

On Bayesian Adaptive Video Super Resolution

Ce Liu *Member, IEEE*, Deqing Sun *Member, IEEE*

Abstract—Although multi-frame super resolution has been extensively studied in past decades, super resolving real-world video sequences still remains challenging. In existing systems, either the motion models are oversimplified, or important factors such as blur kernel and noise level are assumed to be known. Such models cannot capture the intrinsic characteristics that may differ from one sequence to another. In this paper, we propose a Bayesian approach to adaptive video super resolution via simultaneously estimating underlying motion, blur kernel and noise level while reconstructing the original high-res frames. As a result, our system not only produces very promising super resolution results outperforming the state of the art, but also adapts to a variety of noise levels and blur kernels. To further analyze the effect of noise and blur kernel, we perform a two-step analysis using the Cramer-Rao bounds. We study how blur kernel and noise influence motion estimation with aliasing signals, how noise affects super resolution with perfect motion, and finally how blur kernel and noise influence super resolution with unknown motion. Our analysis results confirm empirical observations, in particular that an intermediate size blur kernel achieves the optimal image reconstruction results.

Index Terms—Super resolution, optical flow, blur kernel, noise level, aliasing

1 INTRODUCTION

Multi-frame super resolution, namely estimating the high-res frames from a low-res sequence, is one of the fundamental problems in computer vision and has been extensively studied for decades. The problem becomes particularly interesting as high-definition devices such as high definition television HDTV (1920×1080) dominate the market. The resolution of various display has increased dramatically recently, including the New iPad (2048×1536), 2012 Macbook Pro (2880×1800), and ultra high definition television UHD TV (3840×2048 or 4K, 7680×4320 or 8k). As a result, there is a great need for converting low-resolution, low-quality videos into high-resolution, noise-free videos that can be pleasantly viewed on these high-resolution devices.

Although a lot of progress has been made in the past 30 years, super resolving real-world video sequences still remains an open problem. Most of the previous work assumes that the underlying motion has a simple parametric form, and/or that the blur kernel and noise levels are known. But in reality, the motion of objects and cameras can be arbitrary, the video may be contaminated with noise of unknown level, and motion blur and point spread functions can lead to an unknown blur kernel.

Therefore, a practical super resolution system should simultaneously estimate optical flow [12], noise level [23] and blur kernel [16] in addition to reconstructing the high-res image. As each of these problems has been well studied in computer vision, it is natural to combine all these components in a single framework without making simplified assumptions.

In this paper, we propose a Bayesian framework for adaptive video super resolution that incorporates high-res

image reconstruction, optical flow, noise level and blur kernel estimation. Using a sparsity prior for the high-res image, flow fields and blur kernel, we show that super resolution computation is reduced to each component problem when other factors are known, and the MAP inference iterates between optical flow, noise estimation, blur estimation and image reconstruction. As shown in Figure 1 and later examples, our system produces promising results on challenging real-world sequences despite various noise levels and blur kernels, accurately reconstructing both major structures and fine texture details. In-depth experiments demonstrate that our system outperforms the state-of-the-art super resolution systems [1], [31], [36] on challenging real-world sequences.

We are also interested in theoretical aspects of super resolution, namely to what extent the original high-res information can be recovered under a given condition. Although previous work [3], [19] on the limits of super resolution provides important insights into the increasing difficulty of recovering the signal as a function of the up-sampling factor, most of the bounds are obtained for the entire signal with frequency perspective ignored. Intuitively, high frequency components of the original image are much harder to recover as the blur kernel, noise level and/or up-sampling factor increases.

In a preliminary conference version of the paper [22], we theoretically analyzed the performance using Wiener filter theory. With known ground truth motion, Our analysis predicts that a small blur kernel always produces better image reconstruction results. However we empirically observed that a medium-sized blur kernel achieves the best super resolution results.

When the motion is unknown, our system estimates the motion from low-res, aliased images. Aliasing causes problem to motion estimation and is better suppressed by a large blur kernel. A large blur kernel however boosts the noise more in the image reconstruction process. In this paper, we perform a two-step analysis to consider motion estimation. Our theoretical results confirm our empirical observations that the blur kernel has a two-fold effect on

- Ce Liu is with Microsoft Research New England.
E-mail: celiu@microsoft.com
- Deqing Sun is with Harvard University.
E-mail: dqsun@seas.harvard.edu

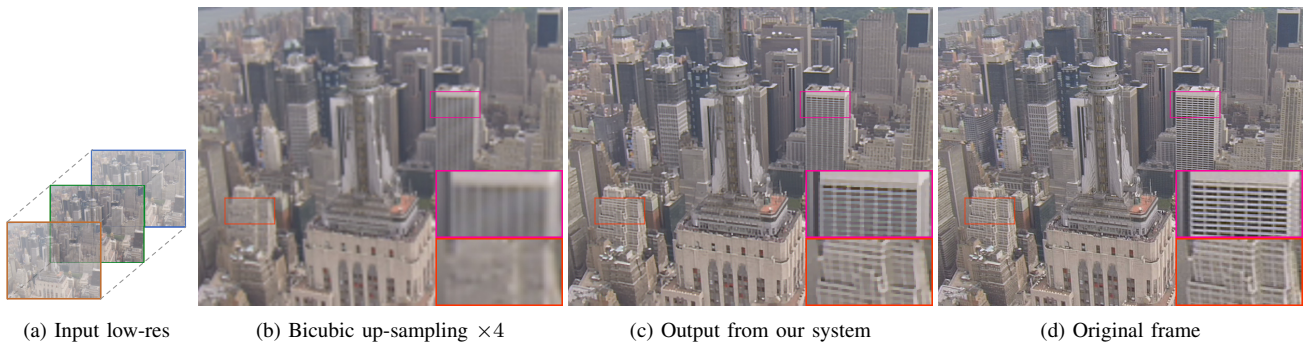


Fig. 1. Our video super resolution system is able to recover image details after $\times 4$ up-sampling.

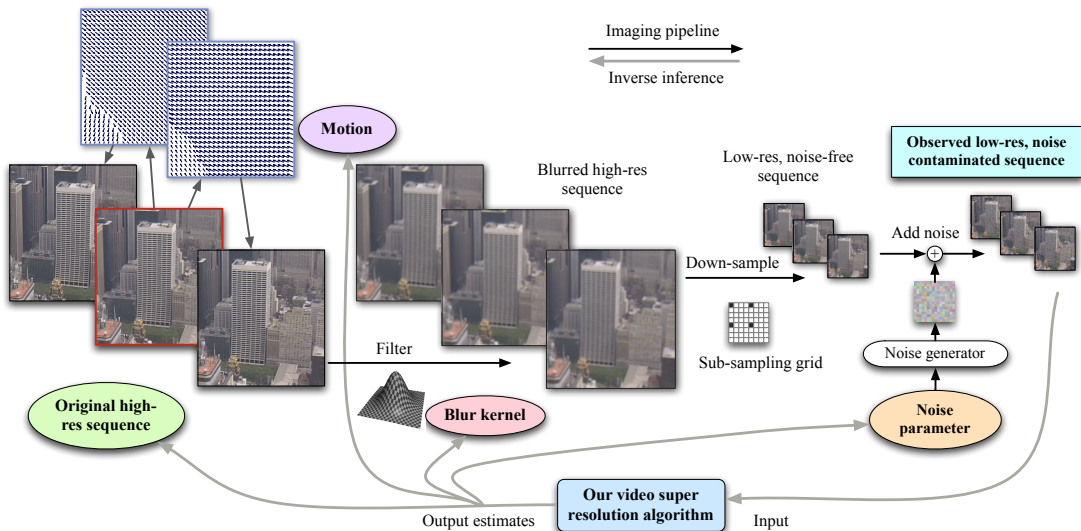


Fig. 2. **Video super resolution diagram.** The original high-res video sequence is generated by warping the source frame (enclosed by a red rectangle) both forward and backward with some motion fields. The high-res sequence is then smoothed with a blur kernel, down-sampled and contaminated with noise to generate the observed sequence. Our adaptive video super resolution system not only estimates the high-res sequence, but also the underlying motion (on the lattice of original sequence), blur kernel and noise level.

the image reconstruction and a medium-size blur kernel can reach a tradeoff between aliasing suppression and noise reduction.

The paper is organized as follows. After reviewing related work in Sect 2, we introduce our Bayesian super resolution framework in Sect 3. We prove the performance bounds in Sect 4, and show experimental results in Sect 5. After in-depth discussion in Sect 6, we conclude our paper in Sect 7.

2 RELATED WORK

Since the seminal work by Tsai and Huang [37], significant progress has been made in super resolution. We refer readers to [26] for a comprehensive literature review.

Early super resolution work focused on dealing with the ill-posed nature of reconstructing a high-res image from a sequence of low-res frames [13]. The lack of constraints is often addressed by spatial priors on the high-res image [30]. Hardie *et al.* [11] jointly estimated the translational motion and the high-res image, while Bascole *et al.* [4] also considered the motion blur using an affine

motion model. But these motion models are too simple to reflect the nature of real-world sequences.

To deal with the complex motion of faces, Baker and Kanade [2] proposed to use optical flow for super resolution, although in fact a parametric motion model was adopted. Fransens *et al.* [10] proposed a probabilistic formulation and jointly estimated the image, flow field and Gaussian noise statistics within an EM framework. They assumed that the blur kernel was known, and used Gaussian priors for both images and flow fields. However, Gaussian priors tend to over-smooth sharp boundaries in images and flows.

While most of these motion-based super resolution models use somewhat standard flow estimation techniques, recent advances in optical flow have resulted in much more reliable methods based on sparsity priors *e.g.* [6]. Accurate motion estimation despite strong noise has inspired Liu and Freeman [21] to develop a high quality video denoising system that removes structural noise in real video sequences. In this paper, we also want to incorporate recent advances in optical flow for more accurate super resolution.

Inspired by the successful non-local means method for video denoising, Takeda *et al.* [36] avoided explicit sub-

pixel motion estimation and used 3D kernel regression to exploit the spatiotemporal neighboring relationship for video up-sampling. However, their method still needs to estimate a pixel-wise motion at regions with large motion. In addition, its data model does not include blur and so its output needs to be postprocessed by a deblurring method.

While most methods assume the blur kernel is known, some work considers estimating the blur kernel under simple settings. Nguyen *et al.*[24] used the generalized cross-correlation method to identify the blur kernel using quadratic formulations. Sroubek *et al.*[32] estimated the image and the blur kernel under translational motion models by joint MAP estimation. However, their models can barely generalize to real videos due to the oversimplified motion models.

Significant improvements on blur estimation from real images have been made in the blind deconvolution community. Levin *et al.*[18] showed that joint MAP estimation of the blur kernel and the original image favors a non-blur explanation, *i.e.*, a delta blur function and the blurred image. Their analysis assumes no spatial prior on the blur kernel, while Joshi *et al.*[14] used a smoothness prior for the blur kernel and obtained reliable estimates. Moreover, Shan *et al.*[31] applied the recent development in image deconvolution to super resolution and obtained promising improvement, but their method only works on a single frame and does not estimate the noise statistics.

