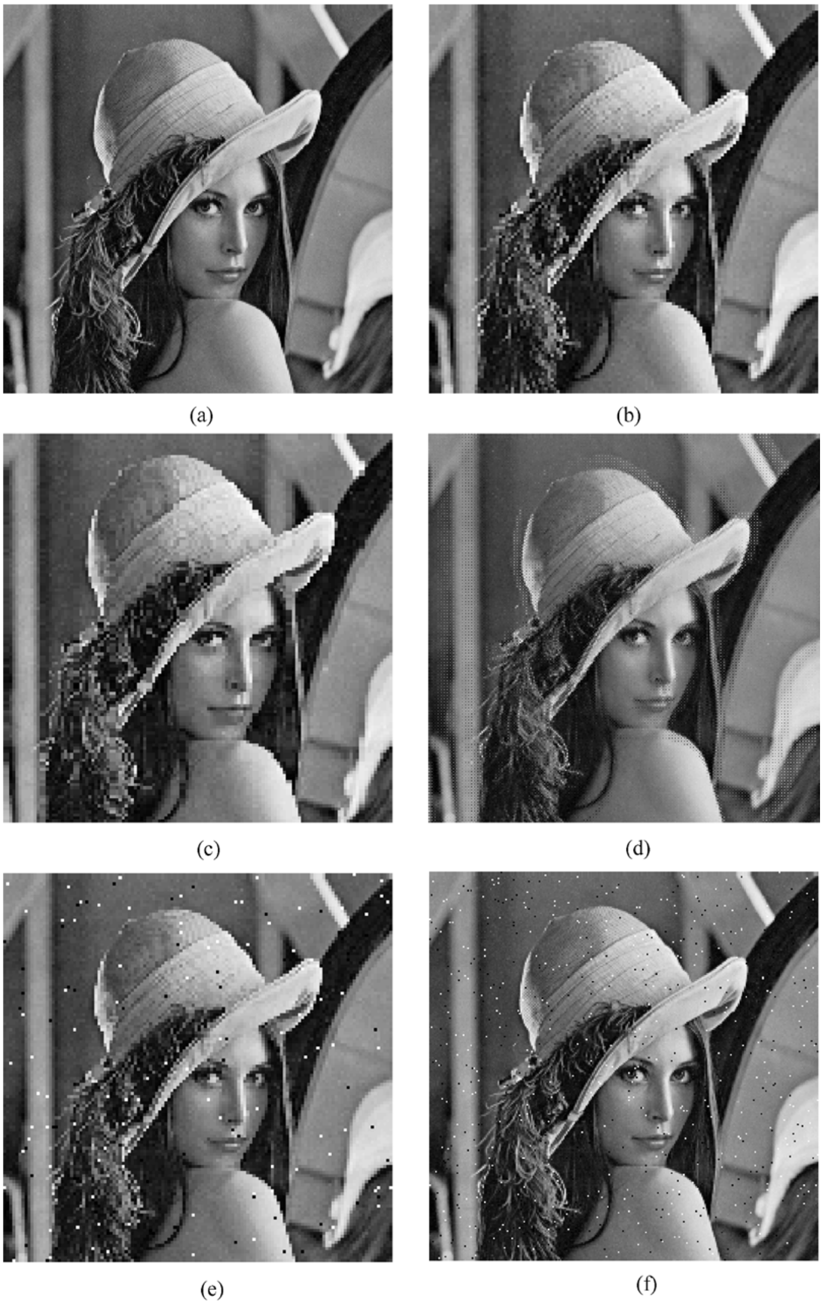

(a) (b) (c) (d) (e) (f)

Fig. 2. Simulation results of outlier effects on super-resolved images. The original HR image in (a) was warped with translational motion and down sampled resulting in four images such as (b). (c) Image acquired with downsampling and zoom (affine motion). (d) Reconstruction of these four LR images with least-squares approach. (e) One of four LR images acquired by adding salt and pepper noise to set of images in (b). (f) Reconstruction of images in (e) with least-squares approach. (a) Original HR frame. (b) LR frame. (c) LR Frame with zoom. (d) Least-squares result. (e) LR frame with salt and pepper outlier. (f) Least-squares result.

The gradient of the cost in (10) is

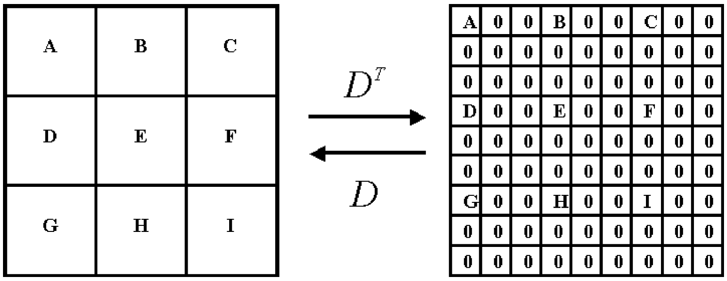
$$\begin{aligned}\underline{G}_p &= \frac{\partial}{\partial \underline{Z}}\left[\sum_{k=1}^{N} \|DF_k\underline{Z} - \underline{Y}_k\|_p^p\right]\\ &= \sum_{k=1}^{N} F_k^T D^T sign(DF_k\underline{Z} - \underline{Y}_k) \odot |DF_k\underline{Z} - \underline{Y}_k|^{p-1} \quad (11)\end{aligned}$$

where operator $\odot$ is the element-by-element product of two vectors.

The vector $\underline{\widehat{Z}}$ which minimizes the criterion (10) will be the solution to $\underline{G}_p = \underline{0}$. There is a simple interpretation for the solution: The vector $\underline{\widehat{Z}}$ is the weighted mean of all measurements at a given pixel, after proper zero filling and motion compensation.

To appreciate this fact, let us consider two boundary values of p. If $p = 2$, then

$$\underline{G}_2 = \sum_{k=1}^{N} F_k^T D^T (DF_k\underline{\widehat{Z}}_n - \underline{Y}_k) = \underline{0} \qquad (12)$$



A	B	C
D	E	F
G	H	I

A	0	0	B	0	0	C	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
D	0	0	E	0	0	F	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
G	0	0	H	0	0	I	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Fig. 3. Effect of upsampling D^T matrix on a 3 × 3 image and downsampling matrix D on the corresponding 9 × 9 upsampled image (resolution enhancement factor of three). In this figure, to give a better intuition, the image vectors are reshaped as matrices.

which is proved in [5] to be the pixelwise average of measurements after image registration. If $p = 1$ then the gradient term will be

$$\underline{G}_1 = \sum_{k=1}^{N} F_k^T D^T sign(DF_k\underline{\widehat{Z}} - \underline{Y}_k) = \underline{0}. \qquad (13)$$

We note that $F_k^T D^T$ copies the values from the LR grid to the HR grid after proper shifting and zero filling, and DF_k copies a selected set of pixels in HR grid back on the LR grid (Fig. 3 illustrates the effect of upsampling and downsampling matrices D^T, and D). Neither of these two operations changes the pixel values. Therefore, each element of $\underline{G}_1$, which corresponds to one element in $\underline{\widehat{Z}}$, is the aggregate of the effects of all LR frames. The effect of each frame has one of the following three forms:

1) addition of zero, which results from zero filling;
2) addition of $+1$, which means a pixel in $\underline{\widehat{Z}}$ was larger than the corresponding contributing pixel from frame $\underline{Y}_k$;
3) addition of -1, which means a pixel in $\underline{\widehat{Z}}$ was smaller than the corresponding contributing pixel from frame $\underline{Y}_k$.

A zero gradient state ($\underline{G}_1 = \underline{0}$) will be the result of adding an equal number of -1 and $+1$, which means each element of $\underline{\widehat{Z}}$ should be the median value of corresponding elements in the LR frames. $\underline{\widehat{X}}$, the final super-resolved picture, is calculated by deblurring $\underline{\widehat{Z}}$.

So far, we have shown that $p = 1$ results in pixelwise median and $p = 2$ results in pixelwise mean of all measurements after motion compensation. According to (11), if $1 < p < 2$, then both $sign(DF_k\underline{Z}_n - \underline{Y}_k)$ and $|DF_k\underline{Z}_n - \underline{Y}_k|^{p-1}$ terms appear in $\underline{G}_p$. Therefore, when the value of p is near one, $\underline{\widehat{Z}}$ is a weighted mean of measurements, with much larger weights around the measurements near the median value, while when the value of p is near two the weights will be distributed more uniformly.

In this subsection we studied $L_p, 1 \leq p \leq 2$ norm minimization family. As $p \longrightarrow 1$, this estimator takes the shape of median estimator, which has the highest breakpoint value, making it the most robust cost function. For the rest of this paper, we choose L_1 to minimize the measurement error[1] (note that we left out the study of $L_p, 0 \leq p < 1$ norm minimization family as they are not convex functions).

In the square or under-determined cases ($N = r^2$ and $N < r^2$ respectively), there is only one measurement available for each HR pixel. As median and mean operators for one or two measurements give the same result, L_1 and L_2 norm minimizations will result in identical answers. Also, in the under-determined cases, certain pixel locations will have no estimate at all. For these cases, it is essential for the estimator to have an extra term, called regularization term, to remove outliers. The next section discusses different regularization terms and introduces a robust and convenient regularization term.

C. Robust Regularization

Super resolution is an ill-posed problem [11], [21]. For the under-determined cases (i.e., when fewer than r^2 frames are available), there exist an infinite number of solutions which satisfy (2). The solution for square and over-determined cases is not stable, which means small amounts of noise in measurements will result in large perturbations in the final solution. Therefore, considering regularization in super-resolution algorithm as a means for picking a stable solution is very useful, if not necessary. Also, regularization can help the algorithm to remove artifacts from the final answer and improve the rate of convergence. Of the many possible regularization terms, we desire one which results in HR images with sharp edges and is easy to implement.

