Video Super-Resolution Using Controlled Subpixel Detector Shifts

Moshe Ben-Ezra, Assaf Zomet, and Shree K. Nayar

Abstract—Video cameras must produce images at a reasonable frame-rate and with a reasonable depth of field. These requirements impose fundamental physical limits on the spatial resolution of the image detector. As a result, current cameras produce videos with a very low resolution. The resolution of videos can be computationally enhanced by moving the camera and applying super-resolution reconstruction algorithms. However, a moving camera introduces motion blur, which limits super-resolution quality. We analyze this effect and derive a theoretical result showing that motion blur has a substantial degrading effect on the performance of super-resolution. The conclusion is that, in order to achieve the highest resolution, motion blur should be avoided. Motion blur can be minimized by sampling the space-time volume of the video in a specific manner. We have developed a novel camera, called the "jitter camera," that achieves this sampling. By applying an adaptive super-resolution algorithm to the video produced by the jitter camera, we show that resolution can be notably enhanced for stationary or slowly moving objects, while it is improved slightly or left unchanged for objects with fast and complex motions. The end result is a video that has a significantly higher resolution than the captured one.

Index Terms—Sensors, jitter camera, jitter video, super-resolution, motion blur.

1 Why Is High-Resolution Video Hard?

Improving the spatial resolution of a video camera is different from doing so with a still camera. Merely increasing the number of pixels of the detector reduces the amount of light received by each pixel and, hence, increases the noise. With still images, this can be overcome by prolonging the exposure time. In the case of video, however, the exposure time is limited by the desired frame-rate. The amount of light incident on the detector can also be increased by widening the aperture, but with a significant reduction of the depth of field. The spatial resolution of a video detector is therefore limited by the noise level of the detector, the frame-rate (temporal resolution), and the required depth of field.[1] Our purpose is to make a judicious use of a given detector that will allow a substantial increase of the video resolution by a resolution-enhancement algorithm.

Fig. 1 shows a *continuous* space-time video volume. A slice of this volume at a given time instance corresponds to the image appearing on the image plane of the camera at this time. This volume is sampled both spatially and temporally, where each pixel integrates light over time and space. Conventional video cameras sample the volume in a simple way, as shown in Fig. 1a, with a regular 2D grid of pixels integrating over regular temporal intervals and at fixed spatial locations. An alternative sampling of the space-time volume is shown in Fig. 1b. The 2D grid of pixels integrates over the same temporal intervals, but at different spatial locations. Given a 2D image detector, how should we sample the space-time volume to obtain the highest spatial resolution?[2]

There is a large body of work on resolution enhancement by varying spatial sampling, commonly known as super-resolution reconstruction [4], [5], [7], [9], [13], [18]. Super-resolution algorithms typically assume that a set of displaced images are given as input. With a video camera, this can be achieved by moving the camera while capturing the video. However, the camera's motion introduces motion blur. This is a key point in this paper: In order to use super-resolution with a conventional video camera, the camera must move, but when the camera moves, it introduces motion blur which reduces resolution.

It is well-known that an accurate estimation of the motion blur parameters is nontrivial and requires strong assumptions about the camera motion during integration [2], [13], [16], [20]. In this paper, we show that *even when an accurate estimate of the motion blur parameters is available*, motion blur has a significant influence on the super-resolution result. We derive a *theoretical lower bound*, indicating that the expected performance of *any* super-resolution reconstruction algorithm deteriorates as a function of the motion blur magnitude. The conclusion is that, in order to achieve the highest resolution, motion blur should be *avoided*.

To achieve this, we propose the "jitter camera," a novel video camera that samples the space-time volume at different locations without introducing motion blur. This is done by instantaneously shifting the detector (e.g., CCD) between temporal integration periods, rather than continuously moving the entire video camera during the integration

1. The optical transfer function of the lens also imposes a limit on resolution. In this paper, we ignore this limit as it is several orders of magnitudes above the current resolution of video.

2. Increasing the temporal resolution [19] is not addressed in this paper.

- *M. Ben-Ezra is with Siemens Corporate Research, 755 College Rd. East, Princeton NJ, 08540. E-mail: moshe.ben-ezra@siemens.com.*
- *A Zomet and S.K. Nayar are with the Computer Science Department, Columbia University, 1214 Amsterdam Ave, MC 0401, New York, NY 10027. E-mail: {zomet, nayar}@cs.columbia.edu.*

Manuscript received 10 Apr. 2004; revised 6 Oct. 2004; accepted 4 Nov. 2004; published online 14 Apr. 2005.
Recommended for acceptance by M. Srinivasan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0171-0404.

0162-8828/05/$20.00 © 2005 IEEE Published by the IEEE Computer Society

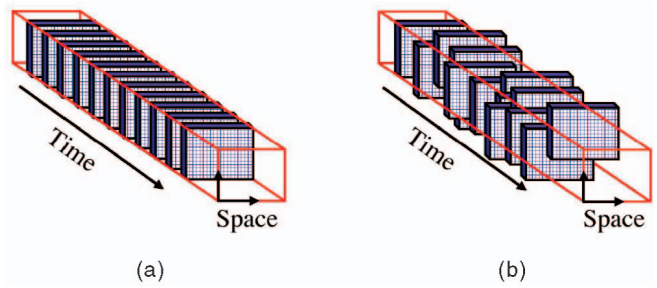

Fig. 1. Conventional video cameras sample the continuous space-time volume at regular time intervals and fixed spatial grid locations as shown in (a). The space-time volume can be sampled differently, for example, by varying the location of the sampling grid as shown in (b) to increase the resolution of the video. A moving video only approximates (b) due to motion blur.

periods. We have built a jitter camera and developed an adaptive super-resolution algorithm to handle complex scenes containing multiple moving objects. By applying the algorithm to the video produced by the jitter camera, we show that resolution can be enhanced significantly for stationary or slowly moving objects, while it is improved slightly or left unchanged for objects with fast and complex motions. The end result is a video that has higher resolution than the captured one.

2 How Bad Is Motion Blur for Super-Resolution?

The influence of motion blur on super-resolution is well understood when all input images undergo the same motion blur [1], [10]. It becomes more complex when the input images undergo *different* motion blurs and details that appear blurred in one image appear sharp in another image. We address the influence of motion blur for any combination of blur orientations.

Super-resolution algorithms estimate the high resolution image by modeling and inverting the imaging process. Analyzing the influence of motion blur requires a definition for super-resolution "hardness" or the "invertibility" of the imaging process. We use a linear model for the imaging process [1], [7], [9], [13], where the intensity of a pixel in the input image is presented as a linear combination of the intensities in the unknown high resolution image:

$$\vec{y} = A\vec{x} + \vec{z}, \tag{1}$$

where $\vec{x}$ is a vectorization of the unknown discrete high resolution image, $\vec{y}$ is a vectorization of all the input images, and the imaging matrix A encapsulates the camera displacements, blur, and decimation [7]. The random variable $\vec{z}$ represents the uncertainty in the measurements due to noise, quantization error, and model inaccuracies.

Baker and Kanade [1] addressed the invertibility of the imaging process in a noise-free scenario, where $\vec{z}$ represents the quantization error. In this case, each quantized input pixel defines two inequality constraints on the super-resolution solution. The combination of constraints forms a *volume of solutions* that satisfy all quantization constraints. Baker and Kanade suggest using the *volume of solutions* as a measure of uncertainty in the super-resolution solution. Their paper [1] shows the benefits in measuring the *volume of solutions* over the standard matrix conditioning analysis.

We measure the influence of motion blur by the volume of solutions. To keep the analysis simple, the following assumptions are made: First, the motion blur in each input image is induced by a constant velocity motion. Different input images may have different motion blur orientations. Second, the optical blur is shift-invariant. Third, the input images are related geometrically by a 2D translation. Fourth, the number of input pixels equals the number of output pixels. Under the last assumption, the dimensionality n^2 of $\vec{x}$ equals the dimensionality of $\vec{y}$. Since the uncertainty due to quantization is an n^2-dimensional unit cube, the volume of solutions for a given imaging matrix A can be computed from the absolute value of its determinant

$$vol(A) = \left|\frac{1}{|A|}\right|. \tag{2}$$

In Appendix A, we derive a simplified expression for $|A|$ as a function of the imaging parameters. This allows for an efficient computation of $vol(A)$, as well as a derivation of a lower bound on $vol(A)$ as a function of the extent of motion blur.

