

# A Study of the Temporal Bandwidth of Video and its Implications in Compressive Sensing

Michael B. Wakin\*

Colorado School of Mines  
Technical Report 2012-08-15

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Videos with one spatial dimension</b>	<b>3</b>
2.1	Problem setup . . . . .	3
2.2	Temporal bandwidth analysis . . . . .	4
2.3	Sampling implications . . . . .	6
2.4	Experiments within our model assumptions . . . . .	7
2.5	Experiments beyond our model assumptions . . . . .	9
<b>3</b>	<b>Videos with two spatial dimensions</b>	<b>12</b>
3.1	Problem setup . . . . .	12
3.2	Temporal bandwidth analysis . . . . .	13
3.3	Sampling implications . . . . .	15
3.4	Experiments . . . . .	15
<b>4</b>	<b>Sampling and interpolation principles</b>	<b>16</b>
4.1	Interpolation theory . . . . .	16
4.2	Experiments within our model assumptions . . . . .	17
4.3	Experiments beyond our model assumptions . . . . .	18
4.4	A manifold-based interpretation . . . . .	18
<b>5</b>	<b>Reconstruction of temporally bandlimited videos from streaming measurements</b>	<b>19</b>
5.1	Measurement process . . . . .	19
5.2	Simplifying the linear equations . . . . .	20
5.3	Experiments . . . . .	21
5.4	Toward a general purpose reconstruction algorithm . . . . .	23

---

\*MBW is with the Department of Electrical Engineering and Computer Science at the Colorado School of Mines, Golden. Email: mwakin@mines.edu. This research was supported by ONR Grant N00014-09-1-1162. A summary of this research was presented at the SIAM Conference on Imaging Science in May 2012.

# 1 Introduction

The task of recovering a video from compressive measurements can be very challenging due to the sheer dimensionality of the data to be recovered. In streaming measurement systems such as the Rice “single-pixel camera” [1, 2], this challenge can be particularly daunting, because we may record as little as one measurement per time instant. Presented with this information, it would be possible to follow standard Compressive Sensing (CS) formulations and set up a system of linear equations that relate the measurements to the underlying high-rate video voxels. However, since the video can change from one instant to the next, the number of unknown voxels in such a formulation would unfortunately scale with the number of measurements collected. Moreover, if the inter-measurement time interval were to go to zero, the size of the reconstruction problem would apparently continue to grow.

There are reasons to believe that we should be able control this explosion of dimensionality. Intuitively, for example, we know that as the inter-measurement time interval goes to zero, the video should change very little between the times when adjacent measurements are collected. In this document, we argue that underlying high-rate video voxels—despite growing in number—actually have a bounded complexity, and that one way of characterizing their relatively few degrees of freedom is to study the temporal bandlimitedness of the video. We argue analytically that, at the imaging sensor, many videos should have limited temporal bandwidth due to the spatial lowpass filtering that is inherent in typical imaging systems. Under modest assumptions about the motion of objects in the scene, this spatial filtering prevents the temporal complexity of the video from being arbitrarily high. Consequently, we conclude that under various interpolation kernels (sinc, linear, etc.), the unknown high-rate voxels can be represented as a linear combination of a low-rate (e.g., Nyquist-rate) “anchor” set of sampled video frames.

Our bandwidth analysis uses fairly standard arguments from imaging processing and communications theory but goes deeper than the “constant velocity” model commonly assumed for object motion. Typical analysis under the constant velocity model reflects that the video’s temporal bandwidth will be proportional to the object’s (constant) speed. For more representative motion models, we show that that proportionality still basically holds true, but with a  $2\times$  higher constant of proportionality, and with an additional additive factor due to the bandwidth of the signal tracing the object’s path. Although there is room for debate in the assumptions we make, our analysis reveals an interesting tradeoff between the spatial resolution of the camera, the speed of any moving objects, and the temporal bandwidth of the video. We use a study of synthetic and real videos to quantify how well this bandlimited/interpolation model fits. The mathematics that appear in our analysis are similar to those that have appeared in the analysis of plenoptic [3] and plenacoustic [4, 5] functions, which arise from measuring light and sound, respectively, at various positions in a room.

