Increasing Space-Time Resolution in Video

Eli Shechtman, Yaron Caspi, and Michal Irani

Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
76100 Rehovot, Israel
{elishe,caspi,irani}@wisdom.weizmann.ac.il

Abstract. We propose a method for constructing a video sequence of high space-time resolution by combining information from multiple low-resolution video sequences of the same dynamic scene. Super-resolution is performed simultaneously in time and in space. By "temporal super-resolution" we mean recovering rapid dynamic events that occur faster than regular frame-rate. Such dynamic events are not visible (or else observed incorrectly) in any of the input sequences, even if these are played in "slow-motion".
The spatial and temporal dimensions are very different in nature, yet are inter-related. This leads to interesting visual tradeoffs in time and space, and to new video applications. These include: (i) treatment of *spatial* artifacts (e.g., motion-blur) by increasing the *temporal* resolution, and (ii) combination of input sequences of different space-time resolutions (e.g., NTSC, PAL, and even high quality still images) to generate a high quality video sequence.

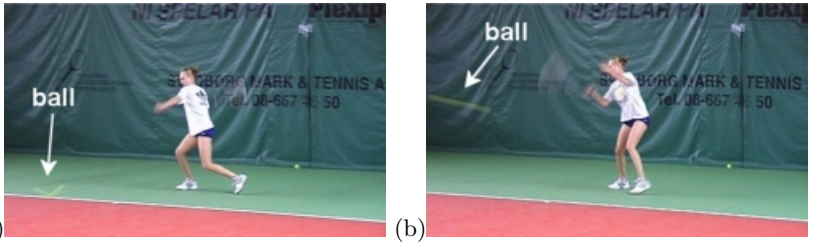
Keywords. Super-resolution, space-time analysis.

1 Introduction

A video camera has limited spatial and temporal resolution. The spatial resolution is determined by the spatial density of the detectors in the camera and by their induced blur. These factors limit the minimal size of spatial features or objects that can be visually detected in an image. The temporal resolution is determined by the frame-rate and by the exposure-time of the camera. These limit the maximal speed of dynamic events that can be observed in a video sequence.

Methods have been proposed for increasing the spatial resolution of images by combining information from multiple low-resolution images obtained at sub-pixel displacements (e.g. [1,2,5,6,9,10,11,12,14]. See [3] for a comprehensive review). These, however, usually assume static scenes and do not address the limited temporal resolution observed in dynamic scenes. In this paper we extend the notion of super-resolution to the *space-time* domain. We propose a unified framework for increasing the resolution both in time and in space by combining information from multiple *video sequences* of dynamic scenes obtained at (sub-pixel) spatial and (sub-frame) temporal misalignments. As will be shown, this enables new visual capabilities of dynamic events, gives rise to visual tradeoffs between time

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2350, pp. 753–768, 2002.



(a) (b)

Fig. 1. Motion blur. *Distorted shape due to motion blur of very fast moving objects (the tennis ball and the racket) in a real tennis video. The perceived distortion of the ball is marked by a white arrow. Note, the "V"-like shape of the ball in (a), and the elongated shape of the ball in (b). The racket has almost "disappeared".*

and space, and leads to new video applications. These are substantial in the presence of very fast dynamic events.

Rapid dynamic events that occur faster than the frame-rate of video cameras are not visible (or else captured incorrectly) in the recorded video sequences. This problem is often evident in sports videos (e.g., tennis, baseball, hockey), where it is impossible to see the full motion or the behavior of the fast moving ball/puck. There are two typical visual effects in video sequences which are caused by very fast motion. One effect (motion blur) is caused by the exposure-time of the camera, and the other effect (motion aliasing) is due to the temporal sub-sampling introduced by the frame-rate of the camera:

(i) *Motion Blur:* The camera integrates the light coming from the scene during the exposure time in order to generate each frame. As a result, fast moving objects produce a noted blur along their trajectory, often resulting in distorted or unrecognizable object shapes. The faster the object moves, the stronger this effect is, especially if the trajectory of the moving object is not linear. This effect is notable in the distorted shapes of the tennis ball shown in Fig. 1. Note also that the tennis racket also "disappears" in Fig. 1.b. Methods for treating motion blur in the context of image-based super-resolution were proposed in [2, 12]. These methods however, require prior segmentation of moving objects and the estimation of their motions. Such motion analysis may be impossible in the presence of severe shape distortions of the type shown in Fig. 1. We will show that by increasing the *temporal resolution* using information from multiple video sequences, *spatial artifacts* such as motion blur can be handled without the need to separate static and dynamic scene components or estimate their motions.

(ii) *Motion-Based (Temporal) Aliasing:* A more severe problem in video sequences of fast dynamic events is false visual illusions caused by aliasing in time. Motion aliasing occurs when the trajectory generated by a fast moving object is characterized by frequencies which are higher than the frame-rate of the camera (i.e., the temporal sampling rate). When that happens, the high temporal frequencies are "folded" into the low temporal frequencies. The observable result is a distorted or even false trajectory of the moving object. This effect is illus-

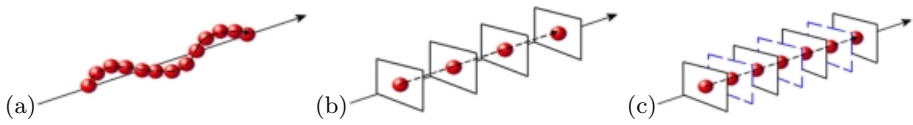


Fig. 2. Motion aliasing. *(a) shows a ball moving in a sinusoidal trajectory. (b) displays an image sequence of the ball captured at low frame-rate. The perceived motion is along a straight line. This false perception is referred to in the paper as "motion aliasing". (c) Illustrates that even using an ideal temporal interpolation will not produces the correct motion. The filled-in frames are indicated by the blue dashed line.*

trated in Fig. 2, where a ball moves fast in sinusoidal trajectory of high frequency (Fig. 2.a). Because the frame-rate is much lower (below Nyquist frequency of the trajectory), the *observed* trajectory of the ball is a straight line (Fig. 2.b). Playing that video sequence in "slow-motion" will not correct this false visual effect (Fig. 2.c). Another example of motion-based aliasing is the well-known visual illusion called the "wagon wheel effect": When a wheel is spinning very fast, beyond a certain speed it will appear to be rotating in the "wrong" direction.