On the theory side, there has been important work on the limit of super resolution as the up-sampling factor increases [3], [19]. Their analysis focused on the stability of linear systems while ignoring the frequency aspects of the limit. In fact, many useful tools have been developed in the signal processing community to analyze the performance of linear systems w.r.t. a particular frequency component. Robinson and Milanfar [27] derived the Cramer-Rao bounds (CRB) [15] for each frequency bands using translational motion model. Their analysis does not consider the aliasing effect and their results suggest that a small blur kernel always produces the best performance. Empirically we find that a medium-sized blur kernel can achieve the optimal performance.

Similar to our iterative system, we perform a two-step analysis of the CRB for motion estimation and image reconstruction. First, we analyze the estimation of motion on the low-res input images with high frequency aliasing. Second, we analyze the performance of image reconstruction with errors in the estimated motion. Our analysis is closer to the estimation procedure and consistent with the empirical observations.

3 A BAYESIAN MODEL FOR SUPER RESOLUTION

Given the low-res sequence $\{J_t\}$, our goal is to recover the high-res sequence $\{I_t\}$. Due to computational issues, we aim at estimating I_t using adjacent frames $J_{t-N}, \dots, J_{t-1}, J_t, J_{t+1}, \dots, J_{t+N}$. To make the notations succinct, we will omit t from now on. Our problem becomes to estimate I given a series of images $\{J_{-N}, \dots, J_N\}$. In addition, we will derive the equations

using gray-scale images for simplicity although our implementation is able to handle color images.

The model of obtaining low-res sequence is illustrated in Figure 2. A full generative model that corresponds to Figure 2 is shown in Figure 3. At time $i = 0$, frame I is smoothed and down-sampled to generate J_0 with noise. At time $i = -N, \dots, N$, $i \neq 0$, frame I is first warped according to a flow field w_i , and then smoothed and down-sampled to generate J_i with noise and outlier R_i (we need to model outliers because optical flow cannot perfectly explain the correspondence between two frames). The unknown parameters in the generative models include the smoothing kernel K , which corresponds to point spread functions in the imaging process, or smoothing filter when video is down-sampled, and parameter θ_i that controls the noise and outlier when I is warped to generate adjacent frames.

We use Bayesian MAP to find the optimal solution

$$\{I^*, \{w_i\}^*, K^*, \{\theta_i\}^*\} = \underset{I, \{w_i\}, K, \{\theta_i\}}{\operatorname{argmax}} p(I, \{w_i\}, K, \{\theta_i\} | \{J_i\}), \quad (1)$$

where the posterior is the product of prior and likelihood:

$$p(I, \{w_i\}, K, \{\theta_i\} | \{J_i\}) \propto p(I)p(K) \prod_i p(w_i) \prod_i p(\theta_i) \cdot p(J_0 | I, K, \theta_0) \prod_{i \neq 0} p(J_i | I, K, w_i, \theta_i). \quad (2)$$

Sparsity on derivative filter responses is used to model the priors of image I , optical flow field w_i and blur kernel K

$$p(I) = \frac{1}{Z_I(\eta)} \exp \{-\eta \|\nabla I\|\}, \quad (3)$$

$$p(w_i) = \frac{1}{Z_w(\lambda)} \exp \left\{ -\lambda \left(\|\nabla u_i\| + \|\nabla v_i\| \right) \right\}, \quad (4)$$

$$p(K_x) = \frac{1}{Z_K(\xi)} \exp \{-\xi \|\nabla K_x\|\}, \quad (5)$$

where ∇ is the gradient operator, $\|\nabla I\| = \sum_q \|\nabla I(q)\| = \sum_q (|I_x(q)| + |I_y(q)|)$ ($I_x = \frac{\partial}{\partial x} I$, $I_y = \frac{\partial}{\partial y} I$) and q is the pixel index. The same notation holds for u_i and v_i , the horizontal and vertical components of the flow field w_i . For computational efficiency, we assume the kernel K is x- and y-separable: $K = K_x \otimes K_y$, where K_y has the same pdf as K_x . $Z_I(\eta)$, $Z_w(\lambda)$ and $Z_K(\xi)$ are normalization constants only dependant on η , λ and ξ , respectively.

To deal with outliers, we assume an exponential distribution for the likelihood

$$p(J_i | I, K, \theta_i) = \frac{1}{Z(\theta_i)} \exp \left\{ -\theta_i \left\| J_i - \mathbf{SKF}_{w_i} I \right\| \right\}, \quad (6)$$

where the parameter θ_i reflects the noise level of frame i and $Z(\theta_i) = (2\theta_i)^{-\dim(I)}$. Matrices \mathbf{S} and \mathbf{K} correspond to down-sampling and filtering with blur kernel K , respectively. \mathbf{F}_{w_i} is the warping matrix corresponding to flow w_i . Naturally, the conjugate prior for θ_i is a Gamma distribution

$$p(\theta_i; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} \exp\{-\theta_i \beta\}. \quad (7)$$

Now that we have the probability distributions for both prior and likelihood, and the Bayesian MAP inference

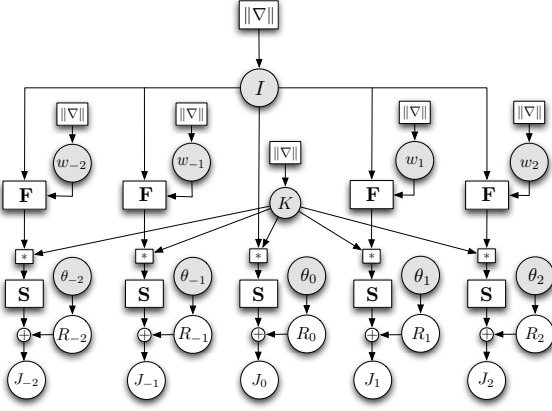


Fig. 3. The graphical model of video super resolution. The circular nodes are variables (vectors), whereas the rectangular nodes are matrices (matrix multiplication). We do not put priors η , λ , ξ , α and β on I , w_i , K , and θ_i for succinctness.

is performed using coordinate descend. Note that in this model there are only five free parameters: η , λ , ξ , α and β .

3.1 Image Reconstruction

Given the current estimates of the flow field w_i , the blur kernel K and the noise level θ_i , we estimate the high-res image by solving

$$I^* = \underset{I}{\operatorname{argmin}} \theta_0 \|\mathbf{SK}I - J_0\| + \eta \|\nabla I\| + \sum_{i=-N, i \neq 0}^N \theta_i \|\mathbf{SKF}_{w_i}I - J_i\|. \quad (8)$$

To use gradient-based methods, we replace the L1 norm with a differentiable approximation $\phi(x^2) = \sqrt{x^2 + \epsilon^2}$ ($\epsilon = 0.001$), and denote the vector $\Phi(|I|^2) = [\phi(I^2(q))]$.

This objective function can be solved by the iterated reweighted least squares (IRLS) method [20], which iteratively solves the following linear system:

$$\begin{aligned} & \left[\theta_0 \mathbf{K}^T \mathbf{S}^T \mathbf{W}_0 \mathbf{SK} + \eta \left(\mathbf{D}_x^T \mathbf{W}_s \mathbf{D}_x + \mathbf{D}_y^T \mathbf{W}_s \mathbf{D}_y \right) + \right. \\ & \quad \left. \sum_{i=-N, i \neq 0}^N \theta_i \mathbf{F}_{w_i}^T \mathbf{K}^T \mathbf{S}^T \mathbf{W}_i \mathbf{SKF}_{w_i} \right] I \\ & = \theta_0 \mathbf{K}^T \mathbf{S}^T \mathbf{W}_0 J_0 + \sum_{i=-N, i \neq 0}^N \theta_i \mathbf{F}_{w_i}^T \mathbf{K}^T \mathbf{S}^T \mathbf{W}_i J_i, \quad (9) \end{aligned}$$

where the matrices \mathbf{D}_x and \mathbf{D}_y correspond to the x- and y- derivative filters. IRLS iterates between solving the above least square problem (through conjugate gradient) and estimating the diagonal weight matrices

$$\begin{aligned} \mathbf{W}_0 &= \operatorname{diag}(\Phi'(|\mathbf{SK}I - J_0|^2)), \\ \mathbf{W}_s &= \operatorname{diag}(\Phi'(|\nabla I|^2)), \\ \mathbf{W}_i &= \operatorname{diag}(\Phi'(|\mathbf{SKF}_{w_i}I - J_i|^2)) \end{aligned} \quad (10)$$

based on the current estimate.

3.2 Motion and Noise Estimation

Given the high-res image and the blur kernel, we jointly estimate the flow field and the noise level in a coarse-to-fine fashion on a Gaussian image pyramid. At each pyramid level noise level and optical flow are estimated iteratively. The Bayesian MAP estimate for the noise parameter θ_i has the following closed-form solution

$$\theta_i^* = \frac{\alpha + N_q - 1}{\beta + N_q \bar{x}}, \quad \bar{x} = \frac{1}{N_q} \sum_{q=1}^{N_q} \left| (J_i - \mathbf{SKF}_{w_i}I)(q) \right|, \quad (11)$$

where \bar{x} is sufficient statistics. When noise is known, the flow field w_i is estimated as

$$w_i^* = \underset{w_i}{\operatorname{argmin}} \theta_i \|\mathbf{SKF}_{w_i}I - J_i\| + \lambda \|\nabla u_i\| + \lambda \|\nabla v_i\|, \quad (12)$$

where we again approximate $|x|$ by $\phi(x^2)$. Notice that this optical flow formulation is different from the standard ones: the flow is established from high-res I to low-res J_i .

By first-order Taylor expansion

$$\mathbf{F}_{w_i + dw_i}I \approx \mathbf{F}_{w_i}I + \mathbf{I}_x du_i + \mathbf{I}_y dv_i, \quad (13)$$

where $\mathbf{I}_x = \operatorname{diag}(\mathbf{F}_{w_i}I_x)$ and $\mathbf{I}_y = \operatorname{diag}(\mathbf{F}_{w_i}I_y)$, we can approximate the first (data) term in Eqn. 12 as

$$\begin{aligned} \|\mathbf{SKF}_{w_i}I - J_i\| &\approx (\mathbf{F}_{w_i}I + \mathbf{I}_x du_i + \mathbf{I}_y dv_i - J_i)^T \\ &\quad \tilde{\mathbf{W}}_i (\mathbf{F}_{w_i}I + \mathbf{I}_x du_i + \mathbf{I}_y dv_i - J_i), \quad (14) \end{aligned}$$

where $\tilde{\mathbf{W}}_i = \mathbf{K}^T \mathbf{S}^T \mathbf{W}_i \mathbf{SK}$, the second (spatial) term for the horizontal flow as

$$\|\nabla u_i\| \approx (u_i + du_i)^T \mathbf{L} (u_i + du_i)^T, \quad (15)$$

where

$$\mathbf{L} = \mathbf{D}_x^T \operatorname{diag}(\Phi'(|\nabla u_i|^2)) \mathbf{D}_x + \mathbf{D}_y^T \operatorname{diag}(\Phi'(|\nabla v_i|^2)) \mathbf{D}_y \quad (16)$$

is a weighted Laplacian matrix, and similarly for the third term. Taking derivative w.r.t. the unknown flow increment (du_i, dv_i) and setting it to be zero, we can derive

$$\begin{aligned} & \begin{bmatrix} \mathbf{I}_x^T \tilde{\mathbf{W}}_i \mathbf{I}_x + \zeta_i \mathbf{L} & \mathbf{I}_x^T \tilde{\mathbf{W}}_i \mathbf{I}_y \\ \mathbf{I}_y^T \tilde{\mathbf{W}}_i \mathbf{I}_x & \mathbf{I}_y^T \tilde{\mathbf{W}}_i \mathbf{I}_y + \zeta_i \mathbf{L} \end{bmatrix} \begin{bmatrix} du_i \\ dv_i \end{bmatrix} = \\ & - \begin{bmatrix} \zeta_i \mathbf{L} u_i \\ \zeta_i \mathbf{L} v_i \end{bmatrix} - \begin{bmatrix} \mathbf{I}_x^T \\ \mathbf{I}_y^T \end{bmatrix} (\tilde{\mathbf{W}}_i \mathbf{F}_{w_i}I - \mathbf{K}^T \mathbf{S}^T \mathbf{W}_i J_i), \quad (17) \end{aligned}$$

where $\zeta_i = \frac{\lambda}{\theta_i}$. Again, we use IRLS [20] to solve the above equation iteratively.