A regularization term compensates the missing measurement information with some general prior information about the desirable HR solution, and is usually implemented as a penalty factor in the generalized minimization cost function (5)

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\sum_{k=1}^{N} \rho(\underline{Y}_k, D_k H_k F_k \underline{X}) + \lambda \Upsilon(\underline{X}) \right] \qquad (14)$$

where λ, the regularization parameter, is a scalar for properly weighting the first term (similarity cost) against the second term (regularization cost) and Υ is the regularization cost function.

One of the most widely referenced regularization cost functions is the Tikhonov cost function [10], [11]

$$\Upsilon_T(\underline{X}) = \|\Gamma \underline{X}\|_2^2 \qquad (15)$$

where Γ is usually a highpass operator such as derivative, Laplacian, or even identity matrix. The intuition behind this regularization method is to limit the total energy of the image (when Γ is the identity matrix) or forcing spatial smoothness (for derivative or Laplacian choices of Γ). As the noisy and edge pixels both contain high-frequency energy, they will be removed in the regularization process and the resulting denoised image will not contain sharp edges.

[1] L_1 norm minimization is the ML estimate of data in the presence of Laplacian noise. The statistical analysis presented in [20] justifies modeling the super-resolution noise in the presence of different sources of outliers as Laplacian probability density function (PDF) rather than Gaussian PDF.

Certain types of regularization cost functions work efficiently for some special types of images but are not suitable for general images (such as maximum entropy regularizations which produce sharp reconstructions of point objects, such as star fields in astronomical images [22]).

One of the most successful regularization methods for denoising and deblurring is the total variation (TV) method [23]. The TV criterion penalizes the total amount of change in the image as measured by the L_1 norm of the magnitude of the gradient and is defined as

$$\Upsilon_{TV}(\underline{X}) = \|\nabla \underline{X}\|_1$$

where ∇ is the gradient operator. The most useful property of TV criterion is that it tends to preserve edges in the reconstruction [22]–[24], as it does not severely penalize steep local gradients.

Based on the spirit of TV criterion, and a related technique called the bilateral filter (Appendix A), we introduce our robust regularizer called bilateral TV, which is computationally cheap to implement, and preserves edges. The regularizing function looks like

$$\Upsilon_{BTV}(\underline{X}) = \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|} \|\underline{X} - S_x^l S_y^m \underline{X}\|_1 \tag{16}$$

where matrices (operators) S_x^l, and S_y^k shift $\mathbf{X}$ by l, and k pixels in horizontal and vertical directions respectively, presenting several scales of derivatives. The scalar weight α, $0<\alpha<1$, is applied to give a spatially decaying effect to the summation of the regularization terms.

It is easy to show that this regularization method is a generalization of other popular regularization methods. If we limit m, l to the two cases of $m=1, l=0$ and $m=0, l=1$ with $\alpha=1$, and define operators Q_x and Q_y as representatives of the first derivative ($Q_x = I - S_x$ and $Q_y = I - S_y$) then (16) results in

$$\Upsilon_{BTV}(\underline{X}) = \|Q_x \underline{X}\|_1 + \|Q_y \underline{X}\|_1 \tag{17}$$

which is suggested in [25] as a reliable and computationally efficient approximation to the TV prior [23].

To compare the performance of bilateral TV ($P \geq 1$) to common TV prior ($P=1$), we set up the following denoising experiment. We added Gaussian white noise of mean zero and variance 0.045 to the image in Fig. 4(a) resulting in the noisy image of Fig. 4(b). If $\underline{X}$ and $\underline{Y}$ represent the original and corrupted images then following (14), we minimized

$$\widehat{\underline{X}} = \underset{\underline{X}}{\operatorname{ArgMin}} \left[\|\underline{Y} - \underline{X}\|_2^2 + \lambda \Upsilon(\underline{X}) \right] \tag{18}$$

to reconstruct the noisy image. Tikhonov denoising resulted in Fig. 4(c), where Γ in (15) was replaced by matrix realization of the Laplacian kernel

$$\Gamma_{\text{kernel}} = \frac{1}{8} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{19}$$

Although a relatively large regularization factor ($\lambda = 4.5$) was chosen for this reconstruction which resulted in the loss of sharp edges, yet the noise has not been removed efficiently. The result of using TV prior ($P = 1$, $\lambda = 0.009$) for denoising is shown in Fig. 4(d). Fig. 4(e) shows the result of applying bilateral TV prior ($P = 3$, $\lambda = 0.009$). [2] Notice the effect of each reconstruction method on the pixel indicated by an arrow in Fig. 4(a). As this pixel is surrounded by nonsimilar pixels, TV prior considers it as a heavily noisy pixel and uses the value of immediate neighboring pixels to estimate its original value. On the other hand, bilateral TV considers a larger neighborhood. By bridging over immediate neighboring pixels, the value of similar pixels are also considered in graylevel estimation of this pixel, therefore the smoothing effect in Fig. 4(e) is much less than Fig. 4(d). Fig. 4(f) compares the performance of TV and bilateral TV denoising methods in estimating graylevel value of the arrow indicated pixel. Unlike bilateral TV regularization, increasing the number of iterations in Tikhonov and TV regularizations will result in more undesired smoothing. This example demonstrates the tendency of other regularization functionals to remove point like details from the image. The proposed regularization not only produces sharp edges but also retains point like details.

To compare the performance of our regularization method to the Tikhonov regularization method, we set up another experiment. We corrupted an image by blurring it with a Gaussian blur kernel followed by adding Gaussian additive noise. We reconstructed the image using Tikhonov and our proposed regularization terms (this scenario can be thought of as a super-resolution problem with resolution factor of one). If $\underline{X}$ and $\underline{Y}$ represent

$$\widehat{\underline{X}} = \underset{\underline{X}}{\operatorname{ArgMin}} \left[\sum_{k=1}^{N} \|D_k H_k F_k \underline{X} - \underline{Y}_k\|_1 + \lambda \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|} \|\underline{X} - S_x^l S_y^m \underline{X}\|_1 \right]. \tag{21}$$

[2] The criteria for parameter selection in this example (and other examples discussed in this paper) was to choose parameters which produce visually most appealing results. Therefore, to ensure fairness, each experiment was repeated several times with different parameters and the best result of each experiment was chosen as the outcome of each method. Fig. 4(c) is an exception, where we show that Tikhonov regularization fails to effectively remove noise even with a very large regularization factor.

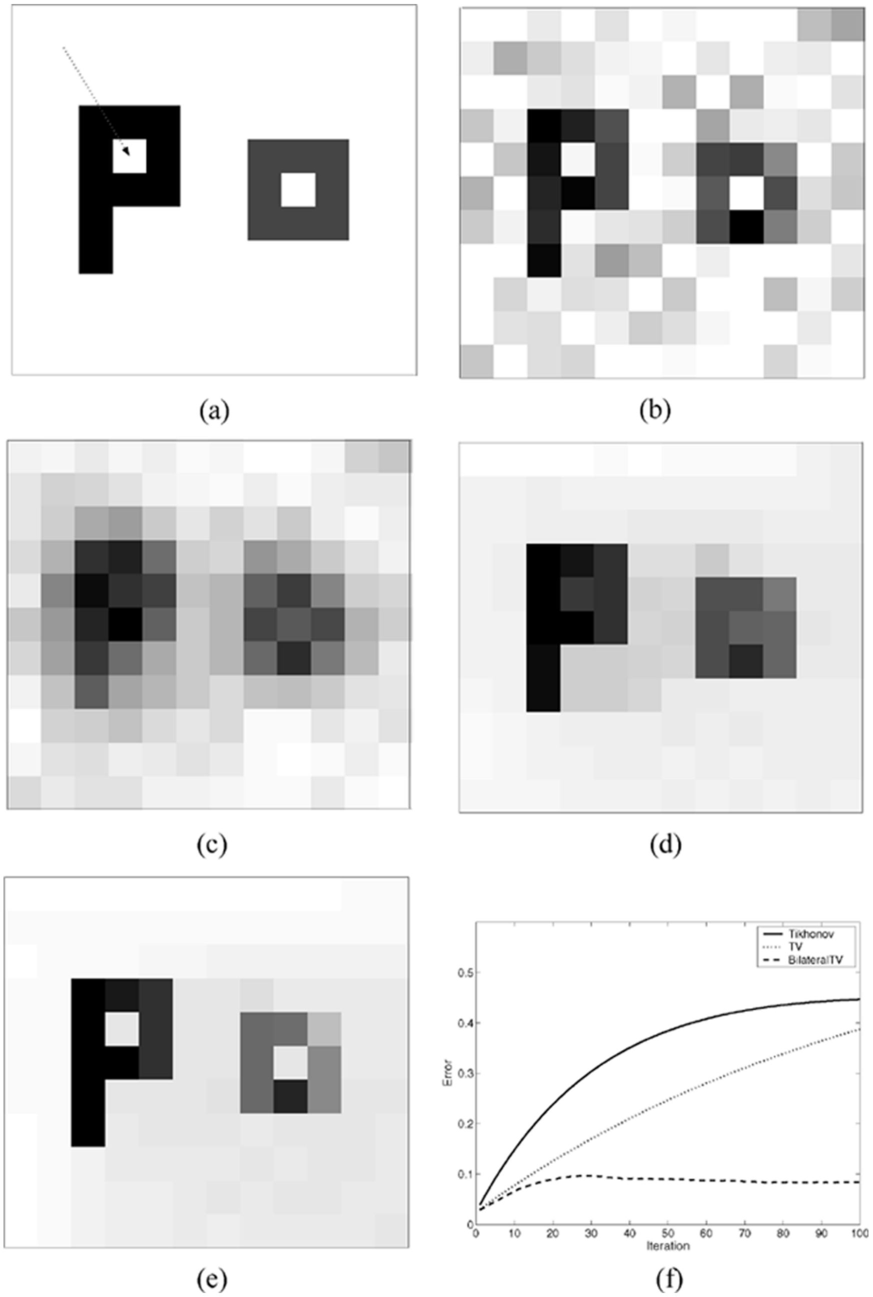


Fig. 4. (a)-(e) Simulation results of denoising using different regularization methods. (a) Original. (b) Noisy. (c) Reconstruction using Tikhonov. (d) Reconstruction using TV. (e) Reconstruction using bilateral TV. (f) Error in gray-level value estimation of the pixel indicated by arrow in (a) versus the iteration number in Tikhonov (solid line), TV (dotted line), and bilateral TV (broken line) denoising.