After arguing that many videos have limited temporal bandwidth, we revisit the CS reconstruction problem. We explain how the problem can be reformulated by setting up a system of linear equations that relate the number of measurements to the underlying degrees of freedom of the video (specifically, the anchor frames). Notably, the number of anchor frames does not depend on the temporal measurement rate, and so the reconstruction problem can be much more tractable than in the original formulation described above. We demonstrate experimentally that the CS reconstruction performance depends on the type of factors that impact the video’s temporal bandwidth (e.g., spatial resolution of the camera, the speed of any moving objects, etc.), and we explore possible tradeoffs in the design of a reconstruction algorithm. Using even a linear interpolation kernel, we

also show that the reconstruction performance is significantly better than what can be achieved by raw “aggregation” of measurements, which corresponds to using a rectangular, nearest neighbor interpolation kernel.

In the hypothetical case where no residual correlations remain among the anchor frames, there would be little room for improvement upon this reconstruction approach. In real-world videos, however, we do expect that significant correlations may remain. The question of how to exploit these correlations has been addressed in the more conventional CS recovery literature. For example, one could employ a linear dynamical system model to capture the restricted degrees of freedom of the anchor frames [6]. Alternatively, we are currently developing an algorithm that will involve estimating the motion of the video and then formulating the recovery of the anchor frames in a motion-compensated sparse basis [7]; the details of this algorithm will be reported in a future publication. As better solutions continue to be developed for exploiting such correlations, we expect that they could be combined with the methods discussed in this paper to permit better reconstruction from streaming or “single-pixel” measurements.

## 2 Videos with one spatial dimension

### 2.1 Problem setup

#### 2.1.1 Signal model

We start our analysis by considering “videos” that have just one spatial dimension. We will use the variable  $t \in \mathbb{R}$  to index time (which we measure in seconds), and we will use the variable  $x \in \mathbb{R}$  to index spatial position (which for convenience we measure in an arbitrary real-valued unit we call “pix”). Though we will begin by considering continuous-space, continuous-time videos, the “pix” unit is intended to symbolize what might be the typical pixel size in a subsequent discretization of this video. One could easily replace pix with meters or any other arbitrary unit of distance.

Now, let  $g(x)$  denote a 1D function of space (think of this as a continuous-space “still image”), and consider a continuous-space, continuous-time video  $f(x, t)$  in which each “frame” of the video merely consists of a shifted version of this prototype frame. More formally, suppose that

$$f(x, t) = g(x - h(t))$$

where  $h(t)$  is some function that controls how much (in pix) the prototype frame is shifted at each time step.

#### 2.1.2 Bandwidth considerations

Because we have an interest in video imaging systems with high temporal sampling rates, our purpose in this document is to describe some conditions under which the model video  $f(x, t)$  will have finite temporal bandwidth. This limited bandwidth, in turn, suggests that such videos can be characterized in terms of a temporal Nyquist-rate sampling of frames, that one may characterize the performance of various temporal interpolation procedures, etc.

We suggest that the limited temporal bandwidth of  $f(x, t)$  could arise in plausible scenarios under which the prototype frame  $g(x)$  and translation signal  $h(t)$  have limited complexity. For example, in a physical imaging system, we may envision  $f(x, t)$  as the video that exists at the imaging sensor prior to sampling. It may be reasonable to expect that, due to optical blurring and due to the implicit filtering that occurs from the spatial extent of each light integrator, the prototype

frame  $g(x)$  may have limited spatial bandwidth. Similarly, if the camera motion is constrained or due to the physics governing the movement of objects in the scene, one might expect that the translation signal  $h(t)$  may have limited slope and/or limited temporal bandwidth. In the sections that follow, we will explain how such scenarios can allow us to bound the approximate temporal bandwidth of  $f(x, t)$ .