Neither the motion-based aliasing nor the motion blur can be treated by playing such video sequences in "slow-motion", even when sophisticated temporal interpolations are used to increase the frame-rate (as in format conversion or "re-timing" methods [8,13]). This is because the information contained in a single video sequence is insufficient to recover the missing information of very fast dynamic events (due to excessive blur and subsampling). Multiple video sequences, on the other hand, provide additional samples of the dynamic space-time scene. While none of the individual sequences provides enough visual information, combining the information from all the sequences allows to generate a video sequence of high space-time resolution (Sec. 2), which displays the correct dynamic events. Thus, for example, a reconstructed high-resolution sequence will display the correct motion of the wagon wheel despite it appearing incorrectly in *all* of the input sequences (Sec. 4).

The spatial and temporal dimensions are very different in nature, yet are inter-related. This introduces visual tradeoffs between space and times, which are unique to spatio-temporal super-resolution, and are not applicable in traditional spatial (i.e., image-based) super-resolution. For example, output sequences of different space-time resolutions can be generated for the same input sequences. A large increase in the temporal resolution usually comes at the expense of a large increase in the spatial resolution, and vice versa.

Furthermore, input sequences of different space-time resolutions can be meaningfully combined in our framework. In traditional image-based super-resolution there is no incentive to combine input images of different spatial resolutions, since a high-resolution image will subsume the information contained in a low-resolution image. This, however, is not the case here. Different types of cameras of different space-time resolutions may provide *complementary* information. Thus, for example, we can combine information obtained by high-quality still cameras

(which have very high spatial-resolution, but extremely low "temporal resolution"), with information obtained by standard video cameras (which have low spatial-resolution but higher temporal resolution), to obtain an improved video sequence of high spatial and high temporal resolution. These issues and other space-time visual tradeoffs are discussed in Sec. 4.

2 Space-Time Super-Resolution

Let S be a dynamic space-time scene. Let $\{S_i^l\}_{i=1}^n$ be n video sequences of that dynamic scene recorded by n different video cameras. The recorded sequences have limited spatial and temporal resolution. Their limited resolutions are due to the space-time imaging process, which can be thought of as a process of blurring followed by sampling in time and in space.

The blurring effect results of the fact that the color at each pixel in each frame (referred to as a "space-time point" and marked by the small boxes in Fig. 3.a) is an integral (a weighted average) of the colors in a space-time *region* in the dynamic scene S (marked by the large pink (bright) and blue (dark) boxes in Fig. 3.a). The temporal extent of this region is determined by the exposure-time of the video camera, and the spatial extent of this region is determined by the spatial point-spread-function (PSF) of the camera (determined by the properties of the lens and the detectors [4]).

The sampling process also has a spatial and a temporal components. The spatial sampling results from the fact that the camera has a discrete and finite number of detectors (the output of each is a single pixel value), and the temporal sampling results from the fact that the camera has a finite frame-rate resulting in discrete frames (typically 25 $frames/sec$ in PAL cameras and 30 $frames/sec$ in NTSC cameras).

The above space-time imaging process inhibits high spatial and high temporal frequencies of the dynamic scene, resulting in video sequences of low space-time resolutions. Our objective is to use the information from all these sequences to construct a new sequence S^h of high space-time resolution. Such a sequence will have smaller blurring effects and finer sampling in space and in time, and will thus capture higher space-time frequencies of the dynamic scene S. In particular, it will capture fine spatial features in the scene and rapid dynamic events which cannot be captured by the low-resolution sequences.

The recoverable high-resolution information in S^h is limited by its spatial and temporal sampling rate (or discretization) of the space-time volume. These rates can be different in space and in time. Thus, for example, we can recover a sequence S^h of very high spatial resolution but low temporal resolution (e.g., see Fig. 3.b), a sequence of very high temporal resolution but low spatial resolution (e.g., see Fig. 3.c), or a bit of both. These tradeoffs in space-time resolutions and their visual effects will be discussed in more detail later in Sec. 4.2.

We next model the geometrical relations (Sec. 2.1) and photometric relations (Sec. 2.2) between the unknown high-resolution sequence S^h and the input low-resolution sequences $\{S_i^l\}_{i=1}^n$.

2.1 The Space-Time Coordinate Transformations

In general a space-time dynamic scene is captured by a 4D representation (x, y, z, t). For simplicity, in this paper we deal with dynamic scenes which can be modeled by a 3D space-time volume (x, y, t) (see in Fig. 3.a). This assumption is valid if one of the following conditions holds: (i) the scene is planar and the dynamic events occur within this plane, or (ii) the scene is a general dynamic 3D scene, but the distances between the recording video cameras are small relative to their distance from the scene. (When the camera centers are very close to each other, there is no relative 3D parallax.) Under those conditions the dynamic scene can be modeled by a 3D space-time representation.