One may notice that it is more expensive to solve Eqn. 17 than ordinary optical flow because in each iteration smoothing and down-sampling as well as the transposes need to be computed. We estimate optical flow from J_i to J_0 on the low-res lattice, and up-sample the estimated flow field to the high-res lattice as initialization for solving Eqn. 17.

3.3 Kernel Estimation

Without loss of generality, we only show how to estimate the x-component kernel K_x given I and J_0 . Let each row of matrix \mathbf{A} be the concatenation of pixels corresponding to the filter K , and define $\mathbf{M}_y : \mathbf{M}_y K_x = K_y \otimes K_x = K$. Estimating K_x leads to

$$K_x^* = \underset{K_x}{\operatorname{argmin}} \theta_0 \|\mathbf{SAM}_y K_x - J_0\| + \xi \|\nabla K_x\|, \quad (18)$$

which is optimized by IRLS.

Although similar Bayesian MAP approach performed poorly for general deblurring problems [18], the spatial smoothness prior on the kernel prevents kernel estimation from converging to the delta function, as shown by [14]. Our experiments also show that our estimation is able to recover the underlying blur kernel.

TABLE 1

The coordinate descent algorithm for Bayesian inference on video super resolution

<p>Input: low-res frames $\{J_i\}_{i=-N}^N$ and upsampling factor s</p> <ul style="list-style-type: none"> • Initialize $k = 1$, $I^{(0)} = J_0 \uparrow s$ (bicubic upsampling) • Loop until $I^{(k-1)} - I^{(k)} < \epsilon$ (<i>outer iteration</i>) <ul style="list-style-type: none"> - Estimate motion $w_i^{(k)}$ by solving Eqn. 17 - Estimate noise $\theta_i^{(k)}$ by solving Eqn. 11 - Initialize $m = 1$, $I^{(k,0)} = I^{(k-1)}$ - Loop until $I^{(k,m)} - I^{(k,m-1)} < \epsilon$ (<i>inner iteration</i>) <ul style="list-style-type: none"> - Compute weight \mathbf{W}_0, \mathbf{W}_s, \mathbf{W}_i using Eqn. 10 - Estimate $I^{(k,m)}$ by solving Eqn. 9 - $m = m + 1$ - Estimate kernel $K_x^{(k)}$ and $K_y^{(k)}$ by solving Eqn. 18 - $k = k + 1$ <p>Output: $I = I^{(k)}$</p>

3.4 Coordinate Descent

Our optimization algorithm iterates between estimating the high-res frame I , flow fields $\{w_i\}$, noise level $\{\theta_i\}$, and blur kernel K . As shown in Table 1, our optimization strategy is coordinate descent, namely sequentially optimizing each of the four sets of variable, and sweep through the entire sets several times until convergence. One sweep is called an *outer* iteration, whereas one IRLS step in optimizing a particular set of variable is called an *inner* iteration.

Although more details of the experiments will be discussed in Sect 5, we show the convergence of our system in Figure 4. In the beginning (the first row), the high-res image I is blurry (initialized as bicubic up-sampling of the low-res input), so the estimated motion $\{w_i\}$ is not very accurate. However, because of the propagation from nearby frames, the image still gets sharper in the end. As soon as a new high-res I is estimated, motion estimation, noise estimation and kernel estimation are performed, and we enter the next inner iteration of estimating I . Clearly, with more accurate estimates of other variables (especially motion), we are able to achieve sharper images.

4 PERFORMANCE BOUNDS

Intuitively super resolution becomes more challenging when noise level increases. It may be futile to perform super resolution if the blur kernel is too large and smoothes out all the high frequency components. Hence we are interested in theoretically analyzing the performance bound. Such an

TABLE 2

Notations for deriving the performance bounds.

$A(\omega)$	magnitude of signal at frequency ω
$A_1 = A(\omega_1)$	magnitude of low frequency signal
$A_2 = A(\omega_2)$	magnitude of aliasing signal
$G_{\sigma_k}(\omega)$	DFT of Gaussian blur kernel
N_H	length of high-res signal
N_L	length of low-res signal
M	downsampling ratio, $M = \frac{N_H}{N_L}$
ω_1	low frequency
$\omega_2 = \omega_1 + N_L$	aliasing high frequency
σ_n	standard deviation of imaging noise
σ_k	standard deviation of Gaussian blur kernel
u_2	translation between two high-res signals

analysis can serve as a good guideline for building up practical systems.

It is difficult, however, to exactly analyze the proposed non-linear system that iteratively estimates the motion and the image. Hence we simplify both the problem setting and the algorithm. The generative imaging process is the same as in Section 3. The input are 1D signals whose spectrum follows the power law for natural images, *i.e.* the magnitude of signal decreases w.r.t. frequency. We assume that the motion is a global translation.

In addition, we analyze the errors produced by one iteration to solve the proposed non-linear system. Given the input low-res signals, the algorithm first performs motion estimation using the input signals and then reconstructs the high-res signal using the estimated motion. For the motion estimation step, we want to analyze how noise and blur kernel affect motion estimation. For the image reconstruction step, we analyze how the imaging noise and the error in the estimated motion affect the image reconstruction. Such a semi-quantitative analysis illustrates the tradeoff we need to consider for building up the system.

The influence of the noise is easy to understand. A small noise level always results in better motion and image estimates. The influence of the blur kernel is more subtle because several factors are involved, particularly aliasing. High frequency components in the original signal become aliasing after downsampling, as shown in Figure 5. The aliasing signals appear to have different motion than the low frequency components at the low resolution grid (see analysis below) and cause errors in motion estimation. We need a large blur kernel to reduce the influence of aliasing. However, a large blur kernel boosts the noise more in image reconstruction. An optimal blur kernel should reach a tradeoff between these two conflicting requirements.

To better describe these relationships, we analyze the Cramer-Rao bounds (CRB) for both the motion estimation and the image reconstruction problems. The CRB gives the minimum mean square error (MSE) that any unbiased estimator can achieve [8], [15]. For certain problems the bounds can be achieved by the maximum likelihood estimator.

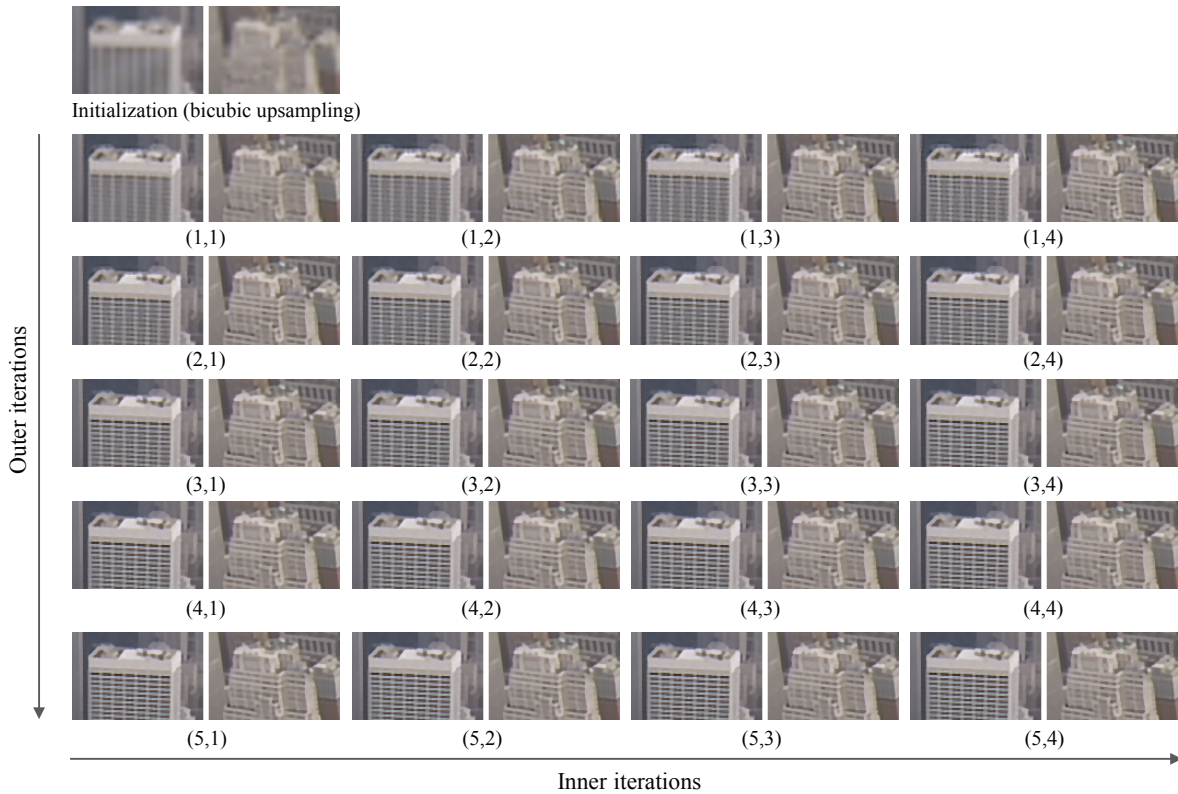


Fig. 4. **The convergence of our video super resolution algorithm.** The outer iteration consists of sweeping through estimating motion, noise level, blur kernel and the high-res frame. The inner iteration here consists of updating high-res frame, namely the iteratively reweighted least square (IRLS) procedure in solving Eqn. 9. The index $(\#i, \#j)$ shows the reconstruction result for outer iteration i and inner iteration j .

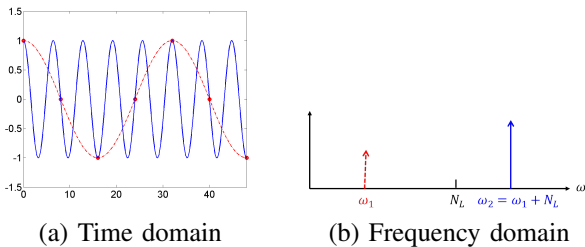


Fig. 5. **Aliasing in downsampling.** With low sampling rates, the high frequency signal (solid blue) appears to be a low frequency signal (dash red) both in the temporal domain (left) and the frequency domain (right). Note that downsampling also decrease the energy of the signal, as shown in the frequency domain.