Since the volume of solutions $vol(A)$ depends on the image size, which is n^2, we define in Appendix A (8) a function $s(A)$ such that

$$vol(A) \propto s(A)^{n^2}.$$

$s(A)$ has two desirable properties for analyzing the influence of motion blur. First, it is independent of the camera's optical transfer function and the detector's integration function, and normalized to one when there is no motion blur and the camera displacements are optimal (Appendix B). Second, $vol(A)$ is exponential in the image size whereas $s(A)$ is normalized to account for the image size.

Fig. 2 shows $s(A)$ as a function of the lengths of the motion blur trajectories. Specifically, let $\vec{l}_j$ be a vector describing the motion blur trajectory for the jth input image: During integration, the projected image moves at a constant velocity from $-\frac{\vec{l}_j}{2}$ to $\frac{\vec{l}_j}{2}$. Each graph in Fig. 2 shows the value of $s(A)$ as a function of the length of the four motion blur trajectories $\{\|\vec{s}_j\|\}_{j=0}^{3}$. The different graphs correspond to different configurations of blur orientations in four input images. The graphs were computed for optimal camera displacements (see Appendix B) and magnification factor 2.

It can be seen that, in **all** selected motion blur configurations, $s(A) \propto vol(A)^{\frac{1}{n^2}}$ increases as a function of the length of the motion blur trajectories $\{\|\vec{l}_j\|\}$. The thick blue line is the lower bound of $s(A)$, whose derivation can be found in Appendix A. This bound is for any configuration of blur orientations and any camera displacements.

The findings above confirm that, at least for our assumptions, *any motion blur is bad for super-resolution* and *the larger the motion blur, the larger the volume of solutions*.

Fig. 3a shows super-resolution results of simulations with and without motion blur. The simulated input images were obtained by displacing, blurring, and subsampling the ground truth image. The blurs and displacements were provided to the super-resolution as input. As can be seen in

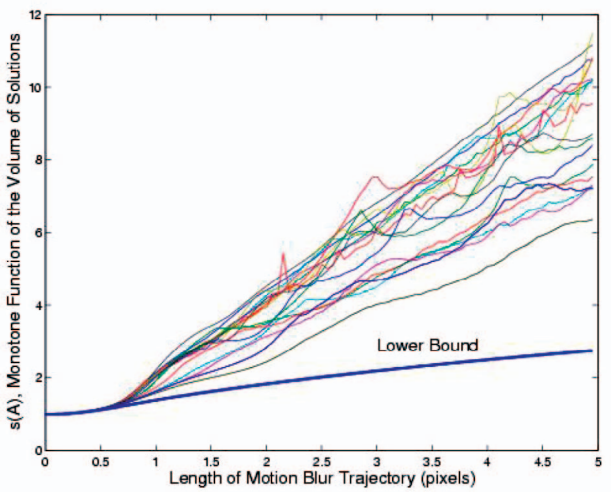

Fig. 2. We measure the super-resolution "hardness" by the volume of plausible high-resolution solutions [1]. The volume of solutions is proportional to $s(A)^{n^2}$, where n^2 is the high resolution image size. The graphs show the value of $s(A)$ as a function of the length of the motion-blur trajectories $\{\|l_j\|\}_{j=0}^{3}$. We show a large number of graphs computed for different configurations of blur orientations. The thick graph (blue line) is the lower bound of $s(A)$ for any combination of motion blur orientations. In all shown configurations, the motion blur has a significant influence on $s(A)$ and, hence, on the volume of solutions. The increase in the volume of solutions can explain the increase in reconstruction error in super-resolution shown in Fig. 3.

Fig. 3, even with motion blur as small as 3.5 pixels, the super-resolution result is degraded such that some of the letters are unreadable. Fig. 3b presents the RMS error in the reconstructed super-resolution image as a function of the extent of the motion blur. It can be seen that the RMS error increases as a function of the motion blur magnitude. This effect is consistent with the theoretical observations made above.

3 Jitter Video: Sampling without Motion Blur

Our analysis showed that sampling with minimal motion blur is important for super-resolution. Little can be done to prevent motion blur when the camera is moving[3] or when objects in the scene are moving. Therefore, our main goal is to sample at different spatial locations while avoiding motion blur in static regions of the image.

The key to avoiding motion blur is *synchronous* and *instantaneous* shifts of the sampling grid between temporal integration periods, rather than a continuous motion during the integration periods. In Appendix B, we show that the volume of solutions can be minimized by properly selecting the grid displacements. For example, in the case of four input images, one set of optimal displacements is achieved by shifting the sampling grid by half a pixel horizontally and vertically. Implementing these abrupt shifts by moving a standard video camera with a variable magnification factor is nontrivial.[4] Hence, we propose to implement the shifts of the sampling grid inside the camera.

3. Small camera shakes can be eliminated by *optical* lens stabilization systems, which stabilize the image before it is integrated.

4. A small uniform image displacement can be approximated by rotating the camera about the X, Y axes. However, the rotation extent depends on the exact magnification factor of the camera, which is hard to obtain. In addition, due to camera's mass, abrupt shifting of the camera is challenging.

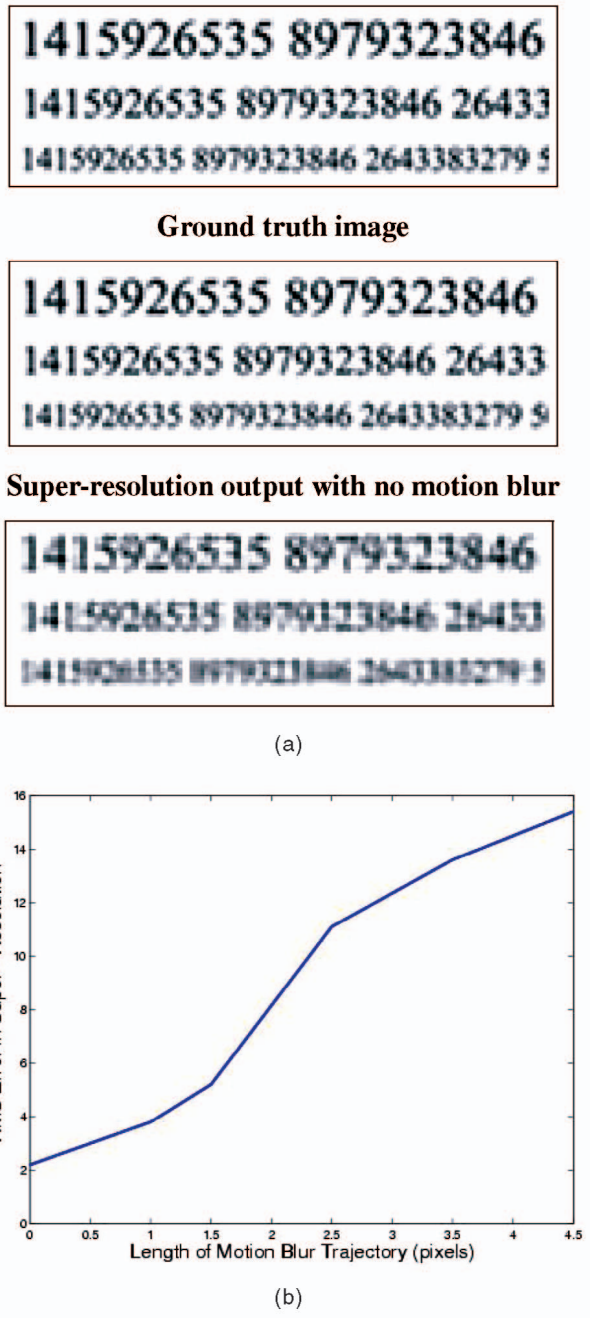

Fig. 3. The effect of motion blur on super-resolution with a known simulated motion blur. (a) The top image is the original ground-truth image. The middle image is the super-resolution result for four simulated input images with no motion blur. This image is almost identical to the ground truth image. The bottom image is a super-resolution result for four simulated input images with motion blur of 3.5 pixels. Two images with horizontal blur and two with vertical blur were used. The algorithm used the known simulated motion blur kernels and the known displacements. The degradation in the super-resolution result due to motion blur is clearly visible. (b) The graph shows the gray level RMS error in the super-resolution image as a function of motion blur trajectory length.