W.l.o.g., let S_1^l be a "reference" sequence whose axes are aligned with those of the continuous space-time volume S (the unknown dynamic scene we wish to reconstruct). S^h is a discretization of S with a higher sampling rate than that of S_1^l. Thus, we can model the transformation T_1 from the space-time coordinate system of S_1^l to the space-time coordinate system of S^h by a scaling transformation (the scaling can be different in time and in space). Let $T_{i\to 1}$ denote the space-time coordinate transformation from the reference sequence S_1^l to the i-th low resolution sequence S_i^l (see below). Then the space-time coordinate transfor-

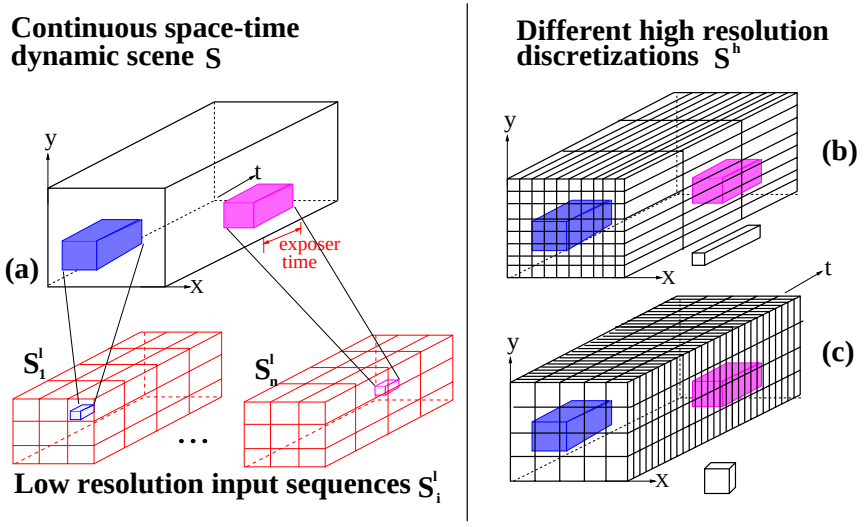


Fig. 3. The space-time imaging process. *(a) illustrates the space-time continuous scene and two of the low resolution sequences. The large pink (bright) and blue (dark) boxes are the support regions of the space-time blur corresponding to the low resolution space-time measurements marked by the respective small boxes. (b,c) show two different possible discretizations of the space-time volume resulting in two different high resolution output sequences. (b) has a low frame-rate and high spatial resolution, (c) has a high frame-rate but low spatial resolution.*

mation of each low-resolution sequence S_i^l is related to that of the high-resolution sequence S^h by $T_i = T_1 \cdot T_{i\to 1}$.

The space-time coordinate transformation between two input sequences ($T_{i\to 1}$) results from the different setting of the different cameras. A *temporal misalignment* between two sequences occurs when there is a time-shift (offset) between them (e.g., if the cameras were not activated simultaneously), or when they differ in their frame rates (e.g., PAL and NTSC). Such temporal misalignments can be modeled by a 1-D affine transformation in time, and is typically at sub-frame time units. The *spatial misalignment* between the two sequences results from the fact that the two cameras have different external and internal calibration parameters. In our current implementation, as mentioned above, because the camera centers are assumed to be very close or else the scene is planar, the spatial transformation can thus be modeled by an inter-camera homography. We computed these space-time coordinate transformations, using the method of [7], which provides high sub-pixel and high sub-frame accuracy.

Note that while the space-time coordinate transformations *between the sequences* ($\{T_i\}_{i=1}^n$) are very simple (a spatial homography and a temporal affine transformation), the motions occurring over time *within* the dynamic scene can be very complex. Our space-time super-resolution algorithm does *not* require knowledge of these motions, only the knowledge of $\{T_i\}_{i=1}^n$. It can thus handle very complex dynamic scenes.

2.2 The Space-Time Imaging Model

As mentioned earlier, the space-time imaging process induces spatial and temporal blurring in the low-resolution sequences. The temporal blur in the low-resolution sequence S_i^l is caused by the exposer-time τ_i of the i-th camera. The spatial blur in S_i^l is due to the spatial point-spread-function (PSF) of the i-th camera, which can be approximated by a 2D spatial Gaussian with std σ_i. (A method for estimating the PSF of a camera can be found in [11].)

Let $B_i = B_{(\sigma_i,\tau_i,p_i^l)}$ denote the combined space-time blur operator of the i-th camera corresponding to the low resolution space-time point $p_i^l = (x_i^l, y_i^l, t_i^l)$. Let $p^h = (x^h, y^h, t^h)$ be the corresponding high resolution space-time point $p^h = T_i(p_i^l)$ (p^h is not necessarily an integer grid point of S^h, but is contained in the continuous space-time volume S). Then the relation between the *unknown* space-time values $S(p^h)$, and the *known* low resolution space-time measurements $S_i^l(p_i^l)$, can be expressed by:

$$S_i^l(p_i^l) = \big(S * B_i^h\big)(p^h) = \int_x \int_y \int_t \underset{p\,=\,(x,\,y,\,t)\,\in\, Support(B_i^h)}{} S(p)\; B_i^h(p - p^h)dp \qquad (1)$$

where $B_i^h = T_i(B_{(\sigma_i,\tau_i,p_i^l)})$ is a point-dependent space-time blur kernel represented in the high resolution coordinate system. Its support is illustrated by the large pink (bright) and blue (dark) boxes in Fig. 3.a. To obtain a linear equation in the terms of the *discrete unknown* values of S^h we used a discrete approximation of Eq. (1). In our implementation we used a non-isotropic approximation in

the temporal dimension, and an isotropic approximation in the spatial dimension (see [6] for a discussion of the different discretization techniques in the context of image-based super-resolution). Eq. (1) thus provides a linear equation that relates the unknown values in the high resolution sequence S^h to the *known* low resolution measurements $S_i^l(p_i^l)$.