4.1 Performance Bound for Motion Estimation with Aliasing and Noise

4.1.1 Problem Setting

A basic approach to analyze a linear system is to study the response of a particular frequency input [25]. To analyze the effect of aliasing, we pair each low frequency component of the original signal with the corresponding high frequency aliasing component. We study the effects of the noise and the blur kernel on motion estimation in Sections 4.1.2 and 4.1.3. We then combine the analysis for all the pairs to obtain the performance bound for the whole signal in

Section 4.1.4. Such analysis is exact for linear systems and can also be used to analyze non-linear systems [5].

We assume the spectrum of the original signal follows a power-law distribution [28], *i.e.* $|A(\omega)| = |\omega|^{-1.64}$, as shown in Figure 6.

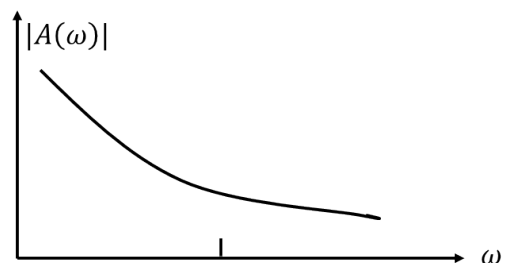


Fig. 6. **Assumed spectra of natural images.** High frequency components tend to have smaller magnitude.

We pair each low frequency component ω_1 with a corresponding high frequency aliasing component ω_2 . *i.e.*, $\omega_2 = \omega_1 + kN_L$ [25], where N_L is the length of the input low-res signal. The lowest aliasing frequency component tends to have a much larger magnitude than the other aliasing frequencies. Hence we assume that there is only one aliasing frequency component $\omega_2 = \omega_1 + N_L$, as shown in Figure 7.

The original signal with two frequency components in the time domain is

$$\begin{aligned} I_1(n) &= \frac{A_1}{N_L} e^{-\frac{i2\pi\omega_1 n}{N_H}} + \frac{A_2}{N_L} e^{-\frac{i2\pi\omega_2 n}{N_H}} \\ &= \frac{A_1}{N_L} W^{-\omega_1 n} + \frac{A_2}{N_L} W^{-\omega_2 n}, \end{aligned} \quad (19)$$

where N_L and N_H are the lengths of the low-res and original signals, ω_1 is the low frequency and ω_2 is the aliasing high frequency, and $\frac{1}{N_L}$ serves as a normalization constant. To make the derivation more succinct, we use $W = e^{\frac{i2\pi}{N_H}}$. We are using complex signals here. For real signals, the DFT coefficient at ω is the conjugate of that at $-\omega$ and we have the same number of unknowns to estimate. The derivation is the same but involves the DFT coefficients at both the positive and the negative frequencies.

The translated signal is

$$I_2(n) = \frac{A_1}{N_L} W^{-\omega_1(n-u_2)} + \frac{A_2}{N_L} W^{-\omega_2(n-u_2)}, \quad (20)$$

where u_2 is the motion on the high-res grid.

In the Discrete Fourier transforms (DFT) domain, the shift in time becomes a change in the phase of the signal. The DFTs of the original signals are

$$\tilde{I}_1(\omega) = M[A_1\delta(\omega - \omega_1) + A_2\delta(\omega - \omega_2)], \quad (21)$$

$$\tilde{I}_2(\omega) = M[A_1\delta(\omega - \omega_1)W^{u_2\omega_1} + A_2\delta(\omega - \omega_2)W^{u_2\omega_2}], \quad (22)$$

where $M = \frac{N_H}{N_L}$ is the downsampling ratio.

The effect of downsampling causes the low frequency and other frequency components to overlap with each other. The DFTs of the low-res signals are

$$\tilde{J}_1(\omega) = [G_{\sigma_k}(\omega_1)A_1 + G_{\sigma_k}(\omega_2)A_2]\delta(\omega - \omega_1) + n_1(\omega), \quad (23)$$

$$\begin{aligned} \tilde{J}_2(\omega) &= [G_{\sigma_k}(\omega_1)A_1W^{u_2\omega_1} + G_{\sigma_k}(\omega_2)A_2W^{u_2\omega_2}]\delta(\omega - \omega_1) \\ &\quad + n_2(\omega), \end{aligned} \quad (24)$$

where n_1 and n_2 are assumed to be additive white Gaussian noise (AWGN) with variance σ_n^2 , and the DFT of the Gaussian blur kernel is $G_{\sigma_k}(\omega) = e^{-\frac{\omega^2\sigma_k^2}{2}}$, where σ_k is the standard deviation of the Gaussian blur kernel.

We can obtain the pixel-wise motion estimate by correlation methods [27] but need to solve for the subpixel motion on the low-res signals. The phase of the low frequency signal is linear w.r.t. the unknown motion. However the phase of the aliasing high frequency component ($\frac{2\pi u_2 \omega_2}{N_H}$) has a nonlinear relationship w.r.t. the motion u_2 , if we treat the aliasing component as a part of the low frequency signal.

4.1.2 Treating Aliasing as Noise

We propose to model aliasing as random noise in the motion estimation process because the magnitude of the aliasing signal is relatively small compared to the low frequency signal. For natural images, their power spectra follow a power law $|A(\omega)|^2 = |\omega|^{-1.64}$ [28] and the magnitude of the low frequency coefficient is larger than the high frequency one (the ratio between $\omega_1 = 1$ and $\omega_2 = 9$ is larger than 30 for $N_H = 16$ and $M = 2$). In addition, the Gaussian blur kernel also attenuates the high frequency components more than the low frequency ones

(the ratio between $\omega_1 = 1$ and $\omega_2 = 9$ is about 2). Hence $|G_{\sigma_k}(\omega_1)A_1| \gg |G_{\sigma_k}(\omega_2)A_2|$ and it is reasonable to treat the aliasing component as AWGN.

Now the problem settings become

$$\begin{aligned} \tilde{J}_1(\omega) &= [G_{\sigma_k}(\omega_1)A_1 + G_{\sigma_k}(\omega_2)A_2]\delta(\omega - \omega_1) + n_1(\omega), \\ &= G_{\sigma_k}(\omega_1)A_1\delta(\omega - \omega_1) + n'_1(\omega), \end{aligned} \quad (25)$$

where $n'_1 = n_1 + G_{\sigma_k}(\omega_2)A_2$ has variance $\sigma_n'^2 = G_{\sigma_k}^2(\omega_2)A_2^2 + \sigma_n^2$. Similarly

$$\tilde{J}_2(\omega) = G_{\sigma_k}(\omega_1)A_1\delta(\omega - \omega_1)W^{u_2\omega_1} + n'_2(\omega), \quad (26)$$

where n'_2 has the same variance as n'_1 .

4.1.3 Cramer Rao bounds (CRB) for motion estimation

The CRB is the inverse of the Fisher information and provides a bound for unbiased estimators [8]. The Fisher information matrix describes the sensitivity of the likelihood function to the unknown parameters. We can obtain the Fisher information by taking the derivatives of the log likelihood function w.r.t. the unknown parameters. The negative log likelihood function for the input low-res signals is

$$\begin{aligned} -\log p(\tilde{J}_1, \tilde{J}_2 | A_1, u_2) &= \frac{1}{2\sigma_n'^2} \left\{ \|\tilde{J}_1(\omega_1) - G_{\sigma_k}(\omega_1)A_1\|^2 \right. \\ &\quad \left. + \|\tilde{J}_2(\omega_1) - G_{\sigma_k}(\omega_1)A_1W^{u_2\omega_1}\|^2 \right\}, \end{aligned} \quad (27)$$

where $\|\cdot\|^2$ evaluates the L2 norms for complex signals.

The Fisher information matrix for the unknown parameters $\theta = \{\text{Re}\{A_1\}, \text{Im}\{A_1\}, u_2\}$ is

$$\mathbf{I}_\theta = \frac{G_{\sigma_k}^2(\omega_1)}{\sigma_n'^2} \begin{pmatrix} 2 & 0 & \frac{A_1\omega_1 2\pi}{N_H} \\ 0 & 2 & \frac{A_1\omega_1 2\pi}{N_H} \\ \frac{A_1\omega_1 2\pi}{N_H} & \frac{A_1\omega_1 2\pi}{N_H} & \frac{A_1^2\omega_1^2 4\pi^2}{N_H^2} \end{pmatrix}, \quad (28)$$

and its inverse is

$$\mathbf{I}_\theta^{-1} = \frac{\sigma_n'^2}{2G_{\sigma_k}^2(\omega_1)} \begin{pmatrix} 1 & -3 & -\frac{N_H}{\pi A_1 \omega_1} \\ -3 & 1 & -\frac{N_H}{\pi A_1 \omega_1} \\ -\frac{N_H}{\pi A_1 \omega_1} & -\frac{N_H}{\pi A_1 \omega_1} & \frac{N_H^2}{\pi^2 A_1^2 \omega_1^2} \end{pmatrix}. \quad (29)$$

We obtain the following CRB for estimating the motion u_2 as

$$\text{var}[\hat{u}_2] \geq I_\theta^{-1}(3, 3) = \frac{N_H^2}{2\pi^2 A_1^2 \omega_1^2} \left(\sigma_n^2 e^{\frac{\omega_1^4 \sigma_k^4}{4}} + A_2^2 e^{\frac{(\omega_1^4 - \omega_2^4) \sigma_k^4}{4}} \right) \quad (30)$$

The effect of the blur kernel (σ_k) on the motion estimation is two-fold. A small blur kernel preserves the effective low frequency components for matching, boosts less the imaging noise (first term), but suppresses less the aliasing component (second term). A large blur kernel, on the other hand, preserves less the effective low frequency components, boosts more the imaging noise, but reduces the aliasing artifacts. An intermediate size blur kernel achieves the optimal performance. In addition, as shown in Figure 8, the optimal blur kernel becomes smaller as the noise level increases. When the noise dominates the aliasing signal, a