Fig. 4 shows two possible ways to shift the sampling grid instantaneously. Fig. 4a shows a purely mechanical design, where the detector (e.g., CCD) is shifted by actuators to change the sampling grid location. If the actuators are fast and are activated *synchronously* with the reading cycle of the detector, then the acquired image will have no motion blur due to the shift of the detector. Fig. 4b shows a mechanical-optical

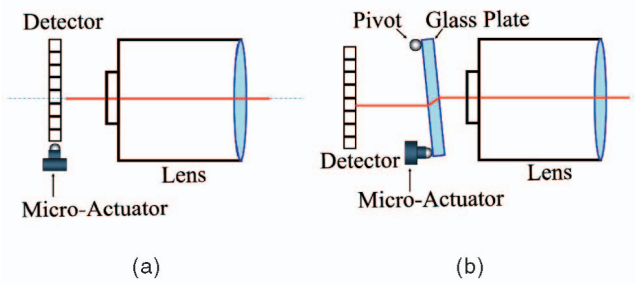

Fig. 4. A jitter video camera shifts the sampling grid accurately and instantaneously. This can be achieved using micro-actuators, which are both fast and accurate. The actuator can shift the detector as shown in (a), or it can be used to operate a simple optical device, such as the tilted glass plate shown in (b), in order to optically move the image with respect to the static detector.

design. A flat thin glass plate is used to shift the image over the detector. An angular change of a $1mm$ thick plate by one degree shifts the image by $5.8\mu m$, which is of the order of a pixel size. Since the displacement is very small relative to the focal length, the change of the optical path length results with negligible effect on the focus (the point spread area is much smaller than the area of a pixel). The mechanical-optical design shown Fig. 4b has been used for high-resolution still-imaging, for example, by Pixera [6], where video-related issues such as motion blur and dynamic scenes do not arise.

An important point to consider in the design of a jitter camera is the quality of the camera lens. With standard video cameras, the lens-detector pair is matched to reduce spatial aliasing in the detector. For a given detector, the matching lens attenuates the spatial frequencies higher than the Nyquist frequency of the detector. For a jitter camera, higher frequencies are useful since they are exploited in the extraction of the high resolution video. Hence, the selected lens should match a detector with a higher (the desired) spatial resolution.

4 The Jitter Camera Prototype

To test our approach, we have built the jitter camera prototype shown in Fig. 5. This camera was built using a standard $16mm$ television lens, a Point-Grey [17] Dragon-Fly

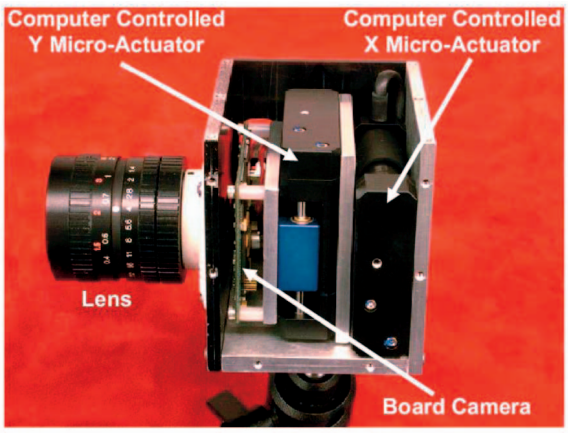

Fig. 5. The jitter camera prototype shown with its cover open. The mechanical micro-actuators are used for shifting the board camera. The two actuators and the board camera are synchronized such that the camera is motionless during integration time.

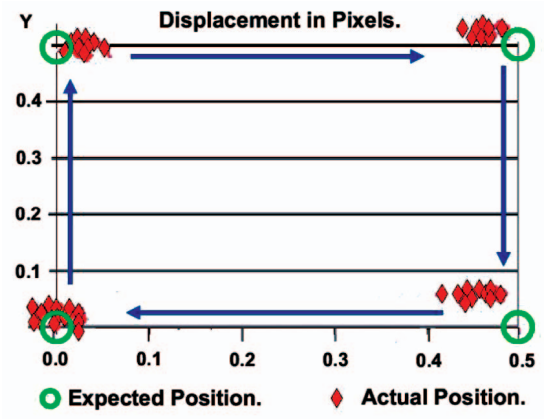

Fig. 6. Accuracy of the jitter mechanism. The detector moves one step at a time along the path shown by the blue arrows. The green circles show the expected position of exactly half a pixel displacement and the red diamonds show the actual position over multiple cycles. We can see that the accuracy was less than a tenth of a pixel. We can also see that the jitter mechanism returns very accurately to its zero position, hence, preventing excessive error accumulation over multiple cycles.

board camera, and two Physik Instrumente [8] micro-actuators. The micro-actuators and the board camera were controlled and synchronized by a Physik Instrumente Mercury stand-alone controllers (not shown).

The jitter camera is connected to a computer using a standard firewire interface and, therefore, it appears to be a regular firewire camera.

We used, in our prototype, two DC-motor actuators, which enable a frame-rate of approximately eight frames per second. Newly developed piezoelectric-based actuators can offer much higher speed than DC-motor based actuators. Such actuators are already used for *camera shake compensation* by Minolta [12], however, they are less convenient for prototyping at this point in time.

The camera operates as follows:

1. At power up, the actuators are moved to a fixed home-position.
2. For each sampling position in [(0,0),(0,0.5),(0.5,0.5),(0.5,0)] pixels do
   - Move the actuators to the next sampling position.
   - Bring the actuators to a full stop.
   - Send a trigger signal to the camera to initiate frame integration and wait during integration duration.
   - When the frame is ready, the camera sends it to the computer over the Firewire interface.
3. End loop.
4. Repeat process from step (2).

To evaluate the accuracy of the jitter mechanism, we captured a sequence of images with the jitter camera, computed the motion between frames to a subpixel accuracy [3], and compared the computed motion to the expected value. The results are shown in Fig. 6. The green circles show the expected displacements and the red diamonds show the actual displacements over multiple cycles. We can see that the accuracy of the jitter mechanism was better than 0.1 pixel. We can also see that, while some error is accumulated along the path, the camera accurately returns to its zero position, thus preventing drift.

The resolution of the computed high-resolution video was $1,280 \times 960$, which has four times the number of pixels compared to the resolution of the input video, which was 640×480. This enhancement upgrades an NTSC grade camera to an HTDV grade camera while maintaining the depth of field and the frame-rate of the original camera.

5 Adaptive Super-Resolution for Dynamic Scenes

Given a video sequence captured by a jitter camera, we would like to compute a high resolution video using super-resolution. We have chosen iterated-back-projection [9] as the super-resolution algorithm. Iterated-back-projection was shown in [4] to produce high quality results and is simple to implement for videos containing complex scenes. The main challenge in our implementation is handling multiple motions and occlusions. Failing to cope with these problems results in strong artifacts that render the output useless.

To address these problems, we compute the image motion in small blocks and detect blocks suspected of having multiple motions. The adaptive super-resolution algorithm maximizes the use of the available data for each block.

5.1 Motion Estimation in the Presence of Aliasing

The estimation of image motion should be robust to outliers, which are mainly caused by occlusions and multiple motions within a block. To address this problem, we use the Tukey M-estimator error function [11]. The Tukey M-estimator depends on a scale parameter σ, the standard deviation of the gray-scale differences of correctly-aligned image regions (*inlier regions*).

Due to the under-sampling of the image, gray-scale image differences in the inlier regions are *dominated by aliasing* and are especially significant near sharp image edges. Hence, we approximate the standard deviation of the gray-scale differences σ in each block from the standard deviation of the aliasing σ_a in the block as $\sigma = \sqrt{2}\sigma_a$. This approximation neglects the influence of noise and makes the simplifying assumption that the aliasing effects in two aligned blocks are statistically uncorrelated. In the following, we describe the approximation for the standard deviation of the aliasing in each block σ_a, using results on the statistics of natural images.