When video cameras of different photometric responses are used to produce the input sequences, then a preprocessing step is necessary that histogram-equalizes all the low resolution sequences. This step is required to guarantee consistency of the relation in Eq. (1) with respect to all low resolution sequences.

2.3 The Reconstruction Step

Eq. (1) provides a single equation in the high resolution unknowns for each low resolution space-time measurement. This leads to the following huge system of linear equations in the unknown high resolution elements of S^h:

$$A\overrightarrow{h} = \overrightarrow{l} \tag{2}$$

where $\overrightarrow{h}$ is a vector containing all the unknown high resolution color values (in YIQ) of S^h, $\overrightarrow{l}$ is a vector containing all the space-time measurements from all the low resolution sequences, and the matrix A contains the relative contributions of each high resolution space-time point to each low resolution space-time point, as defined by Eq. (1).

When the number of low resolution space-time measurements in $\overrightarrow{l}$ is greater than or equal to the number of space-time points in the high-resolution sequence S^h (i.e., in $\overrightarrow{h}$), then there are more equations than unknowns, and Eq. (2) can be solved using LSQ methods. This, however, implies that a large increase in the spatial resolution (which requires very fine spatial sampling in S^h) will come at the expense of a significant increase in the temporal resolution (which also requires fine temporal sampling in S^h), and vice versa. This is because for a given set of input low-resolution sequences, the size of $\overrightarrow{l}$ is fixed, thus dictating the number of unknowns in S^h. However, the number high resolution space-time points (unknowns) can be distributed differently between space and time, resulting in different space-time resolutions (see 4.2).

Directional space-time regularization. When there is an insufficient number of cameras relative to the required improvement in resolution (either in the entire space-time volume, or only in portions of it), then the above set of equations (2) becomes ill-posed. To constrain the solution and provide additional numerical stability (as in image-based super-resolution [9,5]), a space-time regularization term can be added to impose smoothness on the solution S^h in space-time regions which have insufficient information. We introduce a *directional* (or steerable [14]) space-time regularization term which applies smoothness only in directions where the derivatives are low, and does *not* smooth across space-time "edges". In other words, we seek $\overrightarrow{h}$ which minimize the following error term:

$$min\big(||A\overrightarrow{h} - \overrightarrow{l}||^2 + ||W_x L_x \overrightarrow{h}||^2 + ||W_y L_y \overrightarrow{h}||^2 + ||W_t L_t \overrightarrow{h}||^2\big) \tag{3}$$

Where L_j $(j = x, y, t)$ is matrix capturing the second-order derivative operator in the direction j, and W_j is a diagonal weight matrix which captures the degree of desired regularization at each space-time point in the direction j. The weights in W_j prevent smoothing across space-time "edges". These weights are determined by the location, orientation and magnitude of space-time edges, and are approximated using space-time derivatives in the low resolution sequences.

Solving the equations. The optimization problem of Eq. (3) has very large dimensionality. For example, even for a simple case of four low resolution input sequences, each one-second long (25 frames) and of size 128×128 pixels, we get: $128^2 \times 25 \times 4 \approx 1.6 \times 10^6$ equations from the low resolution measurements alone (without regularization). Assuming a similar number of high resolution unknowns poses a severe computational problem. However, matrix A is sparse and local (i.e., all the non zero entries are located in a few diagonals), the system of equations can be solved using "box relaxation" [15].

3 Examples: Temporal Super-Resolution

Empirical Evaluation. To examine the capabilities of temporal super-resolution in the presence of strong motion aliasing and strong motion blur, we first simulated a sports-like scene with a very fast moving object. We recorded a single video sequence of a basketball bouncing on the ground. To simulate high speed of the ball relative to frame-rate and relative to the exposure-time (similar to those shown in Fig. 1), we temporally blurred the sequence using a large (9-frame) blur kernel, followed by a large subsampling in time by factor of 30. This process results in a low temporal-resolution sequences of a very fast dynamic event having an "exposure-time" of about $\frac{1}{3}$ of its frame-time. We generated 18 such low resolution sequences by starting the temporal sub-sampling at arbitrary starting frames. Thus, the input low-resolution sequences are related by *non-uniform* sub-frame temporal offsets. Because the original sequence contained 250 frames, each generated low-resolution sequence contains only 7 frames. Three of the 18 sequences are presented in Fig 4.a-c. To visually display the event captured in each of these sequences, we super-imposed all 7 frames in each sequence. Each ball in the super-imposed image represents the location of the ball at a different frame. None of the 18 low resolution sequences captures the correct trajectory of the ball. Due to the severe motion aliasing, the perceived ball trajectory is roughly a smooth curve, while the true trajectory was more like a cycloid (the ball jumped 5 times on the floor). Furthermore, the shape of the ball is completely distorted in all input image frames, due to the strong motion blur.

We applied the super-resolution algorithm of Sec. 2 on these 18 low-resolution input sequences, and constructed a high-resolution sequence whose frame-rate is 30 times higher than that of the input sequences. (In this case we requested an increase only in the temporal sampling rate). The reconstructed high-resolution sequence is shown in Fig. 4.d. This is a super-imposed display of some of the