the original and corrupted images and H represents the matrix form of the blur kernel then following (14), we minimized

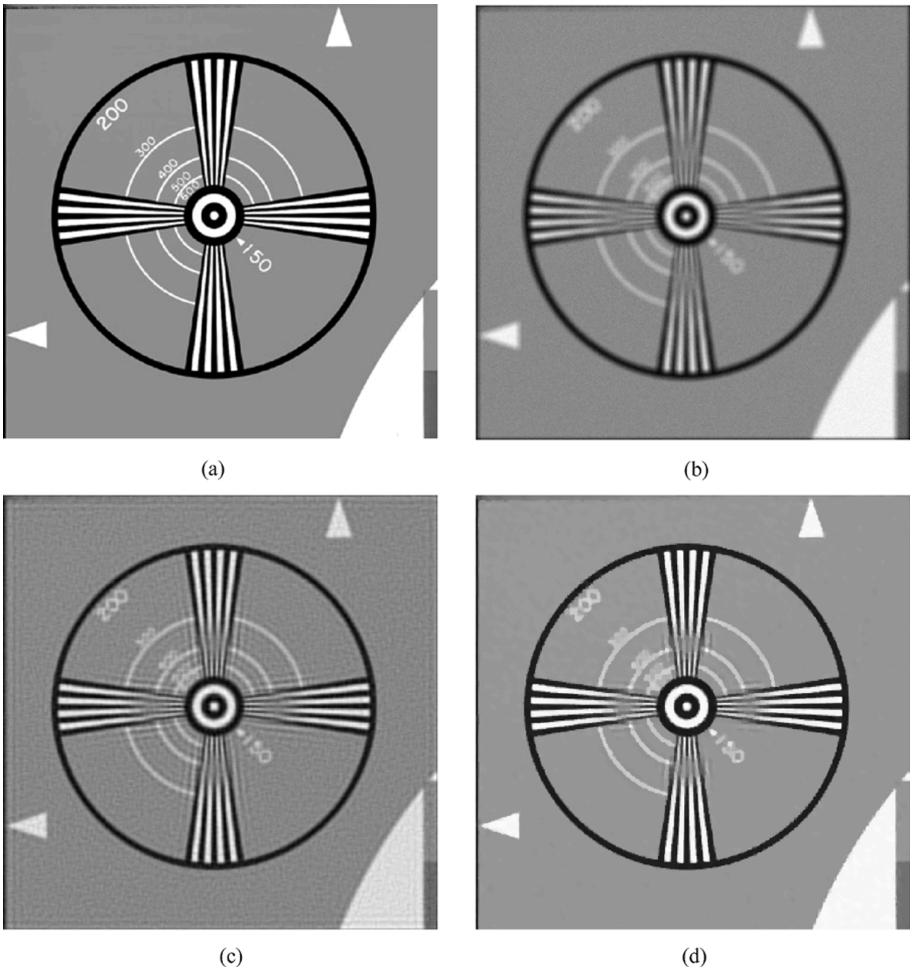
$$\underline{\widehat{X}} = \underset{\underline{X}}{\operatorname{ArgMin}} \left[\|\underline{Y} - H\underline{X}\|_2^2 + \lambda \Upsilon(\underline{X}) \right] \tag{20}$$

to reconstruct the blurred noisy image.

Fig. 5 shows the results of our experiment. Fig. 5(a) shows the original image ($\underline{X}$). Fig. 5(b) is the corrupted $\underline{Y} = H\underline{X} + \underline{V}$, where $\underline{V}$ is the additive noise. Fig. 5(c) is the result of reconstruction with Tikhonov regularization, where Γ in (15) was replaced by the Laplacian kernel (19) and $\lambda = 0.03$. Fig. 5(d) shows the result of applying our regularization criterion (16) with the following parameters $\alpha = 0.7$, $\lambda = 0.17$ and $P = 2$. The best mean-square error (MSE) achieved by Tikhonov regularization was 313 versus 215 for the proposed regularization. The superior edge preserving property of the bilateral prior is apparent in this example.

D. Robust Super-Resolution Implementation

In this subsection, based on the material that was developed in Sections II-B and C, a solution for the robust super-resolution problem will be proposed. Combining the ideas presented thus far, we propose the robust solution of the super-resolution problem as follows [shown in (21), at the bottom of the previous



(a) (b)

(c) (d)

Fig. 5. Simulation results of deblurring using different regularization methods. The mean square error (MSE) of reconstructed image using Tikhonov regularization (c) was 313. The MSE of reconstructed image using bilateral TV (d) was 215. (a) Original. (b) Blurred and noisy. (c) Best Tikhonov regularization. (d) Proposed regularization.

page]. We use steepest descent to find the solution to this minimization problem

$$\underline{\widehat{X}}_{n+1} = \underline{\widehat{X}}_n - \beta\Big\{ \sum_{k=1}^{N} F_k^T H_k^T D_k^T sign(D_k H_k F_k \underline{\widehat{X}}_n - \underline{Y}_k) + \lambda \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|}[I - S_y^{-m} S_x^{-l}] sign(\underline{\widehat{X}}_n - S_x^l S_y^m \underline{\widehat{X}}_n)\Big\} \tag{22}$$

where β is a scalar defining the step size in the direction of the gradient. S_x^{-l} and S_y^{-m} define the transposes of matrices S_x^l and S_y^m respectively and have a shifting effect in the opposite directions as S_x^l and S_y^m.

Simulation results in Section III will show the strength of the proposed algorithm. The matrices F, H, D, S, and their transposes can be exactly interpreted as direct image operators such as shift, blur, and decimation [26]. Noting and implementing the effects of these matrices as a sequence of operators spares us from explicitly constructing them as matrices. This property helps our method to be implemented in an extremely fast and memory efficient way.

Fig. 6 is the block diagram representation of (22). There, each LR measurement Y_k will be compared to the warped, blurred, and decimated current estimate of HR frame X_n. Block G_k represents the gradient back projection operator that compares the kth LR image to the estimate of the HR image in the nth steepest descent iteration. Block $R_{m,l}$ represents the gradient of regularization term, where the HR estimate in the nth steepest descent iteration is compared to its shifted version (l pixel shift in horizontal and m pixel shift in vertical directions).

Details of the blocks G_k and $R_{m,l}$ are defined in Fig. 7(a) and (b). Block $T(PSF)$ in Fig. 7(a) replaces the matrix H_k^T with a simple convolution. Function T flips the columns of PSF kernel in the left-right direction (that is, about the vertical axis), and then flips the rows of PSF kernel in the up-down direction (that is, about the horizontal axis).[3] The D_k^T up-sampling block in Fig. 7(a) can be easily implemented by filling $r-1$ zeros both in vertical and horizontal directions around each pixel (Fig. 3). And, finally, the F_k^T shift-back block in Fig. 7(a), is implemented by inverting the translational motion in the reverse direction. Note that even for the more general affine motion model

[3] If the PSF kernel has even dimensions, one extra row or column of zeros will be added to it to make it odd size (zero columns and rows have no effect in convolution process).

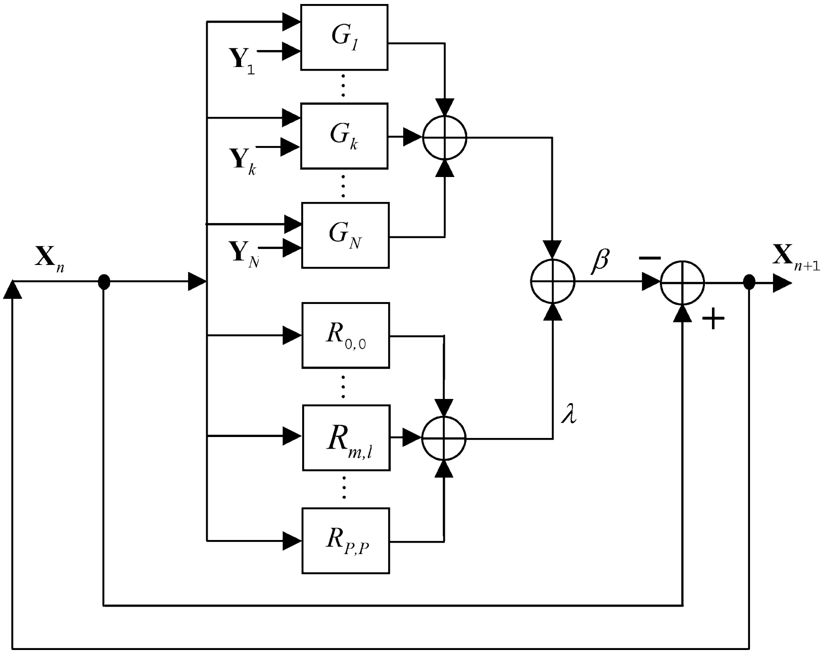
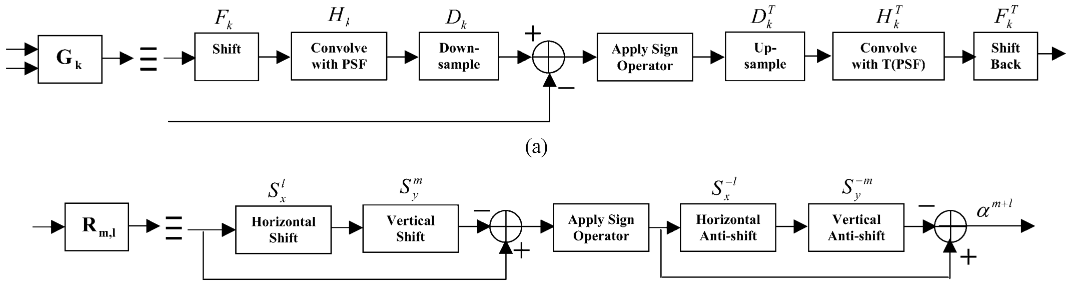


Fig. 6. Block diagram representation of (22), blocks G_k, and $R_{m,l}$ are defined in Fig. 7.