Let f be a high resolution image, blurred and decimated to obtain a low resolution image g:

$$g = (f * h)\downarrow,$$

where $*$ denotes convolution and $\downarrow$ denotes subsampling. Let s be a perfect rect low pass filter. The aliasing in g is given by:

$$(f * h - f * s * h)\downarrow = f * h * (\delta - s)\downarrow .$$

The band-pass filter $h * (\delta - s)$ can, hence, be used to simulate aliasing. For the motion estimation, we need to estimate σ_a, the standard deviation of the response of this filter to blocks of the *unknown* high resolution image. We use the response of this filter to the aliased low resolution input images to estimate σ_a. Let σ_0 be the standard deviation of the filter response to an input block. Testing with a large number of images, we found that σ_a can be approximated to be a linear function of σ_0. Similar results for nonaliased images were shown by Simoncelli [21] for various band-pass filters at different scales. For blocks of size 16×16 pixels, the linear coefficient was in the range $[0.5, 0.7]$. In the experiments, we set $\sigma_a = 0.7\sigma_0$, which was sufficient for our purpose.

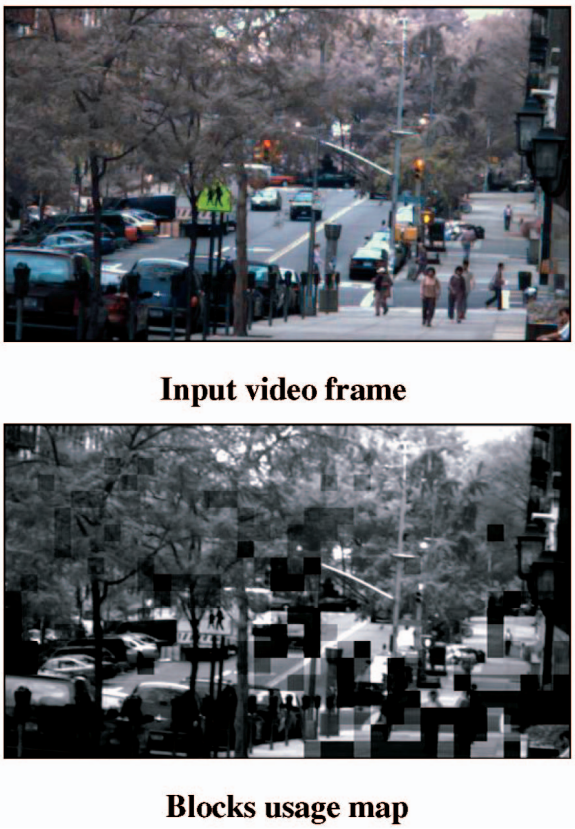

Fig. 7. Adaptation of the super-resolution algorithm to moving objects and occlusions. The image on top shows one frame from a video sequence of a dynamic scene. The image on bottom is a visualization of the number of valid blocks, from four frames, used by the algorithm in each block. We darkened blocks where the algorithm used less than four valid blocks due to occlusions.

5.2 Adaptive Data Selection

We use the scale estimate σ from the previous section to differentiate between blocks with a single motion and blocks that may have multiple motions and occlusions. A block in which the SSD error exceeds 3σ is excluded from the super-resolution calculation. In order to double the resolution (both horizontally and vertically), three additional valid blocks are needed for each block in the current frame. Depending on the timing of the occlusions, these additional blocks could be found in previous frames only, in successive frames only, both, or not at all. We therefore search for valid blocks in both temporal directions and select the blocks which are valid and closest in time to the current frame.

In blocks containing a complex motion, it may happen that less than four valid blocks are found within the temporal search window. In this case, although the super-resolution image is under-constrained, iterated-back-projection produces reasonable results [4]. Fig. 7 shows an example from an outdoor video sequence containing multiple moving objects. On bottom is a visualization of the number of valid blocks used for each block in this frame. Blocks where less than four valid blocks were used are darkened.

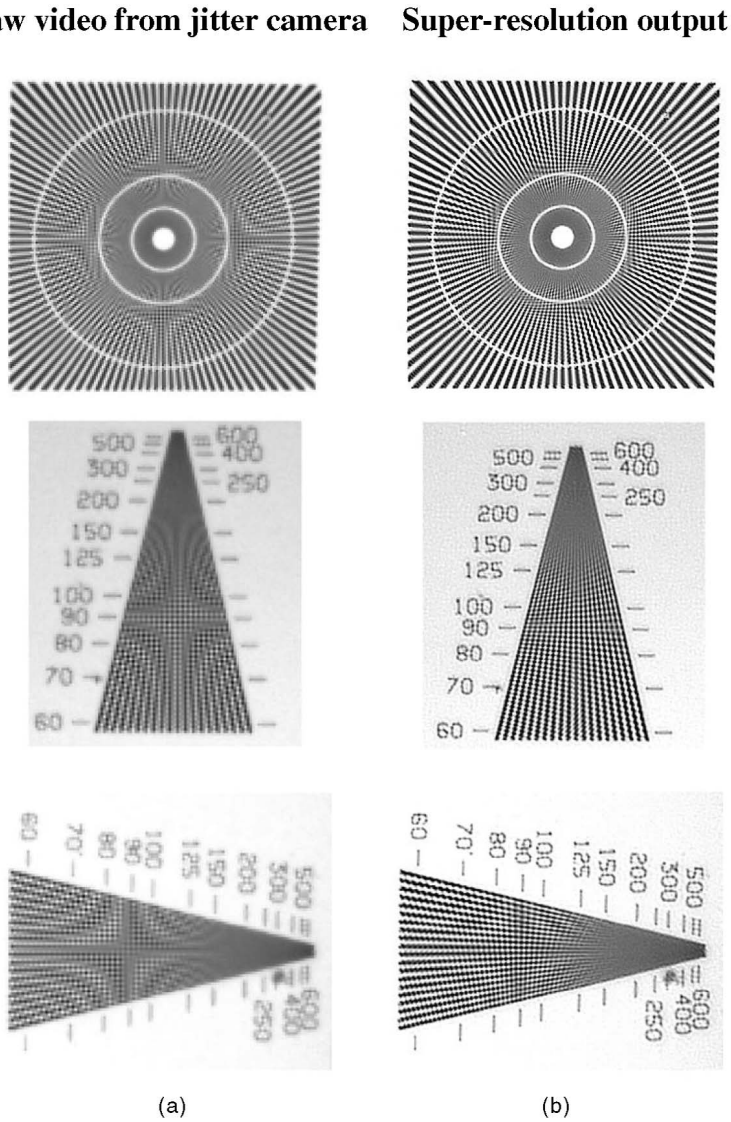

Fig. 8. Resolution test using a standard Kodak test target. The left column shows angular, vertical, and horizontal resolution test targets that were captured by the jitter camera (one of four input images). The right column shows the super-resolution results. Note the strong aliasing in the input images and the clear separation between lines in the super-resolution result images.

6 Experiments

We tested resolution enhancement with our jitter camera for both static and dynamic scenes. The input images were obtained from the raw Bayer-pattern samples using the demosaicing algorithm provided by the camera manufacturer [17]. The images were then transformed to the *CIE-Lab* color space and the super-resolution algorithm [9] was applied to the L-channel only. The low resolution (a,b)-chroma channels were linearly interpolated and combined with the high resolution L-channel.

6.1 Resolution Tests

The resolution enhancement was evaluated quantitatively using a standard Kodak test target. The input to the super-resolution algorithm was four frames from a jitter-camera video sequence. Fig. 8 shows angular, vertical, and horizontal test patterns. The aliasing effects are clearly seen in the input images, where the line separation is not clear even at the lower resolution of 60 lines per inch. In the computed super-resolution images, the spatial resolution is clearly enhanced in all angles and it is possible to resolve separate lines well above 100 lines per inch.

6.2 Color Test

The standard Kodak test target is black and white. In order to check the color performance, we used a test target consisting of a color image and lines of text of different font sizes. Fig. 9a and 9b show one out of four different input images taken by the jitter camera and a magnified part of the image.