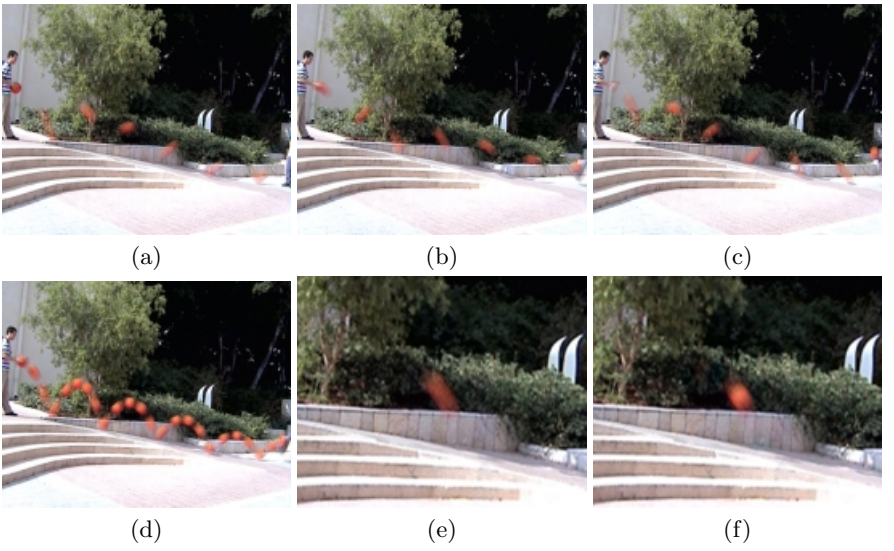


Fig. 4. Temporal super-resolution. *We simulated 18 low-resolution video recordings of a rapidly bouncing ball inducing strong motion blur and motion aliasing (see text). (a)-(c) Display the dynamic event captured by three representative low-resolution sequences. These displays were produced by super-position of all 7 frames in each low-resolution sequences. All 18 input sequences contain severe motion aliasing (evident from the falsely perceived curved trajectory of the ball) and strong motion blur (evident from the distorted shapes of the ball). (d) The reconstructed dynamic event as captured by the generated high-resolution sequence. The true trajectory of the ball is recovered, as well as its correct shape. (e) A close-up image of the distorted ball in one of the low resolution frames. (f) A close-up image of the ball at the exact corresponding frame in time in the high-resolution output sequence. For color sequences see: www.wisdom.weizmann.ac.il/∼vision/SuperRes.html*

reconstructed frames (every 8'th frame). The true trajectory of the bouncing ball has been recovered. Furthermore, Figs. 4(e)-(f) show that this process has significantly reduced effects of motion blur and the true shape of moving ball has been automatically recovered, although no single low resolution frame contains the true shape of the ball. Note that no estimation of the ball motion was needed to obtain these results. This effect is explained in more details in Sec. 4.1.

The above results obtained by temporal super-resolution cannot be obtained by playing any low-resolution sequence in "slow-motion" due to the strong motion aliasing. Such results cannot be obtained either by interleaving frames from the 18 input sequences, due to the non-uniform time shifts between the sequences and due to the severe motion-blur observed in the individual image frames.

A Real Example – The "Wagon-Wheel Effect". We used four independent PAL video cameras to record a scene of a fan rotating clock-wise very

fast. The fan rotated faster and faster, until at some stage it exceeded the maximal velocity that can be captured by video frame-rate. As expected, at that moment all four input sequences display the classical "wagon wheel effect" where the fan appears to be falsely rotating backwards (counter clock-wise). We computed the spatial and temporal misalignments between the sequences at sub-pixel and sub-frame accuracy using [7] (the recovered temporal misalignments are displayed in Fig. 5.a-d using a time-bar). We used the super-resolution method of Sec. 2 to increase the temporal resolution by a factor of 3 while maintaining the same spatial resolution. The resulting high-resolution sequence displays the true forward (clock-wise) motion of the fan, as if recorded by a high-speed camera (in this case, $75 frames/sec$) . Example of a few successive frames from each low resolution input sequence are shown in Fig.5.a-d for the portion where the fan appears to be rotating counter clock-wise. A few successive frames from the reconstructed high temporal-resolution sequence corresponding to the same time are shown in Fig.5.e, showing the correctly recovered (clock-wise) motion. It is difficult to perceive these strong dynamic effects via a static figure (Fig. 5). We therefore urge the reader to view the video clips in www.wisdom.weizmann.ac.il/∼vision/SuperRes.html where these effects are very vivid . Furthermore, playing the input sequences in "slow-motion" (using any type of temporal interpolation) will *not* reduce the perceived false motion effects.

4 Space-Time Visual Tradeoffs

The spatial and temporal dimensions are very different in nature, yet are interrelated. This introduces visual tradeoffs between space and time, which are unique to spatio-temporal super-resolution, and are not applicable to traditional spatial (i.e., image-based) super-resolution.

4.1 Temporal Treatment of Spatial Artifacts

When an object moves fast relative to the exposure time of the camera, it induces observable motion-blur (e.g., see Fig. 1). The perceived distortion is spatial, however the cause is temporal. We next show that by increasing the *temporal* resolution we can handle the *spatial* artifacts caused by motion blur.

Motion blur is caused by the extended temporal blur due to the exposure-time. To decrease effects of motion blur we need to decrease the temporal blur, i.e., recover high temporal frequencies. This requires increasing the frame-rate beyond that of the low resolution input sequences. In fact, to decrease the effect of motion blur, the output temporal sampling rate must be increased so that the distance between the new high resolution temporal samples is *smaller* than the original exposure time of the low resolution input sequences.