(a)

(b)

Fig. 7. Extended block diagram representation of G_k and $R_{m,l}$ blocks in Fig. 6. (a) Block diagram representation of similarity cost derivative (G_k). (b) Block diagram representation of regularization cost derivative.

a similar inverting property (though more complicated) is still valid.

Parallel processing potential of this method, which significantly increases the overall speed of implementation, can be easily interpreted from Fig. 6 (the computation of each G_k or $R_{l,m}$ blocks may be assigned to a separate processor).

Our robust super-resolution approach also has an advantage in the computational aspects over other methods including the one proposed in [14]. In our method, an inherently robust cost function has been proposed, for which a number of computationally efficient numerical minimization methods[4] are applicable. On the contrary, [14] uses steepest descent method to minimize the nonrobust L_2 norm cost function, and robustness is achieved by modifying the steepest descent method, where median operator is used in place of summation operator in computing the gradient term of (12). Implementing the same scheme of substituting summation operator with median operator in computationally more efficient methods such as conjugate gradient is not a straightforward task and besides it is no longer guaranteed that the modified steepest descent and conjugate gradient minimization converge to the same answer.

As an example, Fig. 8(a) and (b) show the result of implementing the proposed method on the same image sets that was used to generate Fig. 2(d) and (f), respectively. The outlier effects have been reduced significantly (more detailed examples are presented in Section III).

In the next section, we propose an alternate method to achieve further improvements in computational efficiency.

E. Fast Robust Super-Resolution Formulation

In Section II-D, we proposed an iterative robust super-resolution method based on (21). Although implementation of (21) is very fast,[5] for real-time image sequence processing, faster methods are always desirable. In this subsection, based on the interpretation of (13) that was offered in Section II-B, we simplify (21) to achieve a faster method.

In this method, resolution enhancement is broken into two consecutive steps:

1) noniterative data fusion;
2) iterative deblurring-interpolation.

[4] Such as conjugate gradient (CG), preconditioned conjugate gradient (PCG), Jacobi, and many others.

[5] Computational complexity and memory requirement is similar to the method proposed in [8].

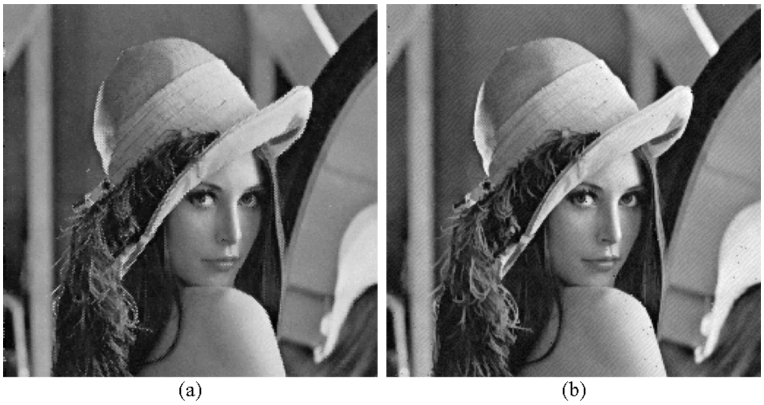


Fig. 8. Reconstruction of the outlier contaminated image in Fig. 2 using (22). (a) Robust reconstruction of the same image that was used to produce Fig. 2(d) and (b) is the robust reconstruction of the same image that was used to produce Fig. 2(f).

As we described in Section II-B, registration followed by the median operation (what we call median shift and add) results in $\underline{\widehat{Z}} = H\underline{\widehat{X}}$. Usage of median operator for fusing LR images is also suggested in [4] and [6].

The goal of the deblurring-interpolation step is finding the deblurred HR frame $\underline{\widehat{X}}$. Note that for the under-determined cases, not all $\underline{\widehat{Z}}$ pixel values can be defined in the data fusion step, and their values should be defined in a separate interpolation step. In this paper, unlike [4], [6] and [13], interpolation and deblurring are done simultaneously.

The following expression formulates our minimization criterion for obtaining $\underline{\widehat{X}}$ from $\underline{\widehat{Z}}$

$$\underline{\widehat{X}} = \underset{\underline{X}}{\text{ArgMin}} \left[\|A(H\underline{X}-\underline{\widehat{Z}})\|_1 + \lambda' \underbrace{\sum_{l=-P}^{P}\sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|}\|\underline{X}-S_x^l S_y^m \underline{X}\|_1 \right] \quad (23)$$

where matrix A is a diagonal matrix with diagonal values equal to the square root of the number of measurements that contributed to make each element of $\underline{\widehat{Z}}$ (in the square case A is the identity matrix). So, the undefined pixels of $\underline{\widehat{Z}}$ have no effect on the HR estimate $\underline{\widehat{X}}$. On the other hand, those pixels of $\underline{\widehat{Z}}$ which have been produced from numerous measurements, have a stronger effect in the estimation of the HR frame $\underline{\widehat{X}}$.

As A is a diagonal matrix, $A^T = A$, and the corresponding steepest descent solution of minimization problem (23) can be expressed as

$$\underline{\widehat{X}}_{n+1} = \underline{\widehat{X}}_n - \beta\Big\{H^T A^T sign(AH\underline{\widehat{X}}_n - A\underline{\widehat{Z}}) + \lambda' \underbrace{\sum_{l=-P}^{P}\sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|}[I - S_y^{-m}S_x^{-l}]sign(\underline{\widehat{X}}_n - S_x^l S_y^m \underline{\widehat{X}}_n)\Big\}. \quad (24)$$

Decimation and warping matrices (D and F) and summation of measurements are not present anymore, which makes the implementation of (24) much faster than (22). Note that physical construction of matrix A is not necessary as it can be implemented as a mask matrix with the size equal to image $\mathbf{X}$.

III. Experiments

In this section, we compare the performance of the resolution enhancement algorithms proposed in this paper to existing resolution enhancement methods. The first example[6] is a controlled simulated experiment. In this experiment, we create a sequence of LR frames by using one HR image [Fig. 9(a)]. First, we shifted this HR image by a pixel in the vertical direction. Then, to simulate the effect of camera PSF, this shifted image was convolved with a symmetric Gaussian low-pass filter of size 4×4 with standard deviation equal to one. The resulting image was subsampled by the factor of 4 in each direction. The same approach with different motion vectors (shifts) in vertical and horizontal directions was used to produce 16 LR images from the original scene. We added Gaussian noise to the resulting LR frames to achieve signal-to-noise ratio (SNR) equal[7] to 18 dB. One of these LR frames is presented in Fig. 9(b). To simulate the errors in motion estimation, a bias equal to one pixel shift in the LR grid was intentionally added to the known motion vectors of three LR frames.

The result of implementing the noniterative resolution enhancement method described in [5] is shown in Fig. 9(c). It is not surprising to see the motion error artifacts in the HR frame as the HR image is the result of zero filling, shifting, and adding the LR measurements. Deblurring this result with Wiener method [Fig. 9(d)] does not remove these artifacts, of course. For reference, Fig. 9(e) shows the result of applying an iterative method based on minimizing L_2 norm, both for the residual and the regularization terms. The following equation describes this minimization criterion:

$$\underline{\widehat{X}} = \text{ArgMin}\left[\sum_{k=1}^{N}\|D_k H_k F_k \underline{X} - \underline{Y}_k\|_2^2 + \lambda\|\Gamma\underline{X}\|_2^2\right] \quad (25)$$

in which Γ is defined in (19) and regularization factor λ was chosen to be 0.4. As L_2 norm is not robust to motion error, motion artifacts are still visible in the result. Note that the relatively high regularization factor which was chosen to reduce the motion artifact has resulted in a blurry image.

[6] This paper (with all pictures and a MATLAB-based software package for resolution enhancement) is available at http://www.ee.ucsc.edu/~milanfar.

[7] SNR is defined as $10\log_{10}(\sigma^2/\sigma_n^2)$, where σ^2, σ_n^2 are variance of a clean frame and noise, respectively.