The camera we used has a single detector with each pixel in the detector measuring a single color channel, either red, green, or blue. In order to obtain the complementary channels in each pixel, an interpolation algorithm is used. There is a wide literature on such interpolation algorithms, typically referred to as "demosaicing." The interested reader may refer to [15] to learn about different demosaicing algorithms and about color artifacts in demosaiced images. In our experiments, for the input images, we utilized the best

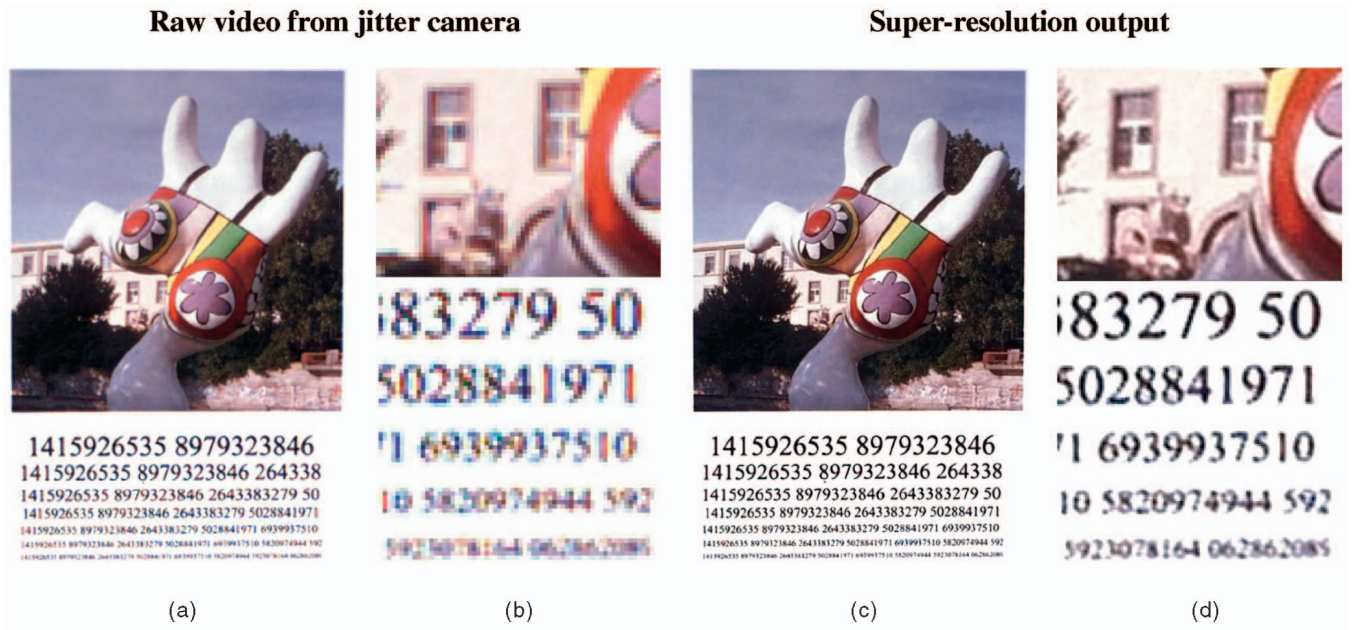

Fig. 9. Resolution test of combined color and text image. (a) and (b) show one out of four different input images taken by the jitter camera together with a magnified part of the image. Note that the last line of the text, which is only six pixels high, is completely unreadable; also, note the demosaicing artifacts in both the text and the image. (c) and (d) show the super-resolution result and a magnified part of it. The resolution is clearly enhanced and it is now possible to read all the text lines that were unreadable in the input images. Moreover, we can see that the demosaicing artifacts have almost vanished while the colors were preserved.

color demosaicing algorithm the Dragonfly camera had to offer (proprietary "rigorous" algorithm). We can see in Fig. 9 that the input image contains color artifact along edges. Fig. 9c and 9d show the super-resolution result image and a magnified part, respectively. The resolution is clearly enhanced and it is now possible to read all the text lines that were unreadable in the input images. Moreover, we can see that the demosaicing artifacts have almost completely disappeared, while the colors were preserved. This is due to the fact that the super-resolution was applied only to the intensity channel while the chromaticity channels were smoothly interpolated.

6.3 Dynamic Video Tests

Several experiments were conducted to test the system's performance in the presence of moving objects and occlusions. Fig. 10 shows magnified parts of a scene with mostly static objects. These objects, such as the crossing pedestrians sign in the first row and the no-parking sign in the second row, were significantly enhanced, revealing new details. Fig. 11 shows magnified parts of scenes with static and dynamic objects. One can see that the adaptive super-resolution algorithm has increased the resolution of stationary objects while preserving or increasing the resolution of moving objects.

7 Conclusions

Super-resolution algorithms can improve spatial resolution. However, their performance depends on various factors in the camera imaging process. We showed that motion blur causes significant degradation of super-resolution results, even when the motion blur function is known. The proposed solution is the jitter camera, a video camera capable of sampling the space-time volume without introducing motion blur. Applying a super-resolution algorithm to jitter camera video sequences significantly enhances their resolution.

Image detectors are becoming smaller and lighter and thus require very little force to jitter. With recent advances, it may be possible to manufacture jitter cameras with the jitter mechanism embedded inside the detector chip. Jittering can then be added to regular video cameras as an option that enables a significant increase of spatial resolution while keeping other factors such as frame-rate unchanged.

Motion blur is only one factor in the imaging process. By considering other factors, novel methods for sampling the space-time volume can be developed, resulting in further improvements in video resolution. In this paper, for example, we limited the detector to a regular sampling lattice and to regular temporal sampling. One interesting direction can be the use of different lattices and different temporal samplings. We therefore consider the jitter camera to be a first step towards a family of novel camera designs that better sample the space-time volume to improve not only spatial resolution, but also temporal resolution and spectral resolution.

Appendix A

The Influence of Motion Blur on the Volume of Solutions

The imaging process of the multiple input images is modeled by a matrix A:

$$\vec{y} = A\vec{x} + \vec{z}. \tag{3}$$

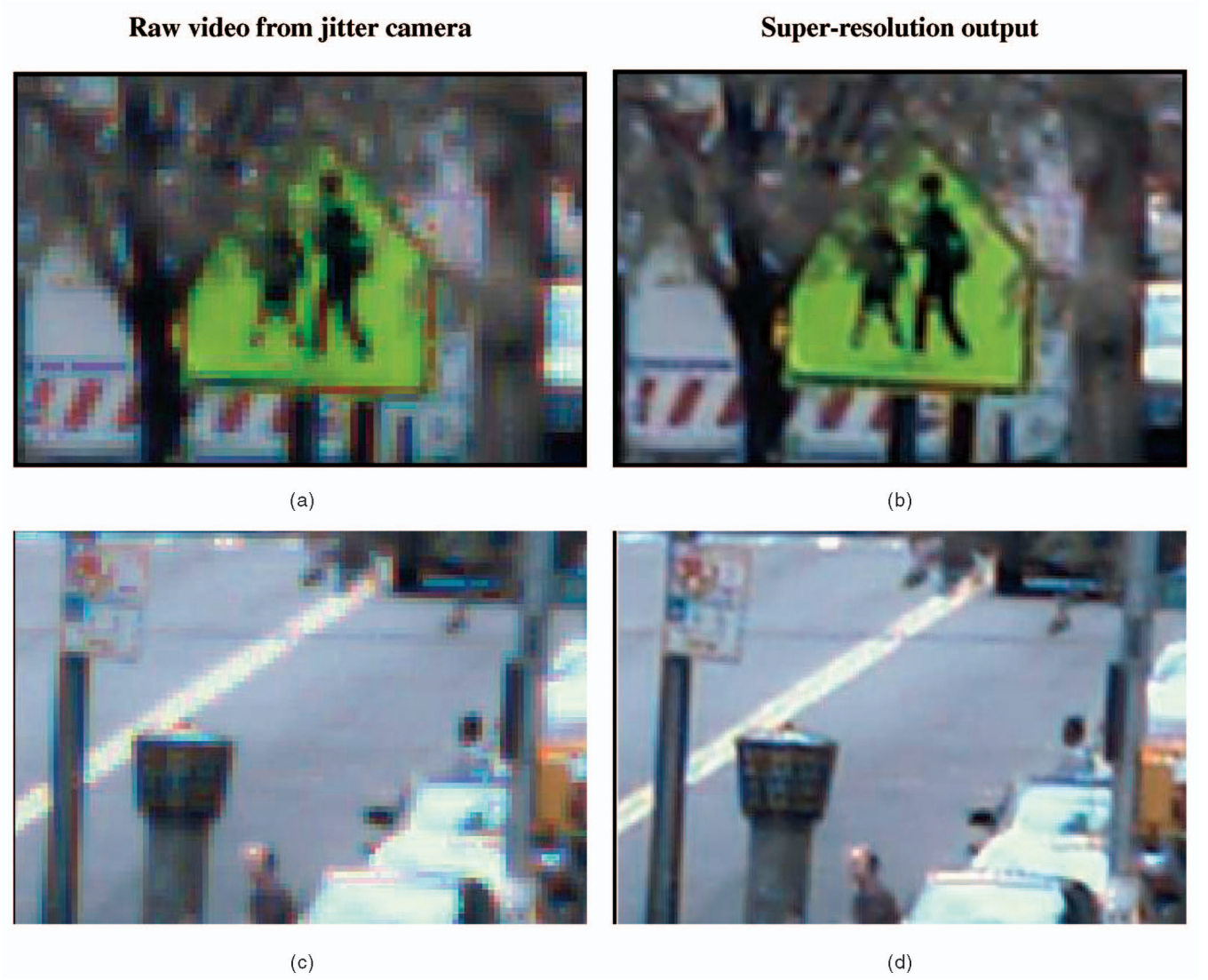

Fig. 10. Jitter camera super-resolution for scenes of mostly stationary objects. The left column shows the raw video input from the jitter camera and the right column shows the super-resolution results. (a) and (b) show a static scene. Note the significant resolution enhancement of the pedestrian on the sign and the fine texture of the tree branches. (c) and (d) show a scene with few moving objects. Note the enhancement of the text on the no-parking sign and some enhancement of the walking person.

$\vec{x}$ is a vectorization of the unknown discrete high resolution image, $\vec{y}$ is a vectorization of all the input images, and $\vec{z}$ is the uncertainty in measurements. A minimal number of input images is assumed such that the dimensionality of $\vec{y}$ is equal to the dimensionality of $\vec{x}$ and matrix A is square.

The volume of solutions corresponding to a square imaging matrix A is computed from the absolute value of its determinant (see (2)):

$$vol(A) = \left|\frac{1}{|A|}\right|.$$

In the following, we derive a simplified expression for the determinant of the imaging matrix A and present the volume of solutions as a function of the camera displacements, motion blurs, optical transfer function, and the integration function of the detector.

Let $\mathbf{f}$ be the $n \times n$ high resolution image (corresponding to $\vec{x}$ in (3)) and let $\{\mathbf{g}_j\}_{j=0}^{m^2-1}$ be the $\frac{n}{m} \times \frac{n}{m}$ input images (corresponding to $\vec{y}$). The imaging process is defined in the image domain by:

$$\mathbf{g}_j = (\mathbf{f} * \mathbf{h}_j) \downarrow_m + \mathbf{z}_j, \tag{4}$$

where $*$ denotes convolution, $\mathbf{h}_j$ encapsulates the sensor displacement and motion blur of the jth image and the optical blur and detector integration of the camera, $\mathbf{z}_j$ represents the quantization error, and $\downarrow_m$ denotes subsampling by a factor of m. In the frequency domain, let Z_j, G_j, H_j, F denote the Fourier transforms of $\mathbf{z}_j, \mathbf{g}_j, \mathbf{h}_j, \mathbf{f}$, respectively. The frequencies of the high resolution image are folded as a result of the subsampling

$$G_j(u,v) = Z_j(u,v) + \sum_{\bar{u} \in U, \bar{v} \in V} Rect_{[-\frac{n}{2},\frac{n}{2}]}(\bar{u},\bar{v}) H_j(\bar{u},\bar{v}) F(\bar{u},\bar{v}), \tag{5}$$

where $U = \{u + \frac{kn}{m}\}_{k=-\infty}^{\infty}$, $V = \{v + \frac{kn}{m}\}_{k=-\infty}^{\infty}$, and $Rect_{[-\frac{n}{2},\frac{n}{2}]}(\bar{u},\bar{v})$ equals 1 when $[-\frac{n}{2} \leq \bar{u}, \bar{v} < \frac{n}{2}]$ and 0 otherwise. This leads to the following result:

Fig. 11. Jitter camera super-resolution for scenes with dynamic and stationary objects. The left column shows the raw video input from the jitter camera and the right column shows the super-resolution results. (a) and (b) show a scene with a large stationary object (boat) and a large moving object (woman's head). As expected, the resolution enhancement is better for the boat. (c) and (d) show a particularly dynamic scene with many moving objects. Note the enhancement of the face of the walking women (center) and the kid on the scooter (left).

Proposition 1. *Let A be the matrix of (3) corresponding to the imaging process above (4) for $m = 2$ (four input images).*

Define $\hat{u} = u - sign(u)\frac{n}{2}$, $\hat{v} = v - sign(v)\frac{n}{2}$, then the determinant of A is given by: $|A| = \prod_{-\frac{n}{4} \le u,v < \frac{n}{4}} |\bar{A}_{u,v}|$, where:

$$\bar{A}_{u,v} = \begin{bmatrix} H_0(u,v) & H_0(\hat{u},v) & H_0(u,\hat{v}) & H_0(\hat{u},\hat{v}) \\ H_1(u,v) & H_1(\hat{u},v) & H_1(u,\hat{v}) & H_1(\hat{u},\hat{v}) \\ H_2(u,v) & H_2(\hat{u},v) & H_2(u,\hat{v}) & H_2(\hat{u},\hat{v}) \\ H_3(u,v) & H_3(\hat{u},v) & H_3(u,\hat{v}) & H_3(\hat{u},\hat{v}) \end{bmatrix}.$$

Proof. Let $\bar{A}$ be a matrix describing the imaging process in the frequency domain

$$\begin{bmatrix} G_0(-\frac{n}{4},\frac{n}{4}) \\ \vdots \\ G_3(\frac{n}{4}-1,\frac{n}{4}-1) \end{bmatrix} = \bar{A} \begin{bmatrix} F(-\frac{n}{2},\frac{n}{2}) \\ \vdots \\ F(\frac{n}{2}-1,\frac{n}{2}-1) \end{bmatrix} + \begin{bmatrix} Z_0(-\frac{n}{4},\frac{n}{4}) \\ \vdots \\ Z_3(\frac{n}{4}-1,\frac{n}{4}-1) \end{bmatrix}.$$

From (5), in the case of $m = 2$, the frequencies $G_0(u,v), \ldots, G_3(u,v)$ are given by linear combinations of only four frequencies $F(\bar{u},\bar{v})$, $\bar{u} \in \{u,\hat{u}\}, \bar{v} \in \{v,\hat{v}\}$ up to the uncertainty Z:

$$\begin{bmatrix} G_0(u,v) \\ G_1(u,v) \\ G_2(u,v) \\ G_3(u,v) \end{bmatrix} = \begin{bmatrix} H_0(u,v) & H_0(\hat{u},v) & H_0(u,\hat{v}) & H_0(\hat{u},\hat{v}) \\ H_1(u,v) & H_1(\hat{u},v) & H_1(u,\hat{v}) & H_1(\hat{u},\hat{v}) \\ H_2(u,v) & H_2(\hat{u},v) & H_2(u,\hat{v}) & H_2(\hat{u},\hat{v}) \\ H_3(u,v) & H_3(\hat{u},v) & H_3(u,\hat{v}) & H_3(\hat{u},\hat{v}) \end{bmatrix} \begin{bmatrix} F(u,v) \\ F(\hat{u},v) \\ F(u,\hat{v}) \\ F(\hat{u},\hat{v}) \end{bmatrix} + \begin{bmatrix} Z_0(u,v) \\ Z_1(u,v) \\ Z_2(u,v) \\ Z_3(u,v) \end{bmatrix}.$$

Hence, the matrix $\bar{A}$ is block diagonal up to a permutation, with blocks corresponding to $\bar{A}_{u,v}$, $-\frac{n}{4} \le u,v < \frac{n}{4}$. It follows that $|\bar{A}| = \prod_{u,v} |\bar{A}_{u,v}|$. Since the Fourier transform preserves the determinant magnitude, $|A| = |\bar{A}| = \prod_{u,v} |\bar{A}_{u,v}|$. □

To analyze the influence of motion blur, we factor the terms in $|\bar{A}_{u,v}|$:

$$H_j(a,b) = O(a,b)C(a,b)M_j(a,b)D_j(a,b)$$

with $a \in \{u,\hat{u}\}, b \in \{v,\hat{v}\}$. $O(a,b)$ is the Fourier transform of the optical transfer function, $C(a,b)$ is the transform of the

detector's integration function, $M_j(a,b)$ is the transform of the motion blur point spread function, and $D_j(a,b)$ is the transform of the sensor displacements $\delta(x-x_j, y-y_j)$.