This indeed was the case in the experiment of Fig. 4. Since the simulated exposure time in the low resolution sequences was 1/3 of frame-time, an increase in temporal sampling rate by a factor > 3 can reduce the motion blur. The larger

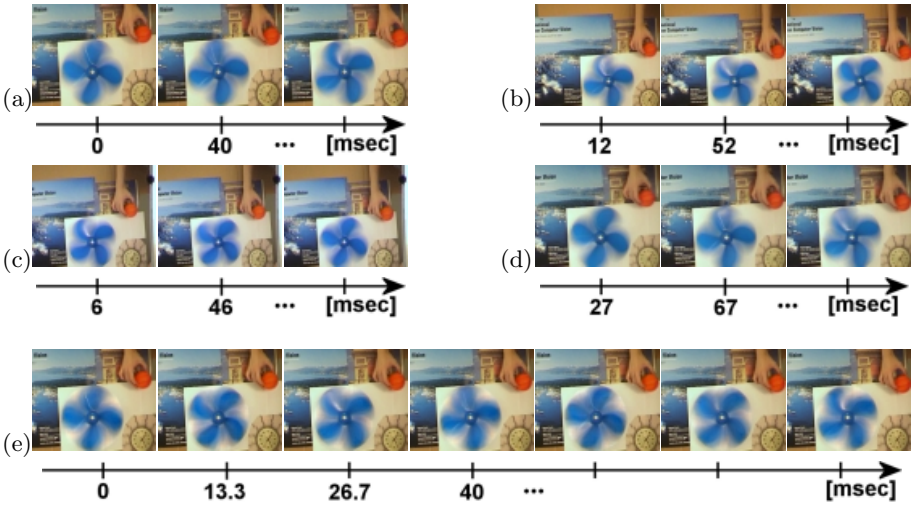


Fig. 5. Temporal super-resolution (the "wagon wheel effect"). *(a)-(d) display 3 successive frames from four PAL video recordings of a fan rotating clock-wise. Because the fan is rotating very fast (almost* 90^o *between successive frames), the motion aliasing generates a false perception of the fan rotating slowly in the opposite direction (counter clock-wise) in all four input sequences. The temporal misalignments between the input sequences were computed at sub-frame temporal accuracy, and are indicated by their time bars. The spatial misalignments between the sequences (e.g., due to differences in zoom and orientation) were modeled by a homography, and computed at sub-pixel accuracy. (e) shows the reconstructed video sequence in which the temporal resolution was increased by a factor of 3. The new frame rate (*$75\frac{frames}{sec}$*) is also indicated by a time bars. The correct clock-wise motion of the fan is recovered. For color sequences see: www.wisdom.weizmann.ac.il/∼vision/SuperRes.html*

the increase the more effective the motion deblurring would be. This increase is limited, of course, by the number of input cameras.

A method for treating motion blur in the context of *image-based* super-resolution was proposed by [2,12]. However, these methods require a prior segmentation of moving objects and the estimation of their motions. These methods will have difficulties handling complex motions or motion aliasing. The distorted shape of the object due to strong blur (e.g., Fig. 1) will pose severe problems in motion estimation. Furthermore, in the presence of motion aliasing, the direction of the estimated motion will not align with the direction of the induced blur. For example, the motion blur in Fig. 4.a-c. is along the true trajectory and not along the perceived one. In contrast, our approach does not require separation of static and dynamic scene components, nor their motion estimation, thus can handle very complex scene dynamics. However, we require multiple cameras.

Temporal frequencies in video sequences have very different characteristics than spatial frequencies, due to the different characteristics of the temporal and

the spatial blur. The typical support of the spatial blur (PSF) is of a few pixels ($\sigma > 1$ *pixel*), whereas the exposure time is usually smaller than a single frame-time ($\tau <$ frame-time). Therefore, if we do not increase the output temporal sampling-rate *enough*, we will not improve the temporal resolution. In fact, if we increase the temporal sampling-rate a little but not beyond $\frac{1}{exposure\ time}$ of the low resolution sequences, we may even introduce *additional* motion blur.

This dictates the number of input cameras needed for an effective decrease in the motion-blur. An example of a case where an insufficient increase in the temporal sampling-rate introduced additional motion-blur is shown in Fig. 6.c3.

4.2 Producing Different Space-Time Outputs

In standard spatial super-resolution the increase in sampling rate is equal in all spatial dimensions. This is necessary in order to maintain the aspect ratio of image pixels, and to prevent distorted-looking images. However, this is not the case in space-time super-resolution. As explained in Sec. 2, the increase in sampling rate in the spatial and temporal dimensions need not be the same. Moreover, increasing the sampling rate in the spatial dimension comes at the expense of increase in the temporal frame rate, and vice-versa. This is because the number of unknowns in the high-resolution space-time volume depends on the space-time sampling rate, whereas the number of equations provided by the low resolution measurements remains fixed.

For example, assume that 8 video cameras are used to record a dynamic scene. One can increase the spatial sampling rate alone by a factor of $\sqrt{8}$ in x and y, or increase the temporal frame-rate alone by a factor of 8, or do a bit of both: increase the sampling rate by a factor of 2 in all three dimensions. Such an example is shown in Fig. 6. Fig. 6.a1 displays one of 8 low resolution input sequences. (Here we used only 4 video cameras, but split them into 8 sequences of even and odd fields). Figs. 6.a2 and 6.a3 display two possible outputs. In Fig. 6.a2 the increase is by a factor of 8 in the temporal axis with no increase in the spatial axes, and in Fig. 6.a3 the increase is by a factor of 2 in all axes x,y,t. Rows (b) and (c) illustrate the corresponding visual tradeoffs. The "$\times 1 \times 1 \times 8$" option (column 2) decreases the motion blur of the moving object (the toothpaste in (c.2)), while the "$\times 2 \times 2 \times 2$" option (column 3) improves the spatial resolution of the static background (b.3), but increases the motion blur of the moving object (c.3). The latter is because the increase in frame rate was only by factor 2 and did not exceed $\frac{1}{exposure\ time}$ of the video camera (see Sec. 4.1). In order to create a significant improvement in all dimensions, more than 4 video cameras are needed.

4.3 Combining Different Space-Time Inputs

So far we assumed that all input sequences were of similar spatial and temporal resolutions. The space-time super-resolution algorithm of Sec. 2 is not restricted to this case, and can handle input sequences of varying space-time resolutions.