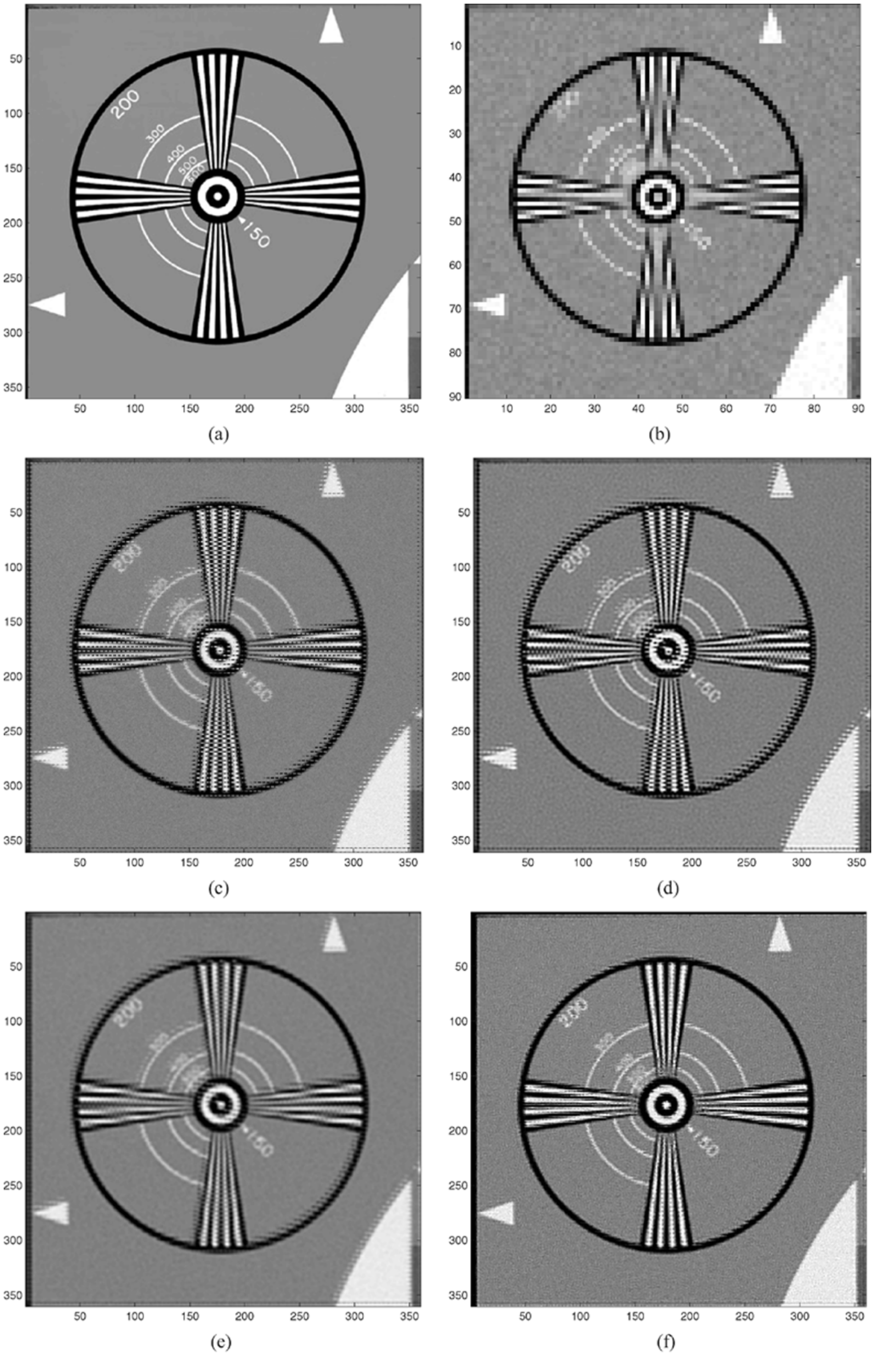
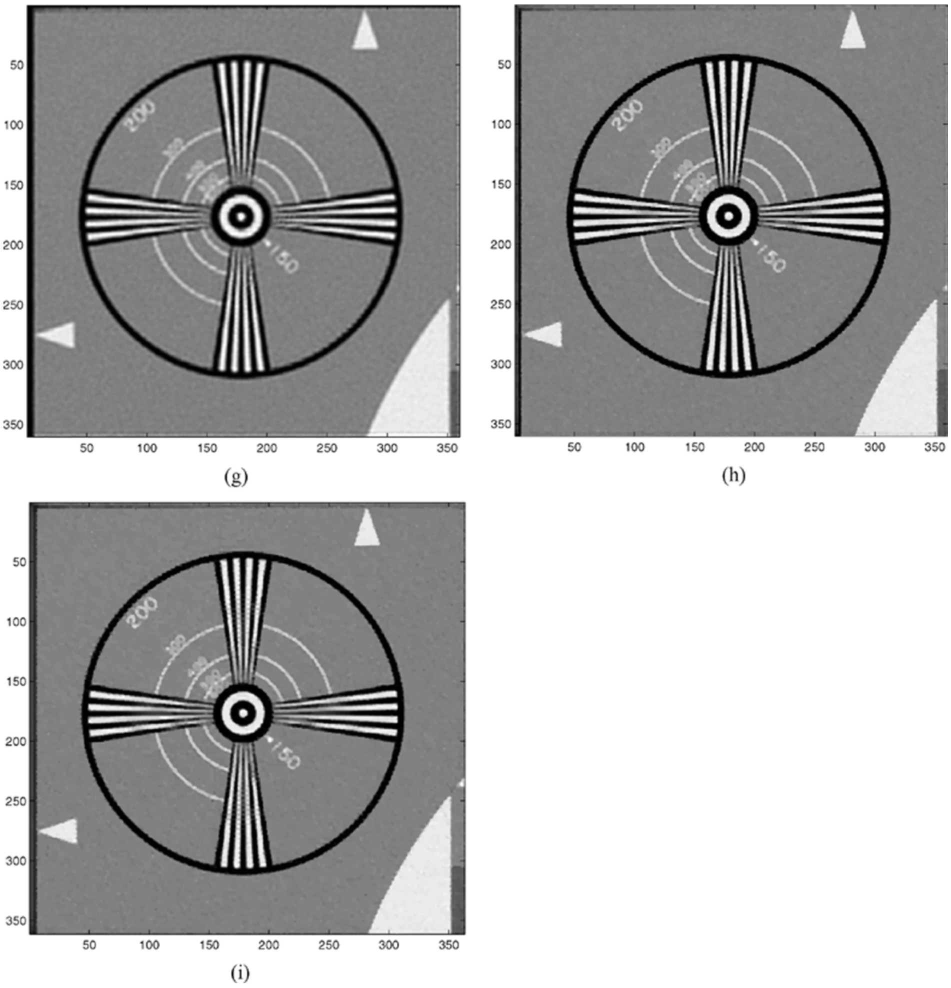

(a) (b) (c) (d) (e) (f)

Fig. 9. Simulation results of different resolution enhancement methods are applied to the (a). (a) Original HR frame. (b) LR frame. (c) Shift and add result [5]. (d) Deconvolved shift and add [5]. (e) $L_2 +$ Tikhonov. (f) Zomet method [14].

The robust super-resolution method which was proposed in [14] resulted in Fig. 9(f). Fig. 9(g) was obtained by simply adding the regularization term defined in (25) to the proposed method of [14] which is far better than the L_2 approach, yet exhibiting some artifacts. Fig. 9(h) shows the implementation of the proposed method described in Section II-D. The selected parameters for this method were as follows: $\lambda = 0.005$, $P = 2$, $\beta = 110$, and $\alpha = 0.6$. Fig. 9(i) shows the implementation of the fast method described in Section II-E. The selected parameters for this method were as follows: $\lambda' = 0.08$, $P = 2$, $\beta = 1$, and $\alpha = 0.6$. Comparing Fig. 9(h) and (i) to other methods, we notice not only our method has removed the outliers more efficiently, but also it has resulted in sharper edges without any ringing effects.

Our second example is a real infrared camera image sequences with no known outliers, courtesy of B. Yasuda and



(g)

(h)

(i)

Fig. 9 (*Continued*). (g) Zomet [14] with regularization. (h) $L_1 +$ bilateral TV. (i) Median shift and add + bilateral TV.

the FLIR research group in the Sensors Technology Branch, Wright Laboratory, WPAFB, OH. We used eight LR frames in our reconstruction to get resolution enhancement factor of four [Fig. 10(a) shows one of the input LR images].[8] Fig. 10(b) shows the cubic spline interpolation of Fig. 10(a) by factor of four. The (unknown) camera PSF was assumed to be a 4×4 Gaussian kernel with standard deviation equal to one. We used the method described in [27] to computed the motion vectors. L_2 norm reconstruction with Tikhonov regularization (25) result is shown in Fig. 10(c) where Γ is defined in (19) and regularization factor λ was chosen to be 0.1. Fig. 10(d) shows the implementation of (22) with the following parameters $\lambda = 0.006$, $P = 2$, $\beta = 81$, and $\alpha = 0.5$. Although modeling noise in these frames as additive Gaussian is a reasonable assumption, our method achieved a better result than the best L_2 norm minimization.

Our third experiment is a real compressed sequence of 20 images (containing translational motion) from a commercial video camera; courtesy of Adyoron Intelligent Systems, Ltd., Tel Aviv, Israel. Fig. 11(a) is one of these LR images and Fig. 11(b) is the cubic spline interpolation of this image by factor of three. We intentionally rotated five frames of this sequence (rotation from 20° to 60°) out of position, creating a sequence of images with relative affine motion. The (unknown) camera PSF was assumed to be a 5×5 Gaussian kernel with standard deviation equal to two. We used the method described in [27] to computed the motion vectors with translational motion assumption. The error in motion modeling results in apparent shadows in L_2 norm reconstruction with Tikhonov regularization [Fig. 11(c)] where Γ is defined in (19) and regularization factor λ was chosen to be 0.5. These shadows are removed in Fig. 11(d), where the method described in Section II-D (22) was used for reconstruction with the following parameters $\lambda = 0.003$, $P = 2$, $\beta = 50$, and $\alpha = 0.7$.

Our final experiment is a factor of three resolution enhancement of a real compressed image sequence captured with a commercial webcam (3Com, Model no. 3718). The (unknown) camera PSF was assumed to be a 3×3 Gaussian kernel with standard deviation equal to 1. In this sequence, two separate sources of motion were present. First, by shaking the camera a global motion was created for each individual frame. Second, an Alpaca statue was independently moved in ten frames out of total 55 input frames. One of the LR input images is shown in Fig. 12(a). Cubic spline interpolation of Fig. 12(a) by factor of three is shown in Fig. 12(b). Fig. 12(c) and (d) are the shift and add results using mean and median operators [minimizing $\underline{\widehat{Z}}$ in (10) with $p = 2$ and $p = 1$, respectively]. Note that the median

[8] Note that this is an under-determined scenario.