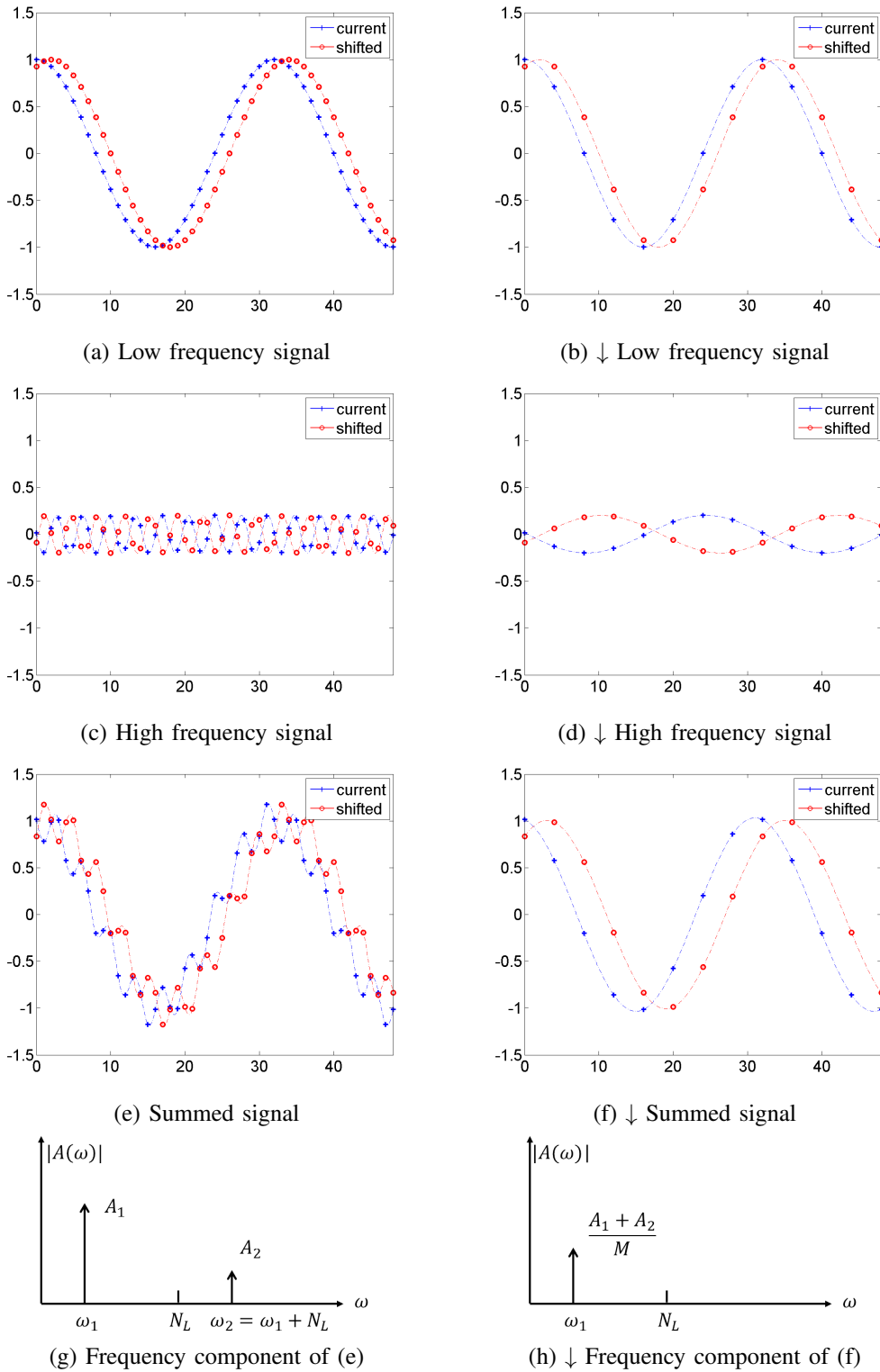


Fig. 7. **Effect of aliasing on motion estimation.** Both the low frequency and the aliasing have the same shift at the original resolution (top row). However, after downsampling, the low frequency signal has the same shift (second row left) but the downsampled aliasing appears to have a different shift (second row right). Aliasing causes incorrect interpretations of these signals and thereby cause errors to the estimated shift (third row right). After downsampling, the aliasing component has the same frequency as the low frequency signal (bottom). Note that downsampling results in a reduction in magnitude in the frequency domain.

small blur kernel will preserve the low frequency signal. When the aliasing signal dominates the noise, a large blur kernel will reduce the aliasing to help motion estimation.

4.1.4 Summing Contributions from All Frequencies

Each frequency pair provides an estimate of the unknown motion. Because of the AWGN assumption, the imaging noises at different frequencies are independent. We can obtain the final motion estimate by computing a weighted average of all the estimates from each frequency pair. The weighted average in uncorrelated noise problem is discussed in Example (6.2) in [15]. The optimal estimator combines the motion estimates at each frequency band according to their inverse variances. The CRB for the variance of the optimal estimator is

$$\text{var}[\hat{u}_2] \geq \left(\sum_{\omega=0}^{N_L-1} \frac{2\pi^2 A^2(\omega) \omega^2 / N_H^2}{\sigma_n^2 e^{\frac{\omega^4 \sigma_k^4}{4}} + A^2(\omega + N_L) e^{\frac{(\omega^4 - (\omega + N_L)^4) \sigma_k^4}{4}}} \right)^{-1} \quad (31)$$

The estimate from the DC frequency will be automatically excluded because the variance from the DC frequency is infinite (we cannot estimate motion using the DC frequency: translating the DC signal by any amount results in the same signal).

4.2 Performance Bound for Image Reconstruction with Errors in Motion

4.2.1 Maximum Likelihood (ML) Estimator with Perfect Motion

Given the perfect motion u_2 , we want to estimate the unknown $\theta_1 = \{\text{Re}\{A_1\}, \text{Im}\{A_1\}, \text{Re}\{A_2\}, \text{Im}\{A_2\}\}$. For this parameter estimation in white Gaussian noise problem, the maximum likelihood estimator achieves the lower bound predicted by the CRB [15]. The negative log likelihood function for the unknown parameters $-\log p(\tilde{J}_1, \tilde{J}_2 | A_1, A_2) =$

$$\frac{1}{2\sigma_n^2} \left\{ \left\| \tilde{J}_1(\omega_1) - G_{\sigma_k}(\omega_1) A_1 - G_{\sigma_k}(\omega_2) A_2 \right\|^2 + \left\| \tilde{J}_2(\omega_1) - G_{\sigma_k}(\omega_1) A_1 W^{u_2 \omega_1} - G_{\sigma_k}(\omega_2) A_2 W^{u_2 \omega_2} \right\|^2 \right\}. \quad (32)$$

We can derive the Fisher information matrix and its derivatives similarly and obtain the CRB for recovering A_1 is

$$\text{var}[\hat{A}_1] \geq \frac{2\sigma_n^2}{(1 - \cos(\frac{2u_2\pi}{M}))} \cdot e^{\frac{\omega_1^4 \sigma_k^4}{4}} \quad (33)$$

which means that, with perfect motion, a smaller blur kernel leads to better results and a higher noise level results in worse performance, as shown in Figure 9.

4.2.2 Performance of the ML Estimator with Motion Error

Given the estimated motion \hat{u}_2 , we want to reconstruct the original signal, including both the low and the high frequency components. Note that although the aliasing high frequency component behaves like noise in the motion

estimation process, it can be estimated once we have obtained an estimate of the motion.

Let $\hat{u}_2 = u_2 + n_{u_2}$. Because we are performing subpixel motion estimation, the motion error n_{u_2} tends to be small and we treat the error as AWGN.

We can perform Taylor expansion around the perfect motion, ignore higher-order term, and incorporate the motion estimation error into the noise term. Note that the motion estimation error has been averaged over all the frequencies and tends to be uncorrelated with the imaging noise at a particular frequency.

$$\begin{aligned} \tilde{J}_2(\omega_1) &= G_{\sigma_k}(\omega_1) A_1 W^{(u_2 + n_{u_2})\omega_1} \\ &+ G_{\sigma_k}(\omega_2) A_2 W^{(u_2 + n_{u_2})\omega_2} + n_2(\omega) \\ &\approx G_{\sigma_k}(\omega_1) A_1 W^{u_2 \omega_1} \left(1 + \frac{2\pi n_{u_2} \omega_1}{N_H}\right) \\ &+ G_{\sigma_k}(\omega_2) A_2 W^{u_2 \omega_2} \left(1 + \frac{2\pi n_{u_2} \omega_2}{N_H}\right) + n_2(\omega) \\ &= G_{\sigma_k}(\omega_1) A_1 W^{u_2 \omega_1} + G_{\sigma_k}(\omega_2) A_2 W^{u_2 \omega_2} + n_2''(\omega), \end{aligned} \quad (34)$$

where the new noise term is $n_2''(\omega) =$

$$n(\omega) + \frac{2\pi}{N_H} \left(G_{\sigma_k}(\omega_1) A_1 \omega_1 + G_{\sigma_k}(\omega_2) \omega_2 A_2 \right) n_{u_2}, \quad (35)$$

with variance

$$\sigma_{n_2''}^2 = \sigma_n^2 + \frac{4\pi^2}{N_H^2} \left(G_{\sigma_k}^2(\omega_1) A_1^2 \omega_1^2 + G_{\sigma_k}^2(\omega_2) \omega_2^2 A_2^2 \right) \text{var}[\hat{u}_2]. \quad (36)$$

We can replace the new noise variance into Eqn. 33 and obtain the CRB for recovering A_1 as

$$\text{var}[\hat{A}_1] \geq \frac{2\sigma_{n_2''}^2}{(1 - \cos(\frac{2u_2\pi}{M}))} \cdot e^{\frac{\omega_1^4 \sigma_k^4}{4}}. \quad (37)$$

Using Eqns (31) and (36), we can obtain the bound for reconstructing the low frequency component in terms of the noise level and the blur kernel in Eqn. 38. A small blur kernel will reduce the influence of noise (first term), but suppresses less the aliasing component (second term). A large blur kernel plays the opposite role. Hence an intermediate size blur kernel achieves the optimal performance, as shown in Figure (10).

Discussions. In this section, we have analyzed how the noise level and the blur kernel affect the performance of super resolution, using CRB analysis from signal processing. Our results confirm the intuition that a higher noise level makes super resolution harder (Eqn 33). We also show the blur kernel has the following influence: a small blur kernel boosts less imaging noise but suppresses less aliasing, while a large blur kernel boosts more imaging noise but suppresses more aliasing (Eqn 38). In the next section, we will empirically validate the prediction of the theoretical analysis. We show that our super resolution system has degraded performance with higher noise levels (Figure 11). We also find that the effect of the blur kernel is consistent with our theoretical analysis (Figure 12).

$$\text{var}[\hat{A}_1] \geq \frac{2\sigma_n^2}{1 - \cos(\frac{2u_2\pi}{M})} e^{-\frac{\omega_1^4 \sigma_k^4}{4}} + \frac{2(A_1^2 \omega_1^2 + A_2^2 \omega_2^2 e^{\frac{(\omega_1^4 - \omega_2^4) \sigma_k^4}{4}})}{(1 - \cos(\frac{2u_2\pi}{M}))} \times \left(\sum_{\omega=0}^{N_L-1} \frac{A^2(\omega) \omega^2}{\sigma_n^2 e^{-\frac{\omega_1^4 \sigma_k^4}{4}} + A^2(\omega + N_L) e^{\frac{(\omega^4 - (\omega + N_L)^4) \sigma_k^4}{4}}} \right)^{-1} \quad (38)$$

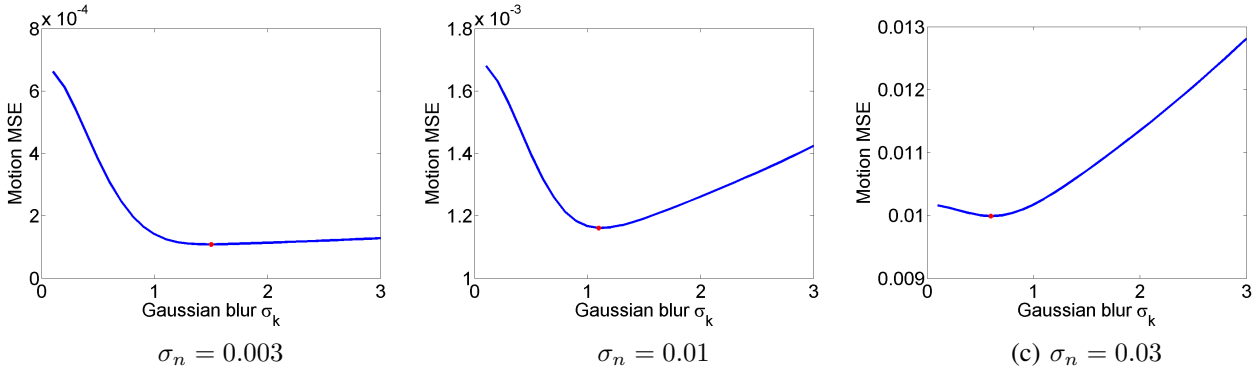


Fig. 8. Effect of Gaussian blur kernel on motion estimation. An intermediate sized blur kernel achieves the optimal performance. Red star marks the position for the optimal blur kernel. As the noise level increases, the optimal blur kernel becomes smaller to preserve the effective signal component. Noise level from left to right: $\sigma_n = 0.003$, $\sigma_n = 0.01$, and $\sigma_n = 0.03$. For the left plot, a very large blur kernel has rather low motion estimation error because of the low noise level ($\sigma_n = 0.003$).