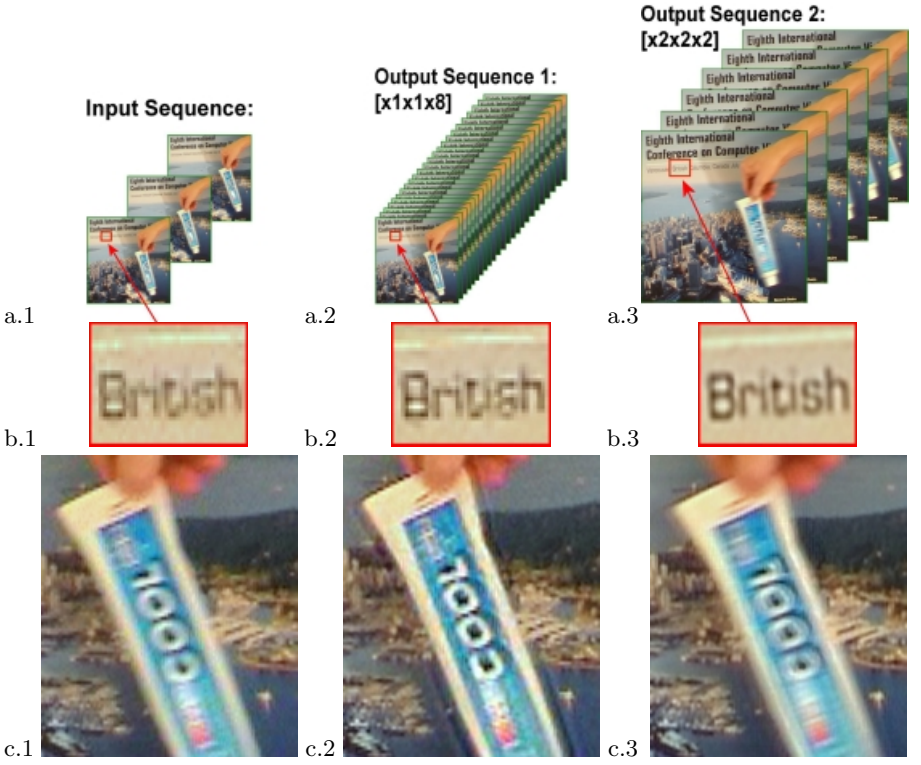


Fig. 6. Tradeoffs between spatial and temporal resolution. *This figure compares the visual tradeoffs resulting from applying space-time super-resolution with different discretization of the space-time volume. (a.1) displays one of eight low-resolution input sequences of a toothpaste in motion against a static background. (b.1) shows a close-up image of a static portion of the scene (the writing on the poster), and (c.1) shows a dynamic portion of the scene (the toothpaste). Column 2 (a.2, b.2, c.2) displays the resulting spatial and temporal effects of applying super-resolution by a factor of 8 in time only. Motion blur of the toothpaste is decreased. Column 3 (a.3, b.3, c.3) displays the resulting spatial and temporal effects of applying super-resolution by a factor of 2 in all three dimensions x, y, t. The spatial resolution of the static portions is increased (see "British" and the yellow line above it in b.3), but the motion blur is also increased (c.3). See text for an explanation of these visual tradeoffs. For color sequences see: www.wisdom.weizmann.ac.il/∼vision/SuperRes.html*

Such a case is meaningless in image-based super-resolution, because a high resolution input image would always contain the information of a low resolution image. In space-time super-resolution however, this is not the case. One camera may have high spatial but low temporal resolution, and the other vice-versa. Thus, for example, it is meaningful to combine information from NTSC and PAL video cameras. NTSC has higher temporal resolution than PAL ($30f/sec$