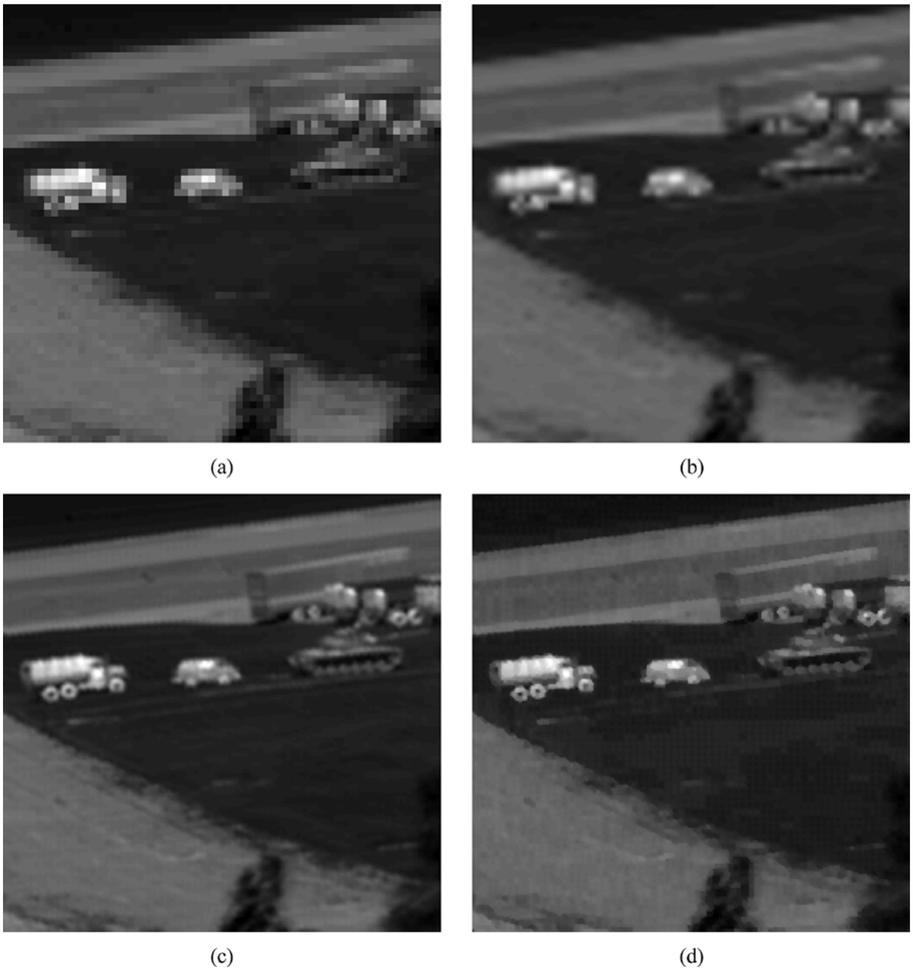


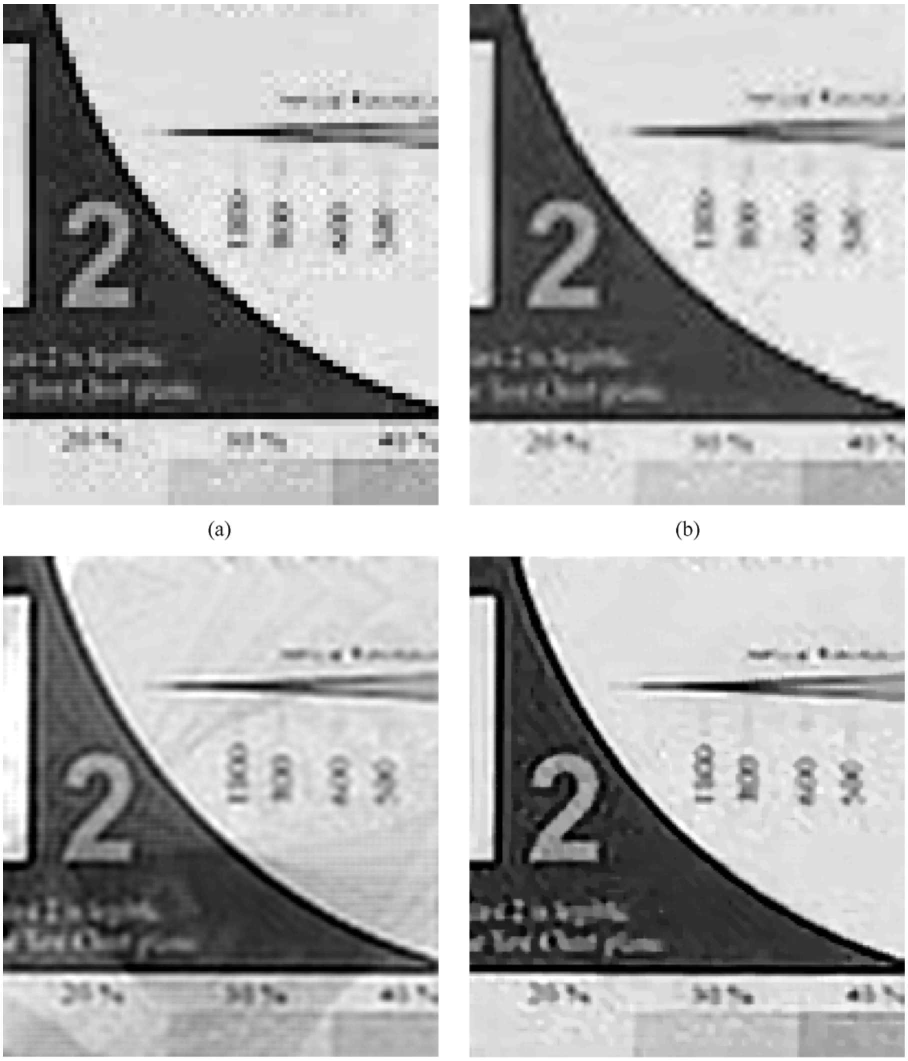
Fig. 10. Results of different resolution enhancement methods applied to Tank sequence. (a) One of eight LR frames. (b) Cubic spline interpolation. (c) $L_2 +$ Tikhonov. (d) $L_1 +$ bilateral TV.

operator has lessened the (shadow) artifacts resulting from the Alpaca motion. L_2 norm reconstruction with Tikhonov regularization (25) result is shown in Fig. 12(e), where Γ is defined in (19) and regularization factor λ was chosen to be one. Fig. 12(f) is the result of minimizing the cost function (as shown at the bottom of the page), where L_2 is the norm minimization of data error term is combined with bilateral TV regularization with the following parameters $\lambda = 0.1$, $P = 2$, $\alpha = 0.7$, and $\beta = 70$ (steepest descent step size). Note that the artifacts resulting from the motion of Alpaca statue is visible in Fig. 12(d)–(g). Robust super-resolution method proposed in [14] is shown in Fig. 12(h). Implementation of the method described in Section II-D (22) with the following parameters $\lambda = 0.003$, $P = 2$, $\beta = 30$, and $\alpha = 0.7$ resulted in Fig. 12(i), with the least outlier effect. And, finally, implementation of the fast method described in Section II-E (24) with the following parameters $\lambda' = 0.04$, $P = 2$, $\beta = 1$, and $\alpha = 0.7$ resulted in Fig. 12(j), which is very similar to the result in Fig. 12(i).

IV. Conclusion

In this paper, we presented an algorithm to enhance the quality of a set of noisy blurred images and produce a HR image with less noise and blur effects. We presented a robust super-resolution method based on the use of L_1 norm both in the regularization and the measurement terms of our penalty function. We showed that our method removes outliers efficiently, resulting in images with sharp edges. Even for images in which the noise followed the Gaussian model, L_1 norm minimization results were as good as L_2 norm minimization

$$\widehat{\underline{X}} = \underset{\underline{X}}{\operatorname{ArgMin}} \left[\sum_{k=1}^{N} \|D_k H_k F_k \underline{X} - \underline{Y}_k\|_2^2 + \lambda \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m \geq 0} \alpha^{|m|+|l|} \|\underline{X} - S_x^l S_y^m \underline{X}\|_1 \right]$$



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

Fig. 11. Results of different resolution enhancement methods applied to ADYORON test sequence. (a) One of 20 LR frames. (b) Cubic spline interpolation. (c) $L_2 +$ Tikhonov. (d) $L_1 +$ bilateral TV.

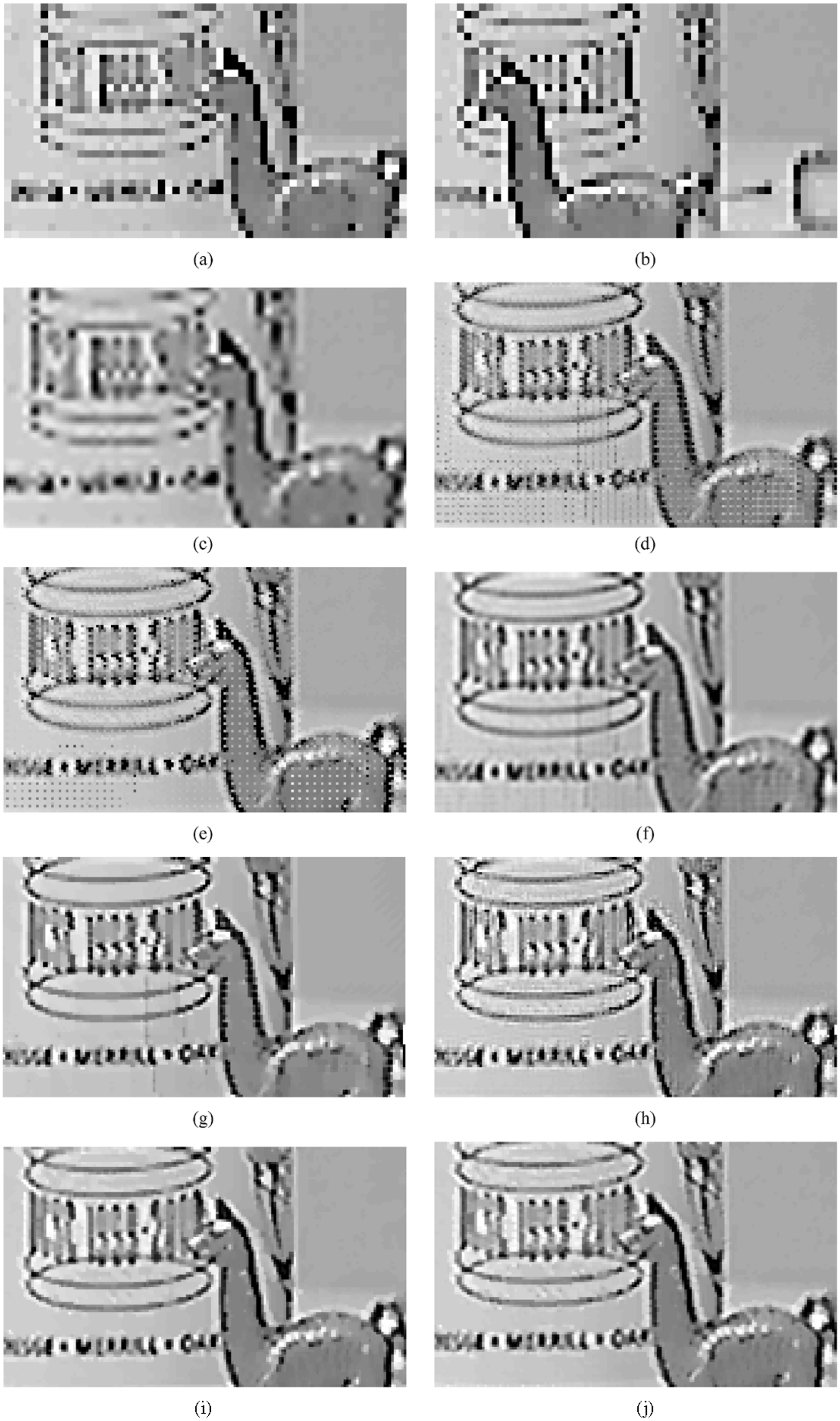
results, which encourages using L_1 norm minimization for any data set. The proposed method was fast and easy to implement.