Let $\{\vec{l}_j\}_{j=0}^3$ be the vectors describing the motion blur path so that, during integration, the projected image g_j moves at a constant velocity from $-\frac{\vec{l}_j}{2}$ to $\frac{\vec{l}_j}{2}$ (measured in the high resolution coordinate system). The transform of the motion blur is given by:

$$M_j(a,b) = sinc(m\vec{l}_j^{\,T}\vec{w}) = \frac{sin(m\vec{l}_j^{\,T}\vec{w})}{\pi m\vec{l}_j^{\,T}\vec{w}}$$

with $\vec{w} = [a,b]^T$.

Let $\{x_j, y_j\}_{j=0}^3$ be the displacements of the input images $\{\mathbf{g}_j\}_{j=0}^3$, respectively, with $x_0 = 0, y_0 = 0$. The Fourier Transform $D_j(a,b)$ of the displacements $\delta(x-x_j, y-y_j)$ is given by:

$$D_j(a,b) = e^{-\frac{2\pi i(ax_j+by_j)}{n}} = e^{-\frac{2\pi i(ux_j+vy_j)}{n}} e^{-\frac{2\pi i((a-u)x_j+(b-v)y_j)}{n}}.$$

$D_j(a,b)$ is expressed as a product of two terms. The first term is common to all pairs (a,b) and, hence, can be factored out of the determinant. Similarly, the terms $O(a,b), C(a,b)$ are common to all images and can be factored out of the determinant. It follows that:

$$|\bar{A}_{uv}| = |\bar{B}_{uv}| \prod_{0\le j\le 3} e^{-\frac{2\pi i(ux_j+vy_j)}{n}} \prod_{a\in\{u,\hat{u}\}b\in\{v,\hat{v}\}} O(a,b)C(a,b), \quad (6)$$

where

$$\bar{B}_{uv} = \begin{bmatrix} M_0(u,v) & M_0(\hat{u},v) & M_0(u,\hat{v}) & M_0(\hat{u},\hat{v}) \\ M_1(u,v) & M_1(\hat{u},v)e^{-i\pi s(u)x_1} & M_1(u,\hat{v})e^{-i\pi s(v)y_1} & M_1(\hat{u},\hat{v})e^{-i\pi\left(s(u)x_1+s(v)y_1\right)} \\ M_2(u,v) & M_2(\hat{u},v)e^{-i\pi s(u)x_2} & M_2(u,\hat{v})e^{-i\pi s(v)y_2} & M_2(\hat{u},\hat{v})e^{-i\pi\left(s(u)x_2+s(v)y_2\right)} \\ M_3(u,v) & M_3(\hat{u},v)e^{-i\pi s(u)x_3} & M_3(u,\hat{v})e^{-i\pi s(v)y_3} & M_3(\hat{u},\hat{v})e^{-i\pi\left(s(u)x_3+s(v)y_3\right)} \end{bmatrix}, \quad (7)$$

and $s(u)$ is an abbreviation for the sign function $s(u) = sign(u)$.

The influence of motion blur on the volume of solutions is therefore expressed in the matrices $\bar{B}_{uv}$. Since the volume of solutions $vol(A) = \left|\frac{1}{|A|}\right|$ depends on the image size, we define

$$s(A) = \left|\left(\prod_{u,v} |\bar{B}_{uv}|\right)^{-\frac{1}{n^2}}\right|, \quad (8)$$

so that, according to Proposition 1 and (6),

$$vol(A) = \left|\prod_{-\frac{n}{4}\le u,v<\frac{n}{4}} |\bar{A}_{u,v}|\right|^{-1} \propto \left|\left(\prod_{u,v} |\bar{B}_{uv}|\right)^{-1}\right| = s(A)^{n^2}. \quad (9)$$

To conclude, $s(A)$ is a relative measure for the volume of solutions that is independent of the optical blur and detector's integration function and is normalized to account for the image size. The generalization of the above results for an arbitrary integer magnification factor m is straightforward and is omitted in order to simplify notations.

The lower bound for $s(A)$ was derived using the following inequality for a $k\times k$ matrix P [14]:

$$|P| \le \left(\frac{\|P\|_F^2}{k}\right)^{\frac{k}{2}}, \quad (10)$$

with $\|\cdot\|_F$ as the Frobenius norm. In order to bound $s(A)$, we define a block-diagonal matrix $\bar{B}$ of size $n^2\times n^2$ with $\bar{B}(mu-m+j, mv-m+k) = \bar{B}_{uv}(j,k)$. Using (10) on (8):

$$s(A) = \left||\bar{B}|^{-\frac{1}{n^2}}\right| \le \left(\frac{\|\bar{B}\|_F^2}{n^2}\right)^{-\frac{1}{2}}. \quad (11)$$

The matrix $\bar{B}$ has m^2n^2 nonzero values, each of the form $e^{ix}sinc(m\vec{s}_j^{\,T}\vec{w}_k)$ for some x. The Frobenius norm of $\bar{B}$ is, hence,

$$\|\bar{B}\|_F^2 = \sum_{j=0}^{m^2-1} \sum_{\vec{w}\in C\times C} sinc^2(m\vec{s}_j^{\,T}\vec{w}), \quad (12)$$

with $C = \{-\frac{1}{2} + \frac{k}{n}\}_{k=0}^{n-1}$. As n goes to infinity, the sums are replaced by integrals:

$$\lim_{n\to\infty} \frac{1}{n^2}|\bar{B}|_F^2 = \sum_{j=0}^{m^2-1} \int_{\vec{w}\in\left[-\frac{1}{2},\frac{1}{2}\right]\times\left[-\frac{1}{2},\frac{1}{2}\right]} sinc^2(m\vec{s}_j^{\,T}\vec{w}). \quad (13)$$

The integrals were solved using a symbolic math software. For a given line magnitude $\|\vec{s}_j\|$, the maximal values of the integrals are obtained when $\vec{s}_j$ is oriented by 45 degrees. The lower bound, appearing in Fig. 2, is therefore the value of (11) using (13) for a 45 degrees oriented blur $\vec{s} = \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T$ and for a magnification factor $m = 2$:

$$s(A) \le \left(4\int_{\vec{w}\in\left[-\frac{1}{2},\frac{1}{2}\right]\times\left[-\frac{1}{2},\frac{1}{2}\right]} sinc^2(m\vec{s}_j^{\,T}\vec{w})\right)^{-\frac{1}{2}}.$$

APPENDIX B
OPTIMAL SPATIAL DISPLACEMENTS

We show that, when there is no motion blur (or the motion blur is common to all images), the four grid displacements $\{(0,0)(1,0)(0,1)(1,1)\}$ (in the high resolution coordinate system) are optimal for super-resolution in terms of the volume of solutions. A similar result was shown in [10] measuring the super-resolution quality using perturbation theory.