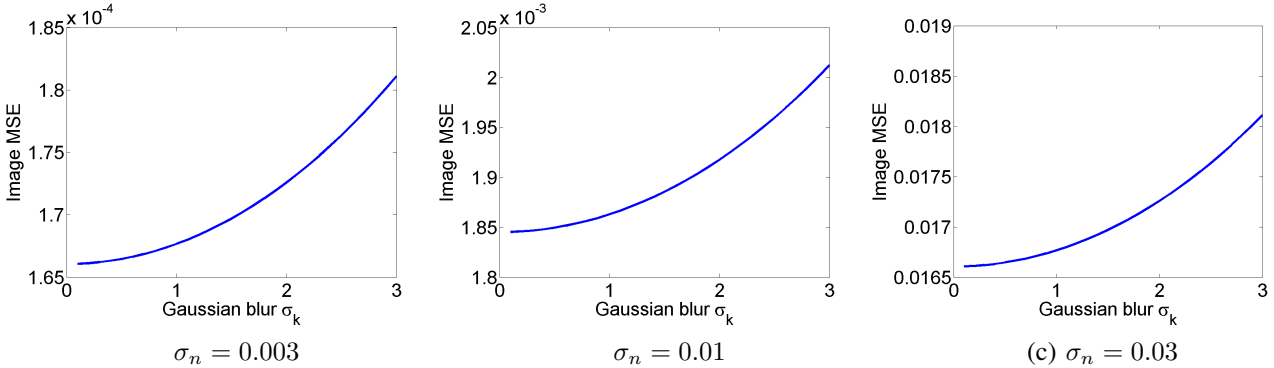


Fig. 9. Effect of Gaussian blur kernel on image reconstruction with perfect motion. A smaller blur kernel produces better image reconstruction results, because such a kernel “boosts” less noise during the inverse filtering process.

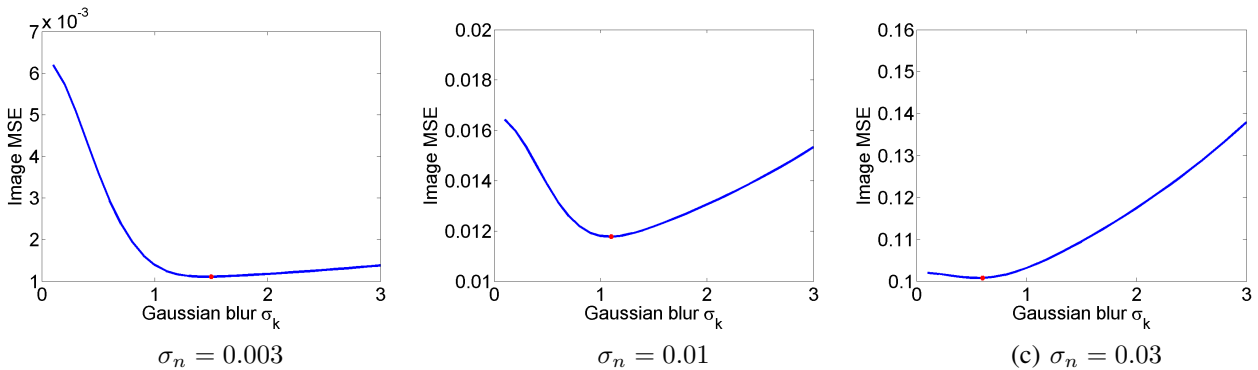


Fig. 10. Effect of Gaussian blur kernel on image reconstruction with unknown motion. Red star marks the position for the optimal blur kernel, which becomes smaller as the noise level increases. An intermediate sized blur kernel achieves the optimal performance.

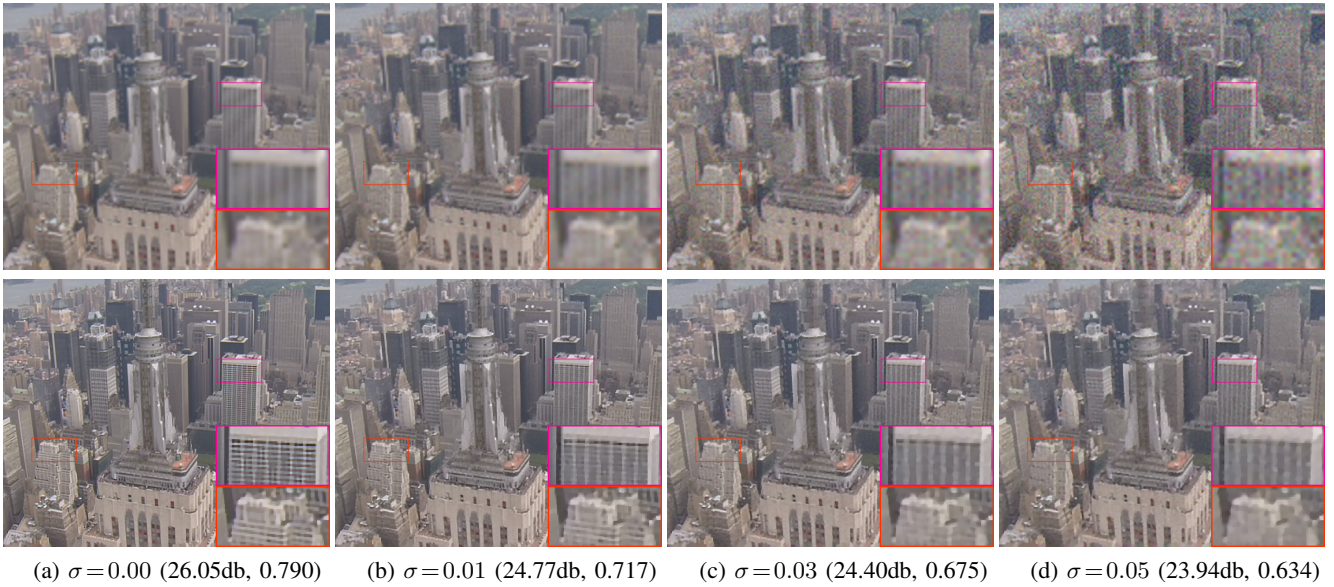


Fig. 11. **Our video super resolution system is robust to noise.** We added synthetic additive white Gaussian noise (AWGN) to the input low-res sequence, with the noise level varying from 0.00 to 0.05 (top row, left to right). The super resolution results are shown in the bottom row. The first number in the parenthesis is PSNR score and the second is SSIM score.

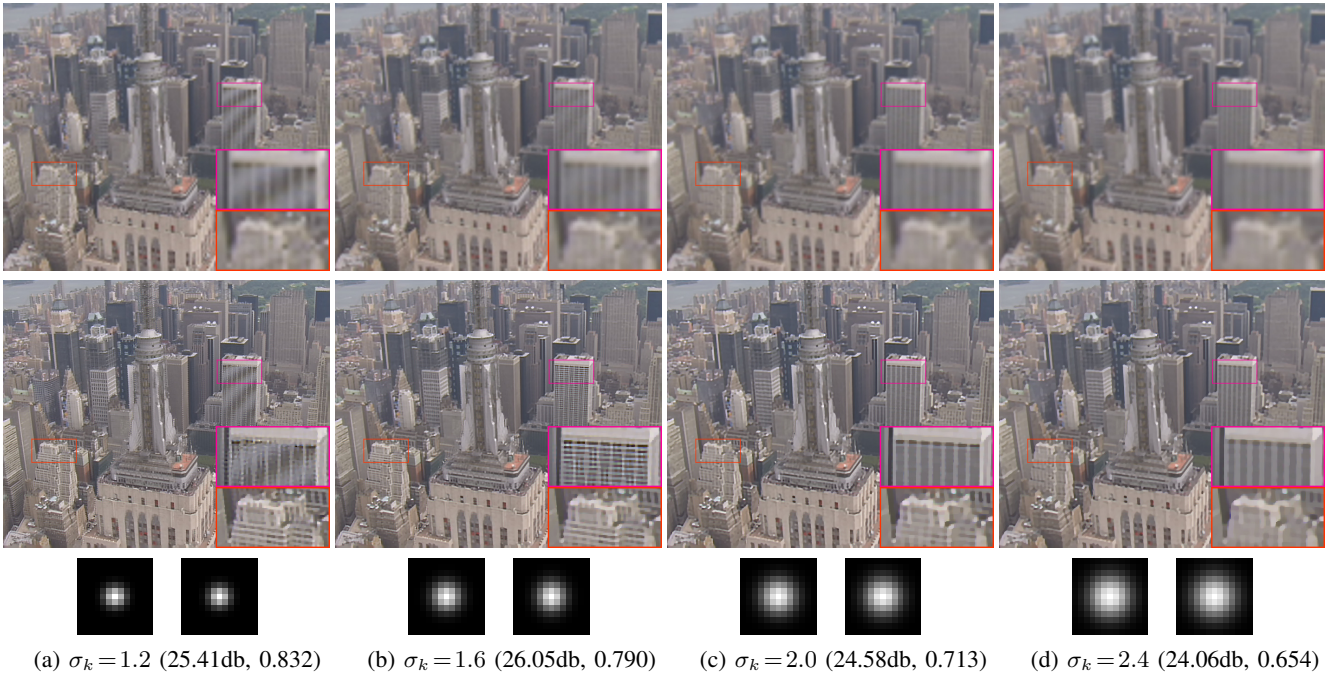


Fig. 12. **Our video super resolution system is able to estimate the PSF.** We varied the standard deviation of the blur kernel (PSF) $\sigma_k = 1.2, 1.6, 2.0, 2.4$, and our system is able to estimate the underlying PSF. Aliasing causes performance degradation for the small blur kernel $\sigma_k = 1.2$ (see text for detail), consistent with the theoretical prediction of our performance analysis. Top: bicubic up-sampling ($\times 4$); middle: output of our system; bottom: the ground truth kernel (left) and estimated kernel (right). The first number in the parenthesis is PSNR score and the second is SSIM score.