We also proposed and mathematically justified a very fast method based on pixelwise "shift and add" and related it to L_1 norm minimization when relative motion is pure translational, and PSF and decimation factor is common and space invariant in all LR images. Note that the mathematical derivation of the proposed shift and add method was independent of the constraint over decimation factor, but we included it as this constraint distinguishes super-resolution problem from the more general problem of multiscale image fusion. In this method, we rounded the displacements in the HR grid so that F_k applies only integer translations. This will not pose a problem as the rounding is done only on the HR grid [5]. Besides, any alternative method will introduce time consuming smoothing interpolation effects which can be harder to overcome.

Analysis of the convergence properties of the steepest descent method is only possible for simplistic cases such as minimizing a quadratic function. Considering quantized images, L_1 norm minimization, and regularization terms make such analysis much harder. We have observed that only five to twenty iterations are required for convergence to the desired solution, where the initialization and the type of involved images play a vital role in determining the required iterations. The outcome of the speed-up method of Section II-E is a very good initialization guess for the more general case of Section II-D.

Although the "cross validation" method can be used to determine the parameter values [12], implementing such method for the L_1 norm is rather more difficult and computationally expensive. Parameters like P can also be learned using a learning algorithm, however such an approach is outside the scope of this paper. We have found that setting P to 2 or 3 works well; using higher values for P will be time consuming while not very useful.

One important extension for our algorithm include incorporation of blur identification algorithms in the super-resolution method. Although many single-frame blind deconvolution algorithms have been suggested in the last 30 years [28] and recently [12] incorporated a single-parameter blur identification algorithm in their super-resolution method, still there is need



(a) (b)

(c) (d)

(e) (f)

(g) (h)

(i) (j)

Fig. 12. Results of different resolution enhancement methods applied to the Alpaca sequence. Outlier effects are apparent in the nonrobust reconstruction methods. (a) Frame 1 of 55 LR frames. (b) Frame 50 of 55 LR frames. (c) Cubic spline interpolation of frame 1. (d) Mean shift and add. (e) Median shift and add. (f) $L_2 +$ Tikhonov. (g) $L_2 +$ bilateral TV. (h) Zomet method [14]. (i) $L_1 +$ bilateral TV. (j) Median shift and add + bilateral.

for more research to provide a super-resolution method along with a more general blur estimation algorithm.

Few papers have addressed resolution enhancement of compressed video sequences [15] and [16]. Compression artifacts resulting from quantization of DCT coefficients can dramatically decrease the performance of super-resolution system. The results of Section II-E may be used to design a very fast none iterative method for reducing the compression artifacts in the super-resolved images.

One of the most apparent effects of DCT-based compression methods, such as *MPEG* for video and *JPEG* for still images, is the blocking artifact. The quantization noise variance of each pixel in a block is space variant. For a block located in a low-frequency content area, pixels near boundaries contain more quantization noise than the interior pixels. On the other hand, for the blocks located in the high-frequency area, pixels near boundaries contain less quantization noise than the interior pixels [29]. This space-variant noise property of the blocks may be exploited to reduce the quantization noise. Because of the presence of motion in video sequences, pixel locations in the blocks change from one frame to the other. So two corresponding pixels from two different frames may be located on and off the boundaries of the blocks in which they are located. Based on the discussion that was presented in the previous paragraph, it is easy to determine which pixel has less quantization noise. It is reasonable to assign a higher weight to those pixels which suffer less from quantization noise in the data fusion step which was explained in Section II-E. The relative magnitude of the weight assigned because of quantization and the weight that was explained in Section II-E will depend on the compression ratio.

Appendix A
Bilateral Filter

The idea of the bilateral filter was first proposed in [30] as a very effective one-pass filter for denoising purposes while keeping sharp edges. Unlike conventional filters such as Gaussian low-pass filter, the bilateral filter defines the closeness of two pixels not only based on geometric distance but also based on photometric distance. Considering one-dimensional (1-D) case (for simplifying the notations), the result of applying bilateral filter for the kth sample in the estimated 1-D signal $\widehat{X}$ is

$$\widehat{X}[k] = \frac{\sum_{m=-M}^{M} W[k,m]Y[k-m]}{\sum_{m=-M}^{M} W[k,m]} \qquad (26)$$

where $\underline{Y} = \underline{X} + \underline{V}$ is the noisy image (vector), and $2 \times M + 1$ is the size of 1-D bilateral kernel. The weight $W[k,m] = W_S[k,m]W_P[k,m]$ considers both photometric and spatial difference of sample k in noisy vector $\underline{Y}$ from its neighbors to define the value of sample k in the estimated vector $\widehat{\underline{X}}$. The spatial and photometric difference weights were arbitrarily defined in [30] as

$$\begin{aligned} W_S[k,m] &= \exp\left\{-\frac{m^2}{2\sigma_S^2}\right\} \\ W_P[k,m] &= \exp\left\{-\frac{[Y[k]-Y[k-m]]^2}{2\sigma_R^2}\right\} \end{aligned} \qquad (27)$$

where parameters σ_S^2 and σ_R^2 control the strength of spatial and photometric property of the filter, respectively.

In [31] it was proved that such filter is a single iteration of the weighted least-squares minimization [shown in (28), at the bottom of the page], with Jacobi method, where S^m implies a shift right of m samples. [31] also showed that using more iterations will enhance the performance of this filter.

Note that if we define the (i,i)th element of the diagonal weight matrix $\mathbf{W_m}$ as

$$\mathbf{W_m}(i,i) = \frac{\alpha^m}{|\underline{X}(i) - S^m\underline{X}(i)|} \qquad 0 < \alpha < 1$$

that is, weighting the estimate with respect to both photometric distance $|\underline{X}(i) - S^m\underline{X}(i)|$ and geometric distance α^m, then (28) will become

$$\widehat{\underline{X}} = \underset{\underline{X}}{\operatorname{ArgMin}} \left[\|\underline{X} - \underline{Y}\|_2^2 + \lambda \sum_{m=1}^{M} \alpha^m \|\underline{X} - S^m\underline{X}\|_1 \right] \qquad (29)$$

which is the 1-D version of the bilateral TV criterion in (16).

Appendix B
Limitations of Zomet Method

A robust super-resolution method was recently proposed by Zomet *et al.* [14], where robustness is achieved by modifying the gradient of the L_2 norm cost function (7)

$$\begin{aligned} \underline{G}_2 &= \sum_{k=1}^{N} \underline{B}_k = \sum_{k=1}^{N} F_k^T H_k^T D_k^T (D_k H_k F_k \underline{X} - \underline{Y}_k) \\ &= \sum_{k=1}^{N} F_k^T H_k^T D_k^T \underline{U}_k \end{aligned} \qquad (30)$$

in which $\underline{B}_k$ is the gradient resulted from frame k and $\underline{U}_k$ represents the residual vector. They substituted (30) with the following approximation:

$$\widehat{\underline{G}}_2 = \text{MED}\{\underline{B}_k\}_{k=1}^N = \text{MED}\{F_k^T H_k^T D_k^T \underline{U}_k\}_{k=1}^N \qquad (31)$$

where MED is a pixelwise median operator. Then, steepest descent minimization was used to calculate $\widehat{\underline{X}}$

$$\widehat{\underline{X}}_{n+1} = \widehat{\underline{X}}_n + \lambda'' \widehat{\underline{G}}_2. \qquad (32)$$

where λ'' is the step size in the direction of gradient.