### 2.1.3 Fourier setup

Let  $F(\omega_x, \omega_t)$  denote the 2D Fourier transform of  $f(x, t)$ , and let  $G(\omega_x)$  denote the 1D Fourier transform of  $g(x)$ . Keeping track of units,  $\omega_x$  is measured in terms of rad/pix, and  $\omega_t$  is measured in terms of rad/s. The following relationships will be of use to us.

First, let  $\mathcal{F}_x\{\cdot\}$  and  $\mathcal{F}_t\{\cdot\}$  denote operators on  $L_2(\mathbb{R}^2)$  that perform 1D Fourier transforms in the spatial and temporal directions, respectively. Due to the separability of the 2D Fourier transform, we know that

$$F(\omega_x, \omega_t) = \mathcal{F}_x\{\mathcal{F}_t\{f\}\}(\omega_x, \omega_t) = \mathcal{F}_t\{\mathcal{F}_x\{f\}\}(\omega_x, \omega_t). \quad (1)$$

Now, due to the shift property of the 1D Fourier transform, we have

$$\mathcal{F}_x\{f\}(\omega_x, t) = \mathcal{F}_x\{g(x - h(t))\}(\omega_x, t) = G(\omega_x)e^{-j\omega_x h(t)}.$$

We note that this function is complex-valued and has constant magnitude in the temporal direction; therefore it has only phase changes in the temporal direction. Following (1),  $F$  will be the result of taking the 1D Fourier transform of this function  $\mathcal{F}_x\{f\}$  in the temporal direction. We have

$$F(\omega_x, \omega_t) = \mathcal{F}_t\{G(\omega_x)e^{-j\omega_x h(t)}\}(\omega_x, \omega_t) = G(\omega_x) \cdot L(\omega_x, \omega_t),$$

where

$$L(\omega_x, \omega_t) := \mathcal{F}_t\{e^{-j\omega_x h(t)}\}(\omega_x, \omega_t). \quad (2)$$

For fixed  $\omega_x$ ,  $L(\omega_x, \omega_t)$  equals the 1D Fourier transform of  $e^{-j\omega_x h(t)}$  with respect to time, evaluated at the frequency  $\omega_t$ .

## 2.2 Temporal bandwidth analysis

The appearance of the  $h(t)$  term within an exponent in (2) can complicate the task of characterizing the bandwidth of  $f(x, t)$  in terms of properties of  $h(t)$ . However, by imposing certain assumptions on  $h(t)$ , this analysis can become tractable. In Section 2.2.1 below, we briefly discuss a “constant velocity” model for  $h(t)$  that is commonly seen in textbook discussions of video bandwidth (see, e.g., [8]). In Section 2.2.2, we then go more deeply into the analysis of a “bounded velocity” model; while we are not aware of any such analysis in the existing image and video processing literature, the concepts do borrow heavily from standard results in communications and modulation theory (see, e.g., [9]). Except where noted, all subsequent discussion in this document will build on this bounded velocity analysis, rather than the constant velocity analysis.

### 2.2.1 Constant velocity model for $h(t)$

Our analysis simplifies dramatically if we assume that  $h(t) = \Gamma t$  for some constant  $\Gamma$  (having units of pix/s). In this case we have  $L(\omega_x, \omega_t) = \delta(\omega_t + \omega_x \Gamma)$ , and so

$$F(\omega_x, \omega_t) = G(\omega_x) \cdot \delta(\omega_t + \omega_x \Gamma),$$



which corresponds to a diagonal line in the 2D Fourier plane with slope (say,  $\Delta\omega_t$  over  $\Delta\omega_x$ ) that depends linearly on  $\Gamma$ .