Proposition 2. *Consider the imaging process as defined in (4). Assume the filters $\{\mathbf{h}_k\}_{k=0}^3$ have the same spatial blur, yet different displacements $\{x_k, y_k\}_{k=0}^3$, i.e., $\mathbf{h}_k = \mathbf{h} * \delta(x-x_k, y-y_k)$ for some filter $\mathbf{h}$. Then, $vol(A)$ in (2) is minimal for displacements $\{(0,0)(1,0)(0,1)(1,1)\}$ in the coordinate system of the high resolution image.*

Proof. Let H be the Fourier transform of $\mathbf{h}$. From Proposition 1 and (6) and (7), it is sufficient to prove the maximality of $|\bar{B}_{u,v}|$ for all frequencies (u,v). In this case, since the images share the same spatial blur, the motion blur can be folded into H and (7) simplifies to:

$$\bar{B}_{u,v} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{-i\pi s(u)x_1} & e^{-i\pi s(v)y_1} & e^{-i\pi(s(u)x_1+s(v)y_1)} \\ 1 & e^{-i\pi s(u)x_2} & e^{-i\pi s(v)y_2} & e^{-i\pi(s(u)x_2+s(v)y_2)} \\ 1 & e^{-i\pi s(u)x_3} & e^{-i\pi s(v)y_3} & e^{-i\pi(s(u)x_3+s(v)y_3)} \end{bmatrix}.$$

The rows of $\bar{B}_{u,v}$ have the same norm for all assignments of $\{(x_k, y_k)\}$. Hence, the determinant is maximized

when the rows are orthogonal. The rows are orthogonal if and only if

$$\forall k,l,(1+e^{\pi is(u)(x_l-x_k)}+e^{\pi is(v)(y_l-y_k)} + e^{\pi is(u)(x_l-x_k)}e^{\pi is(v)(y_l-y_k)})=0$$
$$\Rightarrow \forall k,l,(1+e^{\pi is(u)(x_l-x_k)})(1+e^{\pi is(v)(y_l-y_k)})=0,$$

which is satisfied when, for every k,l, either $|x_l-x_k|=1$ or $|y_l-y_k|=1$. This condition is satisfied by the above displacements $\{(0,0)(1,0)(0,1)(1,1)\}$. □

Note that there are other displacements that maximize $|A|$, for example, $(0,0)(1,0)(x,1)(x+1,1)$ for any $x \in \mathcal{R}$.

ACKNOWLEDGMENTS

This research was conducted at the Columbia Vision and Graphics Center in the Computer Science Department at Columbia University. It was funded in parts by an ONR Contract (N00014-03-1-0023) and a US National Science Foundation ITR Grant (IIS-00-85864).

REFERENCES

[1] S. Baker and T. Kanade, "Limits on Super-Resolution and How to Break Them," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 9, pp. 1167-1183, Sept. 2002.
[2] B. Bascle, A. Blake, and A. Zisserman, "Motion Deblurring and Super-Resolution from an Image Sequence," *Proc. European Conf. Computer Vision,* vol. 2, pp. 573-582, 1996.
[3] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani, "Hierarchical Model-Based Motion Estimation," *Proc. European Conf. Computer Vision,* pp. 237-252, 1992.
[4] D. Capel and A. Zisserman, "Super-Resolution Enhancement of Text Image Sequences," *Proc. Int'l Conf. Pattern Recognition,* vol. I, pp. 600-605, Sept. 2000.
[5] M.C. Chiang and T.E. Boult, "Efficient Super-Resolution via Image Warping," *Image and Vision Computing,* vol. 18, no. 10 pp. 761-771, July 2000.
[6] Pixera Corporation, "Diractor," http://www.pixera.com.
[7] M. Elad and A. Feuer, "Restoration of a Single Superresolution Image from Several Blurred, Noisy, and Undersampled Measured Images," *IEEE Trans. Image Processing,* vol. 6, no. 12, pp. 1646-1658, Dec. 1997.
[8] Physik Instrumente, "M-111 Micro Translation Stage," http://www.physikinstrumente.de, 2005.
[9] M. Irani and S. Peleg, "Improving Resolution by Image Registration," *Graphical Models and Image Processing,* vol. 53, pp. 231-239, 1991.
[10] Z. Lin and H.Y. Shum, "Fundamental Limits of Reconstruction-Based Superresolution Algorithms under Local Translation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 1 pp. 83-97, Jan. 2004.
[11] P. Meer, D. Mintz, D.Y. Kim, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision,* vol. 6, no. 1, pp. 59-70, 1991.
[12] Minolta "Dimage-a1," http://www.dpreview.com/reviews/minoltadimagea1/, 2005.
[13] A.J. Patti, M.I. Sezan, and A.M. Tekalp, "Superresolution Video Reconstruction with Arbitrary Sampling Lattices and Nonzero Aperture Time," *IEEE Trans. Image Processing,* vol. 6, no. 8, pp. 1064-1076, Aug. 1997.
[14] P. Pudlák, "A Note on the Use of Determinant for Proving Lower Bounds on the Size of Linear Circuits," *Information Processing Letters,* vol. 74, nos. 5-6, pp. 197-201, 2000.
[15] R. Ramanath, W. Snyder, G. Bilbro, and W. Sander, "Demosaicking Methods for Bayer Color Arrays," *J. Electronic Imaging,* vol. 11, no. 3, July 2002.
[16] A. Rav-Acha and S. Peleg, "Restoration of Multiple Images with Motion Blur in Different Directions," *Proc. IEEE Workshop Applications of Computer Vision,* pp. 22-28, 2000.
[17] Point Grey Research, "Dragonfly Camera," http://www.ptgrey.com, 2005.
[18] R.R. Schultz and R.L. Stevenson, "Extraction of High-Resolution Frames from Video Sequences," *IEEE Trans. Image Processing,* vol. 5, no. 6, pp. 996-1011, June 1996.
[19] E. Shechtman, Y. Caspi, and M. Irani, "Increasing Space-Time Resolution in Video," *Proc. European Conf. Computer Vision,* vol. I, p. 753, 2002.
[20] H. Shekarforoush and R. Chellappa, "Data-Driven Multichannel Superresolution with Application to Video Sequences," *J. Optical Soc. Am.,* vol. 16, no. 3, pp. 481-492, Mar. 1999.
[21] E.P. Simoncelli, "Modeling the Joint Statistics of Images in the Wavelet Domain," *SPIE,* vol. 3813, pp. 188-195, July 1999.

Moshe Ben-Ezra received the BSc, MSc, and PhD degrees in computer science from the Hebrew University of Jerusalem in 1994, 1996, and 2000, respectively. He was a research scientist at Columbia University from 2002 until 2004 and is now with Siemens Corporate Research at Princeton. His research interests are in computer vision with an emphasis on real-time vision and optics.

Assaf Zomet received the BA, MSc, and PhD degrees from the Hebrew University of Jerusalem, Israel, in 1997, 1999, and 2003, respectively. He is currently a research scientist in the computer science department at Columbia University. His research interests include mosaicing, super-resolution, low-level vision, and novel cameras.

Shree K. Nayar received the PhD degree in electrical and computer engineering from the Robotics Institute at Carnegie Mellon University in 1990. He is currently the TC Chang Professor of Computer Science at Columbia University and heads the Columbia Automated Vision Environment (CAVE), which is dedicated to the development of advanced computer vision systems. His research is focused on three areas: the creation of cameras that produce new forms of visual information, the modeling of the interaction of light with materials, and the design of algorithms that recognize objects from images. His work is motivated by applications in the fields of computer graphics, human-machine interfaces, and robotics. Dr. Nayar has authored and coauthored papers that have received the Best Paper Award at the 2004 CVPR conference, the Best Paper Honorable Mention Award at the 2000 IEEE CVPR conference, the David Marr Prize at the 1995 ICCV, the Siemens Outstanding Paper Award at the 1994 IEEE CVPR Conference, the 1994 Annual Pattern Recognition Award from the Pattern Recognition Society, the Best Industry Related Paper Award at the 1994 ICPR, and the David Marr Prize at the 1990 ICCV. He holds several US and international patents for inventions related to computer vision and robotics. Dr. Nayar was the recipient of the David and Lucile Packard Fellowship for Science and Engineering in 1992, the National Young Investigator Award from the US National Science Foundation in 1993, and the Excellence in Engineering Teaching Award from the Keck Foundation in 1995.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.