We show that for certain imaging scenarios, the approximated gradient (31) is zero in all iterations, which means estimated HR

$$\begin{aligned} \widehat{\underline{X}} &= \underset{\underline{X}}{\operatorname{ArgMin}} \left[\|\underline{X} - \underline{Y}\|_2^2 + \lambda \sum_{m=1}^{M} \|\underline{X} - S^m\underline{X}\|_{W_m}^2 \right] \\ &= \underset{\underline{X}}{\operatorname{ArgMin}} \left[[\underline{X} - \underline{Y}]^T [\underline{X} - \underline{Y}] + \lambda \sum_{m=1}^{M} [\underline{X} - S^m\underline{X}]^T \mathbf{W_m} [\underline{X} - S^m\underline{X}] \right] \end{aligned} \qquad (28)$$

frame of the nth iteration ($\widehat{\underline{X}}_n$) is the same as the initial guess ($\widehat{\underline{X}}_n$) and the method fails. To appreciate this fact, lets start with a square case in which blurring effect is negligible (i.e., H_k is an identity matrix resulting in $\underline{B}_k = F_k^T D_k^T \underline{U}_k$). A quick consultation with Fig. 3 suggests that only one of every r^2 elements in $D_k^T \underline{U}_k$ has a nonzero value. Moreover, recall that F_k^T just registers vector $D_k^T \underline{U}_k$ with respect to the estimated relative motion without changing its value. According to (31), $\widehat{\underline{G}}(i)$ (the ith element of the gradient vector) is equal to $\mathrm{MED}\{\underline{B}_k(i)\}_{k=1}^N$. As $N-1$ elements in $\{\underline{B}_k(i)\}_{k=1}^N$ have zero value, their median will also be zero. Therefore, every element of the approximated gradient vector will be zero. Even for a more general case in which the effect of blur matrix is not negligible (H_k is a matrix form of a $m \times n$ blur kernel), the same approach may be employed to show that unless ($m \times n > r^2/2$), the gradient will remain zero for all iterations.

The ($m \times n > r^2/2$) condition is also valid for the over-determined cases where the distribution of motion vectors is uniform (that is the number of available LR measurements for each pixel in the HR grid is equal). Therefore, this condition does not depend on the number of available LR frames. In particular, consider the identity blur matrix case, where the addition of any new frame Y_ϑ is equivalent to the addition of a new gradient vector $\underline{B}_\vartheta$ with r^2-1 times more zero elements (resulting from upsampling) than nonzero elements to the stack of gradient vectors. Therefore, if

$$\widehat{\underline{G}}(i) = \mathrm{MED}\{\underline{B}_k(i)\}_{k=1}^N = \underline{0}$$

even after addition of r^2 uniformly spread LR frames $\widehat{\underline{G}}'(i) = \mathrm{MED}\{\underline{B}_k(i)\}_{k=1}^{N+r^2}$ will still be zero (as r^2-1 value of r^2 newly added elements are zeros). Generalization of this property to the case of arbitrary number of LR frames with uniform motion distribution is straightforward.

This limitation can be overcome by modifying the MED operator in (31). This modified median operator would not consider those elements of $\underline{B}_k(i)$ which are the result of zero filling. It is interesting to note that such assumption will result in estimating the HR frame as the median of registered LR frames after zero filling, which is the exact interpretation of using L_1 norm minimization discussed in Section II-B.

Acknowledgment

The authors would like to thank the Associate Editor Prof. R. D. Nowak and the three reviewers for valuable comments that helped improve the clarity of presentation of this paper.

References

[1] T. S. Huang and R. Y. Tsai, "Multi-frame image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, pp. 317–339, 1984.

[2] N. K. Bose, H. C. Kim, and H. M. Valenzuela, "Recurcive implementation of total least squares algorithm for image reconstruction from, noisy, undersampled multiframes," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, Minneapolis, MN, Apr. 1993, pp. 269–272.

[3] S. Borman and R. L. Stevenson, "Super-resolution from image sequences—A review," in *Proc. Midwest Symp. Circuits and Systems*, vol. 5, Notre Dame, IN, Apr. 1998.

[4] L. Teodosio and W. Bender, "Salient video stills: Content and context preserved," in *Proc. 1st ACM Int. Conf. Multimedia*, vol. 10, Anaheim, CA, Aug. 1993, pp. 39–46.

[5] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space invariant blur," *IEEE Trans. Image Processing*, vol. 10, pp. 1187–1193, Aug. 2001.

[6] M. C. Chiang and T. E. Boulte, "Efficient super-resolution via image warping," *Image Vis. Comput.*, vol. 18, no. 10, pp. 761–771, July 2000.

[7] S. Peleg, D. Keren, and L. Schweitzer, "Improving image resolution using subpixel motion," *CVGIP: Graph. Models Image Process.*, vol. 54, pp. 181–186, Mar. 1992.

[8] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graph. Models Image Process.*, vol. 53, pp. 231–239, 1991.

[9] H. Ur and D. Gross, "Improved resolution from sub-pixel shifted pictures," *CVGIP: Graph. Models Image Process.*, vol. 54, no. 181–186, Mar. 1992.

[10] M. Elad and A. Feuer, "Restoration of single super-resolution image from several blurred, noisy and down-sampled measured images," *IEEE Trans. Image Processing*, vol. 6, pp. 1646–1658, Dec. 1997.

[11] N. Nguyen, P. Milanfar, and G. H. Golub, "A computationally efficient image superresolution algorithm," *IEEE Trans. Image Processing*, vol. 10, pp. 573–583, Apr. 2001.

[12] ——, "Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement," *IEEE Trans. Image Processing*, vol. 10, pp. 1299–1308, Sept. 2001.

[13] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using delaunay triangulation," *IEEE Trans. Image Processing*, vol. 11, pp. 1427–1441, Dec. 2002.

[14] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super resolution," in *Proc. Int. Conf. Computer Vision and Patern Recognition*, vol. 1, Dec. 2001, pp. 645–650.

[15] Y. Altunbasak, A. Patti, and R. Mersereau, "Super-resolution still and video reconstruction from mpeg-coded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 217–226, Apr. 2002.

[16] C. A. Segall, R. Molina, A. Katsaggelos, and J. Mateos, "Bayesian high-resolution reconstruction of low-resolution compressed video," in *IEEE Int. Conf. Image Processing*, vol. 2, Thessaloniki, Greece, Oct. 2001, pp. 25–28.

[17] G. C. Calafiore, "Outliers robustness in multivariate orthogonal regression," *IEEE Trans. Syst., Man. Cybern.*, vol. 30, no. 6, pp. 674–679, Nov. 2000.

[18] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing:Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993, vol. I.

[20] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to super-resolution," in *Proc. SPIE Conf. Applications of Digital Signal and Image Processing*, San Diego, CA, Aug. 2003, pp. 121–130.

[21] A. M. Tekalp, *Digital Video Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[22] A. Bovik, *Handbook of Image and Video Processing*. New York: Academic, 2000.

[23] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992.

[24] T. F. Chan, S. Osher, and J. Shen, "The digital TV filter and nonlinear denoising," *IEEE Trans. Image Processing*, vol. 10, pp. 231–241, Feb. 2001.

[25] Y. Li and F. Santosa, "A computational algorithm for minimizing total variation in image restoration," *IEEE Trans. Image Processing*, vol. 5, pp. 987–995, June 1996.

[26] A. Zomet and S. Peleg, "Efficient super-resolution and applications to mosaics," in *Proc. Int. Conf. Pattern Recognition*, Sept. 2000.

[27] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierachical model-based motion estimation," in *Proc. Eur. Conf. Computer Vision*, 1992, pp. 237–252.

[28] D. Kondur and D. Hatzinakos, "Blind image deconvolution," *IEEE Signal Processing Mag.*, vol. 13, pp. 43–64, May 1996.

[29] M. Robertson and R. Stevenson, "DCT quantization noise in compressed images," in *IEEE Int. Conf. Image Processing*, vol. 1, Thessaloniki, Greece, Oct. 2001, pp. 185–1888.

[30] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Computer Vision*, New Delhi, India, Jan. 1998, pp. 836–846.

[31] M. Elad, "On the bilateral filter and ways to improve it," *IEEE Trans. Image Processing*, vol. 11, pp. 1141–1151, Oct. 2002.

Sina Farsiu received the B.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1999 and the M.Sc.(Hons) degree in biomedical engineering from the University of Tehran, Tehran, in 2001. He is currently pursuing the Ph.D. degree in electrical engineering at the University of California, Santa Cruz.

His technical interests include signal and image processing, adaptive optics, and artificial intelligence.

Dirk Robinson (S'01) received the B.S. degree in electrical engineering from Calvin College, Grand Rapids, MI, and the M.S. degree in computer engineering from the University of California, Santa Cruz (UCSC), in 1999 and 2001, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at UCSC.

His technical interests include signal and image processing and machine learning.

Michael Elad received the B.Sc, M.Sc., and D.Sc. degrees from the Department of Electrical Engineering at the Technion–Israel Institute of Technology (IIT), Haifa, Israel, in 1986, 1988, and 1997, respectively.

From 1988 to 1993, he served in the Israeli Air Force. From 1997 to 2000, he worked at Hewlett–Packard Laboratories as an R&D Engineer. From 2000 to 2001, he headed the research division at Jigami Corporation, Israel. From 2001 to 2003, he was a Research Associate with the Computer Science Department, Stanford University (SCCM program), Stanford, CA. In September 2003, he joined the Department of Computer Science, IIT, as an Assistant Professor. He was also a Research Associate at IIT from 1998 to 2000, teaching courses in the Electrical Engineering Department. He works in the field of signal and image processing, specializing, in particular, on inverse problems, sparse representations, and over-complete transforms.

Dr. Elad received the Best Lecturer Award twice (in 1999 and 2000). He is also the recipient of the Guttwirth and the Wolf fellowships.

Peyman Milanfar (SM'98) received the B.S. degree in electrical engineering and mathematics from the University of California, Berkeley, and the S.M., E.E., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1988, 1990, 1992, and 1993, respectively.

Until 1999, he was a Senior Research Engineer at SRI International, Menlo Park, CA. He is currently Associate Professor of Electrical Engineering, University of California, Santa Cruz. He was a Consulting Assistant Professor of computer science at Stanford University, Stanford, CA, from 1998 to 2000, where he was also a Visiting Associate Professor from June to December 2002. His technical interests are in statistical signal and image processing and inverse problems.

Dr. Milanfar won a National Science Foundation CAREER award in 2000 and he was Associate Editor for the IEEE SIGNAL PROCESSING LETTERS from 1998 to 2001.