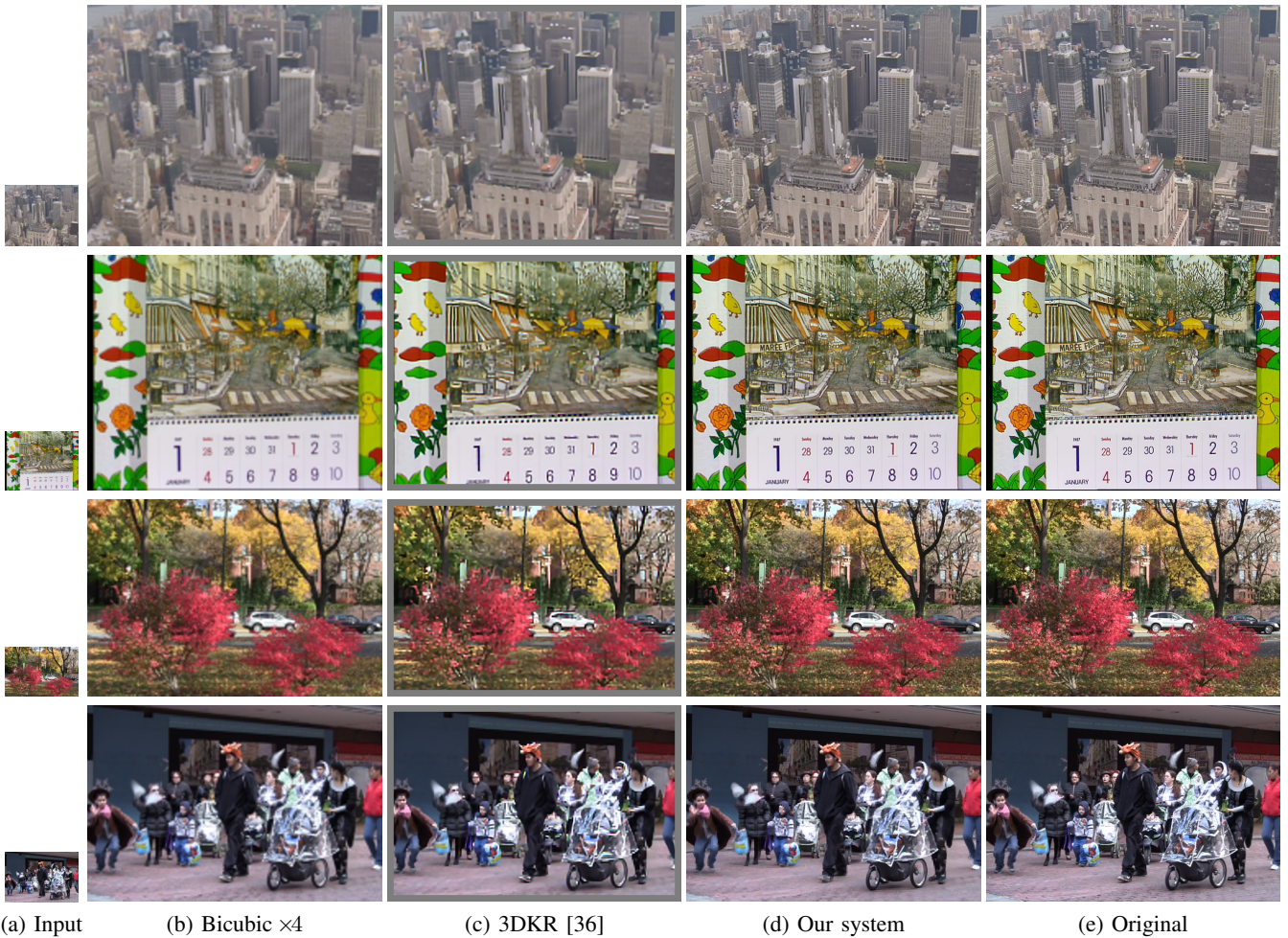


Fig. 13. **Super resolution results.** From top to bottom are *city*, *calendar*, *foliage* and *walk* sequences. The 3DKR implementation does not have valid output for pixels near the image boundaries and we fill in the gaps using gray pixels. **Please view this figure on the screen.**

5 EXPERIMENTAL RESULTS

We will first examine the performance of our system under unknown blur kernel and noise level and then compare it to state-of-the-art video super resolution methods on several real-world sequences. **Please refer to the supplemental materials or the authors' website¹ to view the super resolved sequences. Please enlarge and view Figures 11, 12 and 13 on the screen for better comparison.**

Parameter setting. We empirically set the free parameters as $\eta = 0.02$, $\lambda = 1$, $\xi = 0.7$, $\alpha = 1$ and $\beta = 0.1$.

Performance evaluation. We used the benchmark sequence *city* in video compression society to evaluate the performance. Rich details at different scales make the *city* sequence ideal to observe how different frequency components get recovered. We simulated the imaging process by first smoothing every frame of the original video with a Gaussian filter with standard deviation σ_k . We downsample the smoothed images by a factor of 4, and add white Gaussian noise with standard deviation σ_n . As we vary the blur kernel σ_k and the noise level σ_n for evaluation, we initialize our blur kernel K_x , K_y with a standard normal

distribution and initialize noise parameters θ_i using the temporal difference between frames. We use 15 forward and 15 backward adjacent frames to reconstruct a high-res image.

We first tested how our system performs under various noise levels. We fixed σ_k to be 1.6 and changed σ_n from small (0) to large (0.05). When $\sigma_n = 0$, quantization is the only source of error in the image formation process. As shown in Figure 11, our system is able to produce fine details when the noise level is low ($\sigma_n = 0.00, 0.01$). Our system can still recover major image structure even under very heavy noise ($\sigma_n = 0.05$). These results suggest that our system is robust to unknown noise. Note that the performance drop as the noise level increases is consistent with our theoretical analysis.

Next, we tested how well our system performs under various blur kernels. We gradually increase σ_k from 1.2 to 2.4 with step size 0.4 in generating the low-res input. As shown in Figure 12, the estimated blur kernels match the ground truth well. The optimal performance (in PSNR) of our system occurs for $\sigma_k = 1.6$, consistent with our theoretical analysis that a medium-sized blur kernel achieves the optimal performance. A small blur kernel generates strong

1. <http://research.microsoft.com/en-us/um/people/ce-liu/CVPR2011>

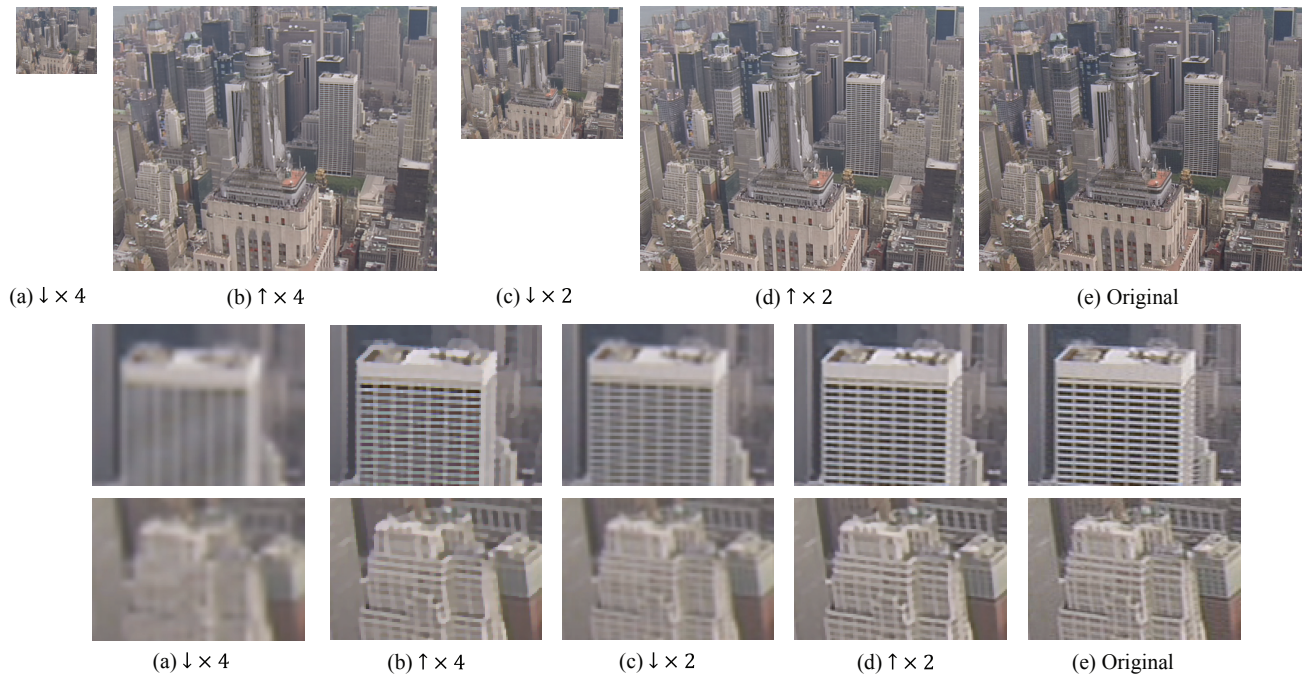


Fig. 14. **Comparison of different down-sampling rate.** (a) and (b): down-sampling and up-sampling by a factor of 4. (c) and (d): down-sampling and up-sampling by a factor of 2. (e): original frame. For (b), 15 forward and 15 backward frames were used, where as 7 forward and 7 backward frames were used for (d). Because it is down-sampling by a factor of two, we simply estimated the optical flow between input frames without re-estimating flow between the underlying high-res and the input frame. The results suggest that our system is able to handle x2 super resolution well.

aliasing, which severely degrades motion estimation and therefore prevents reconstructing the true high-frequency details. A large blur kernel removes too many image details and results in less accurate reconstructed images.

Comparison to the state of the art. We compared our method to two recent methods [31], [36] using the public implementations downloaded from the authors’ websites² and one state-of-the-art commercial software, “Video Enhancer” [1]. Since the 3DKR method [36] produced the best results amongst these methods, we only display their results due to the limited space.

We used three additional real-world video sequences, *calendar*, *foliage* and *walk* for the comparison. The results are listed in Figures 15 and 13. Although the 3DKR method has recovered the major structures of the scene, it tends to over-smooth fine details. In contrast, our system performed consistently well across the test sequences. On the *city* sequence our system recovered the windows of the tall building while 3DKR only reconstructed some blurry outlines. On the *calendar* sequence, we can easily recognize the banner “MAREE FINE” from the output of our system, while the 3DKR method failed to recover such

2. The implementation of the 3DKR method [36] does not include the last deblurring step as described in their paper. We used a state-of-the-art deconvolution method [17] to post-process its output. We used the default parameter setting of the 3DKR code to upscale the low-res video and adjusted the deconvolution method [17] to produce visually the best result for each individual sequence. The 3DKR implementation does not have valid output for pixels near the image boundaries. We filled in the gaps using gray pixels.

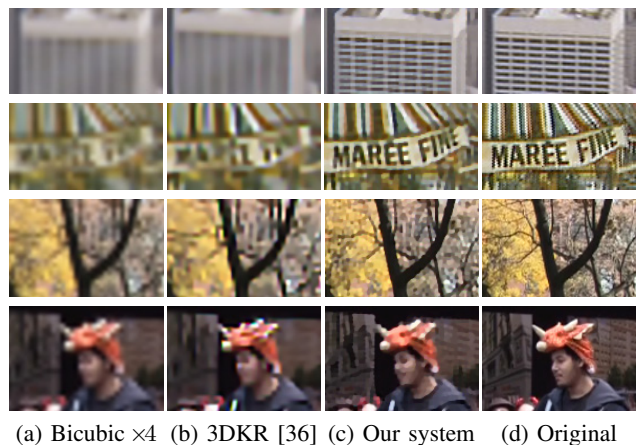


Fig. 15. Closeup of Figure 13. From top to bottom: *city*, *calendar*, *foliage* and *walk*.

detail. Moreover, our system recovered the thin branches in the *foliage* sequence and revealed some facial features for the man in the *walk* sequence. The 3DKR method, however, over-smoothed these details and produced visually less appealing results.