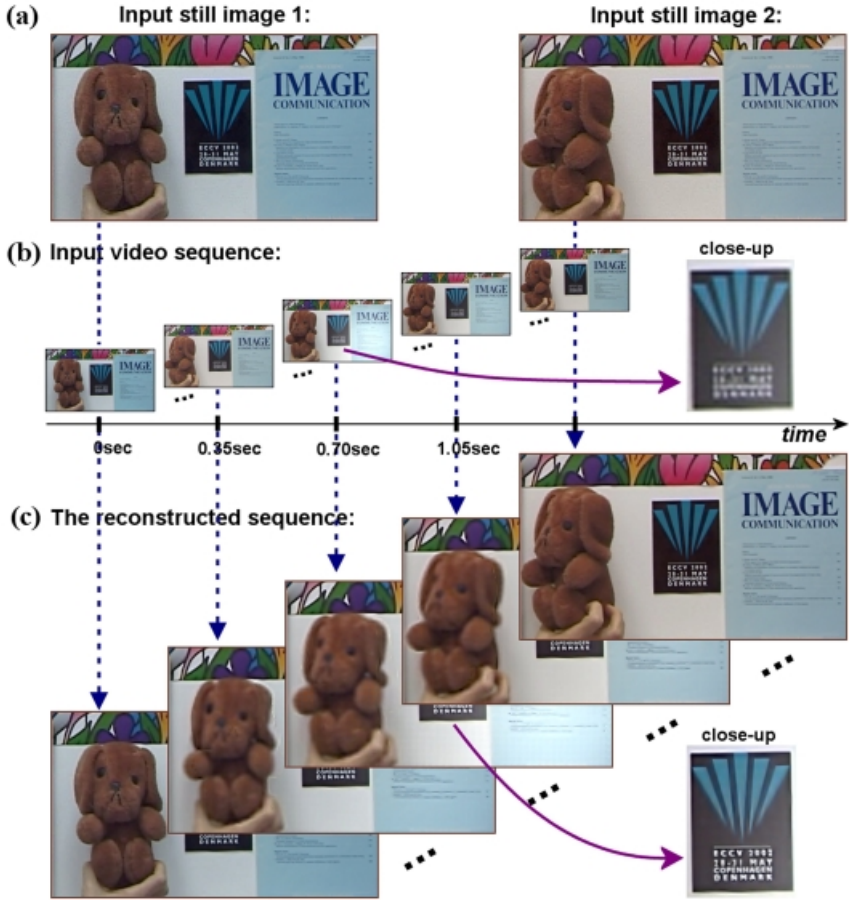


Fig. 7. Combining Still and Video. *A dynamic scene of a rotating toy-dog and varying illumination was captured by: (a) A still camera with spatial resolution of* 1120×840 *pixels, and (b) A video camera with* 384×288 *pixels at 50 f/sec. The video sequence was* $1.4sec$ *long (*70 *frames), and the still images were taken* $1.4sec$ *apart (together with the first and last frames). The algorithm of Sec. 2 is used to generate the high resolution sequence (c). The output sequence has the spatial dimensions of the still images and the frame-rate of the video (*$1120\times840\times50$*). It captures the temporal changes correctly (the rotating toy and the varying illumination), as well the high spatial resolution of the still images (the sharp text). Due to lack of space we show only a portion of the images, but the proportions between video and still are maintained. For color sequences see: www.wisdom.weizmann.ac.il/∼vision/SuperRes.html*

vs. $25f/sec$), but lower spatial resolution (640×480 pixels vs. 768×576 pixels). An extreme case of this idea is to combine information from *still* and *video* cameras. Such an example is shown in Fig. 7. Two high quality still images of high spatial

resolutions (1120×840 pixels) but extremely low "temporal resolution" (the time gap between the two still images was 1.4 sec), were combined with an interlaced (PAL) video sequence using the algorithm of Sec 2. The video sequence has 3 times lower spatial resolution (we used fields of size 384×288 pixels), but a high temporal resolution ($50f/sec$). The goal is to construct a new sequence of high spatial and high temporal resolutions (i.e., 1120×840 pixels at 50 $images/sec$).

The output sequence shown in Fig. 7.c contains the high spatial resolution from the still images (the sharp text) and the high temporal resolution from the video sequence (the rotation of the toy dog and the brightening and dimming of illumination).

In the example of Fig. 7 we used only one input sequence and two still images, thus did not exceed the temporal resolution of the video or the spatial resolution of the stills. However, when multiple video cameras and multiple still images are used, the number of input measurements will exceed the number of output high resolution unknowns. In such cases the output sequence will exceed the spatial resolution of the still images and temporal resolution of the video sequences.

In Fig. 7 the number of unknowns was significantly larger than the number of low resolution measurements (the input video and the two still images). Yet, the reconstructed output was of high quality. The reason for this is the following: In video sequences the data is significantly more redundant than in images, due to the additional time axis. This redundancy provides more flexibility in applying *physically meaningful* directional regularization. In regions that have high spatial resolution but small (or no) motion (such as in the sharp text in Fig. 7), strong *temporal* regularization can be applied without decreasing the space-time resolution. Similarly, in regions with very fast dynamic changes but low spatial resolution (such as in the rotating toy in Fig. 7), strong *spatial* regularization can be employed without degradation in space-time resolution. More generally, because a video sequence has much more data redundancy than an image has, the use of *directional space-time regularization* in video-based super-resolution is physically more meaningful and gives rise to recovery of higher space-time resolution than that obtainable by image-based super-resolution with image-based regularization.

Acknowledgments. The authors wish to thank Merav Galun & Achi Brandt for their helpful suggestions regarding solutions of large scale systems of equations. Special thanks to Ronen Basri & Lihi Zelnik for their useful comments on the paper.

References

1. S. Baker and T. Kanade. Limits on super-resolution and how to break them. In *CVPR*, Hilton Head Island, South Carolina, June 2000.
2. B. Bascle, A. Blake, and A.Zisserman. Motion deblurring and super-resolution from an image sequence. In *ECCV*, pages 312–320, 1996.

3. S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. Technical report, Laboratory for Image and Signal Analysis (LISA), University of Notre Dame, Notre Dame, July 1998.
4. M. Born and E. Wolf. *Principles of Optics*. Permagon Press, 1965.
5. D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *CVPR*, pages 885–891, June 1998.
6. D. Capel and A. Zisserman. Super-resolution enhancement of text image sequences. In *ICPR*, pages 600–605, 2000.
7. Y. Caspi and M. Irani. A step towards sequence-to-sequence alignment. In *CVPR*, pages 682–689, Hilton Head Island, South Carolina, June 2000.
8. G. de Haan. Progress in motion estimation for consumer video format conversion. *IEEE Transactions on Consumer Electronics*, 46(3):449–459, August 2000.
9. M. Elad. Super-resolution reconstruction of images. Ph.D. Thesis, Technion Israel Institute of Technology, December 1996.
10. T.S. Huang and R.Y. Tsai. Multi-frame image restoration and registration. In *Advances in Computer Vision and Image Processing*, volume 1, pages 317–339. JAI Press Inc., 1984.
11. M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP:GM*, 53:231–239, May 1991.
12. A. J. Patti, M. I. Sezan, and A. M. Tekalp. Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. In *IEEE Trans. on Image Processing*, volume 6, pages 1064–1076, August 1997.
13. REALVIZ^{TM}. Retimer. *www.realviz.com/products/rt*, 2002.
14. J. Shin, J. Paik, J. R. Price, and M.A. Abidi. Adaptive regularized image interpolation using data fusion and steerable constraints. In *SPIE Visual Communications and Image Processing*, volume 4310, January 2001.
15. U. Trottenber, C. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, 2000.