We also observe failures from our system. For the fast moving pigeon in the *walk* sequence, our system produced sharp boundaries instead of preserving the original motion blur. Since motion blur has not been taken into account in our system, the sparse spatial prior favors sharp boundaries in reconstructing smooth regions such as motion blur.

TABLE 3
PSNR and SSIM scores. 3DKR-b is the output of the 3DKR method before postprocessing.

PSNR	city	calendar	foliage	walk
Proposed	27.100	21.921	25.888	24.664
3DKR [36]	24.672	19.360	24.887	22.109
3DKR-b [36]	24.363	18.836	24.376	21.938
Enhancer [1]	24.619	19.115	24.476	22.303
Shan <i>et al.</i> [31]	23.828	18.539	22.858	21.018
Bicubic	23.973	18.662	24.393	22.066
SSIM				
Proposed	0.842	0.803	0.845	0.786
3DKR [36]	0.647	0.600	0.819	0.584
3DKR-b [36]	0.637	0.554	0.797	0.554
Enhancer [1]	0.677	0.587	0.803	0.604
Shan <i>et al.</i> [31]	0.615	0.544	0.747	0.554
Bicubic	0.597	0.529	0.789	0.548

Furthermore, motion blur can significantly degrade motion estimation and results in undesired artifacts.

Tables 3 summarizes the PSNR and SSIM scores³ for these methods on the video frames in Figure 13. Our system consistently outperforms other methods across all the test sequences.

Computational performance. Our C++ implementation takes about two hours on an Intel Core i7 Q820 workstation with 16 GB RAMs when super resolving a 720×480 frame using 30 adjacent frames at an up-sampling factor of 4. The computational bottle neck is solving the optical flow equation in Eqn. 17, which takes about one minute for a pair of high-res and low-res frames. Computing flow for all adjacent frames takes more than half an hour. To compare, one IRLS iteration for image reconstruction takes about two minutes.

Up-sampling by a factor of 2. For practical concerns, we only need to do up-sampling by a factor of 2, especially when standard-definition (SD, typically 720×480) videos are super-resolved to high-definition (HD, typically 1920×1080). For the up-sampling by a factor of 2, we can simply take the motion between the low-res input, resize it and magnify it by two as the true motion between the underlying high-res frame and adjacent low-res frames. This omission makes our system run at 2 minutes per frame for 720×480 videos. The difference between $\times 2$ and $\times 4$ super-resolution is illustrated in Figure 14. Clearly, sharper image details were obtained for $\times 2$ super resolution.

Real-world videos without ground truth. We applied our system to several real-world videos. As shown in Figure 16, the enhanced videos are visually more appealing and contain more details than the input.

6 DISCUSSION

When the model works and when it fails. The basic assumption of our model is that the video is generated by reshuffling pixels of a high-res frame. Therefore, our model works the best for slow and smooth motion, and would fail when there is significant lighting changes and occlusion (where the underlying assumption is broken). We also did

3. We discarded rows and columns within 20 pixels to the boundary in computing these numbers because the 3DKR method did not have valid output in these regions.



Fig. 16. **Real-world videos.** Our system is applied to enhance the resolution of real-world videos. Left: input low-res video. Right: $\times 2$ super-resolved output. Better enlarge and view on the screen.

not model motion blur, which often takes place for fast motion and/or long-exposure (for example, low light).

Aliasing: both a friend and enemy of super resolution. In this paper, we discussed in depth how aliasing would affect super resolution. Intuitively, on one hand, if there is no aliasing (namely the smoothing kernel is large enough), then there is little information to propagate from adjacent frames for generate high-frequency details. On the other hand, if the aliasing is too strong, then the false signal from aliasing would affect motion estimation and degrade super resolution. Therefore, the optimum smoothing kernel (with respect to noise level) exists. We analyzed both theoretically and empirically how the reconstruction error is affected by blur size and noise level, and these analysis results match. These results can be used as guidelines for designing super resolution systems.

Future research directions. Future work will incorporate the recent developments in each sub problem, such as high-order image prior model [29], non-local motion prior model [35], feature matching for fast moving objects [7], [33], [40] and advanced inference methods for estimating the spatially-variant blur kernel [9], [39]. Our system cannot deal with large occlusions, for which the layered representation [38] is more suitable. For scenes with changing illuminations, inferring the illumination and super resolving the

surface properties can relax our assumption that every input frame can be generated by reshuffling the center frame. Motion blur can be incorporated into the generative model too. Furthermore, our system does not model compression artifacts, which are ubiquitous in low-bit compressed videos on the web and act like high-frequency false signals. We have developed a non-causal system to jointly estimate the optical flow and the original video sequence using the encoded bit streams [34]. Incorporating the compression process will make our system more robust. Finally it is of great practical value to theoretically predict how much a given video sequence can be super resolved.

7 CONCLUSION

In this paper we have demonstrated that our adaptive video super resolution system based on a Bayesian probabilistic model is able to reconstruct original high-res images with great details. Our system is robust to arbitrary motion, unknown noise level and/or unknown blur kernel because we jointly estimate motion, noise and blur with the high-res image using sparse image/flow/kernel priors. Very promising experimental results suggest that our system consistently outperform the state-of-the-art methods on a variety of real-world sequences. On the theoretical side, we have performed a two step analysis of how noise level and blur kernel affect the performance using the Cramer-Rao bounds. Our analytical results are consistent with our experiments, indicating that they can be good guidelines for analyzing super resolution systems.

ACKNOWLEDGMENTS

The second author would like to thank Dr. Lo-Bin Chang and Mr. Jianshu Chen for their help with the mathematical derivations of the performance bounds.

REFERENCES

- [1] <http://www.infognition.com/videoenhancer/>, Sep. 2010. Version 1.9.5.
- [2] S. Baker and T. Kanade. Super-resolution optical flow. Technical report, CMU, 1999.
- [3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 24(9):1167–1183, 2002.
- [4] B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *European Conference on Computer Vision*, 1996.
- [5] S.A. Billings and K.M. Tsang. Spectral analysis for non-linear systems, part i: Parametric non-linear spectral analysis. *Mechanical Systems and Signal Processing*, 3(4):319339, oct. 1989.
- [6] T. Brox, A. Bruhn, N. Papenberger, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*, 2004.
- [7] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 33(3):500–513, March 2011.
- [8] S. Clayton and R. Nowak. The cramer-rao lower bound. <http://cnx.org/content/m11429/1.4/>, May 2004.
- [9] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W.T. Freeman. Removing camera shake from a single photograph. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 25:787–794, 2006.
- [10] R. Fransens, C. Strecha, and L. J. Van Gool. Optical flow based super-resolution: A probabilistic approach. *Computer Vision and Image Understanding*, 106:106–115, 2007.
- [11] 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 Transaction on Image Processing*, 6(12):1621–1633, Dec. 1997.
- [12] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 16:185–203, Aug. 1981.
- [13] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP*, 53:231–239, 1991.
- [14] N. Joshi, R. Szeliski, and D.J. Kriegman. PSF estimation using sharp edge prediction. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.
- [15] S. M. Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
- [16] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, 1996.
- [17] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 26, 2007.
- [18] A. Levin, Y. Weiss, F. Durand, and W.T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, 2009.
- [19] Z. Lin and H-Y Shum. Fundamental limits of reconstruction based superresolution algorithms under local translation. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 26:83–97, 2004.
- [20] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, MIT, 2009.
- [21] C. Liu and W. T. Freeman. A high-quality video denoising algorithm based on reliable motion estimation. In *European Conference on Computer Vision*, 2010.
- [22] C. Liu and D. Sun. A bayesian approach to adaptive video super resolution. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.
- [23] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 30(2):299–314, 2008.
- [24] N. Nguyen, G. Golub, and P. Milanfar. Blind restoration/superresolution with generalized cross-validation using gauss-type quadrature rules. In *Asilomar Conference on Signals, Systems, and Computers*, 1999.
- [25] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck. *Discrete-time signal processing (2nd ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.
- [26] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, 2003.
- [27] D. Robinson and P. Milanfar. Statistical performance analysis of super-resolution. *IEEE Transaction on Image Processing*, 15(6):1413–1428, jun. 2006.
- [28] S. Roth. *High-Order Markov Random Fields for Low-Level Vision*. PhD thesis, Brown University, 2007.
- [29] S. Roth and M.J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, April 2009.
- [30] R.R. Schultz and R.L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transaction on Image Processing*, 5(6):996–1011, June 1996.
- [31] Q. Shan, Z. Li, J. Jia, and C-K Tang. Fast image/video upsampling. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 27(5):153, 2008.
- [32] F. Sroubek, G. Cristobal, and J Flusser. Simultaneous super-resolution and blind deconvolution. *Journal of Physics: Conference Series*, 124(1), 2008.
- [33] H. Su, Y. Wu, and J. Zhou. Super-resolution without dense flow. *IEEE Transaction on Image Processing*, 21(4):1782–1795, april 2012.
- [34] D. Sun and C. Liu. Non-causal temporal prior for video deblocking. In *ECCV*, 2012.
- [35] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.
- [36] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *IEEE Transaction on Image Processing*, 18(9):1958–1975, Sep. 2009.
- [37] R. Y. Tsai and T. S. Huang. Multiframe image restoration and registration. In *Advances in Computer Vision and Image Processing*, 1984.
- [38] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transaction on Image Processing*, 3(5):625–638, September 1994.

- [39] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. *International Journal of Computer Vision*, 98(2):168–186, 2012.
- [40] L. Xu, J. Jia, and Y. Matsushita. Motion detail preserving optical flow estimation. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 34(9):1744–1757, 2012.



Ce Liu received the BS degree in automation and the ME degree in pattern recognition from the Department of Automation, Tsinghua University in 1999 and 2002, respectively. After receiving the PhD degree from the Massachusetts Institute of Technology in 2009, he is a researcher at Microsoft Research New England, and an adjunct assistant professor at Boston University. From 2002 to 2003, he worked at Microsoft Research Asia as an assistant researcher. His

research interests include computer vision, computer graphics, and machine learning. He has published more than 50 papers in the top conferences and journals in these fields. He received a Microsoft Fellowship in 2005, the Outstanding Student Paper award at the Advances in Neural Information Processing Systems (NIPS) in 2006, a Xerox Fellowship in 2007, and the Best Student Paper award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009. He is a member of the IEEE.



Deqing Sun received the BEng degree in Electronic and Information Engineering from Harbin Institute of Technology, the MPhil degree in Electronic Engineering from the Chinese University of Hong Kong, and the MS and PhD degrees in Computer Science from Brown University. He was a research intern at Microsoft Research New England from October to December 2012. His research interests include computer vision, machine learning, signal and image processing, particularly motion estimation and segmentation and the applications

to video processing. He has been a postdoctoral fellow at Harvard University since August 2012. He is a member of the IEEE.