To see the implications of this in terms of bandwidth, let us suppose that  $G(\omega_x)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $\omega_x \in [-\Omega_x, \Omega_x]$  rad/pix. (This may occur if the profile  $g(x)$  has been blurred in the  $x$  direction, for example.) In this case, it follows that  $F(\omega_x, \omega_t)$  must be bandlimited (or essentially bandlimited) to the range of frequencies  $(\omega_x, \omega_t) \in [-\Omega_x, \Omega_x] \times [-\Gamma\Omega_x, \Gamma\Omega_x]$ . In other words, the temporal bandwidth of the video is no greater than  $\Gamma\Omega_x$  rad/s.

### 2.2.2 Bounded velocity model for $h(t)$

We now consider a more robust “bounded velocity” model for  $h(t)$ ; we note that similar mathematics have appeared in the analysis of plenoptic [3] and plenacoustic [4, 5] functions, which arise from measuring light and sound, respectively, at various positions in a room. We assume that the position function  $h(t)$  has bounded slope, i.e., that for some  $\Gamma > 0$ ,

$$\left| \frac{dh(t)}{dt} \right| \leq \Gamma \text{ pix/s}$$

for all  $t$ . This corresponds to a bound on the speed at which the object can move in the video, without requiring that this speed be constant. We also assume that the translation signal  $h(t)$  is bandlimited, with bandwidth given by  $\Omega_h$  rad/s.

For any fixed  $\omega_x$ , we can recognize  $e^{-j\omega_x h(t)}$  as a phase-modulated (PM) sinusoid having carrier frequency 0 and phase deviation (in radians)

$$\phi(t) = -\omega_x h(t),$$

or equivalently, as a frequency-modulated (FM) sinusoid having carrier frequency 0 and instantaneous frequency (in rad/s)

$$\omega_i(t) = \frac{d\phi(t)}{dt} = -\omega_x \frac{dh(t)}{dt}.$$

For an FM signal, we have

$$\frac{d\phi(t)}{dt} = \omega_d m(t),$$

where  $m(t)$  is known as the modulating signal and  $\omega_d$  is the frequency deviation constant. Thus, we have

$$m(t) = \frac{1}{\omega_d} \frac{d\phi(t)}{dt} = \frac{-\omega_x}{\omega_d} \frac{dh(t)}{dt}.$$

Let us also define the deviation term

$$D := \frac{\omega_d}{\Omega_h} \max |m(t)| = \frac{|\omega_x|}{\Omega_h} \max \left| \frac{dh(t)}{dt} \right| \leq \frac{|\omega_x| \Gamma}{\Omega_h}.$$

From Carson’s bandwidth rule for frequency modulation, we have that for fixed  $\omega_x$ , at least 98% of the total power of  $e^{-j\omega_x h(t)}$  must be concentrated in the frequency range  $\omega_t \in [-2(D+1)\Omega_h, 2(D+1)\Omega_h]$  rad/s. Since  $D \leq \frac{|\omega_x| \Gamma}{\Omega_h}$ , we conclude that at least 98% of the total power of  $e^{-j\omega_x h(t)}$  must be concentrated in the frequency range  $\omega_t \in [-(2|\omega_x| \Gamma + 2\Omega_h), 2|\omega_x| \Gamma + 2\Omega_h]$  rad/s. We note that the dependence of this bandwidth on  $\omega_x$  is essentially linear.

We conclude that  $L(\omega_x, \omega_t)$  will have a characteristic “butterfly shape” with most of its total power concentrated between two diagonal lines that intercept the  $\omega_t$ -axis at  $\pm 2\Omega_h$  and have slope approximately  $\pm 2\Gamma$ . This shape is illustrated in Figure 1(a). Though not shown, the corresponding figure for the constant velocity model discussed in Section 2.2.1 would involve a single diagonal line intersecting the origin and having slope  $-\Gamma$  (which is half as large as the slope that appears in our more general bounded velocity analysis).

To see the implications of this in terms of bandwidth, let us again suppose that  $G(\omega_x)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $\omega_x \in [-\Omega_x, \Omega_x]$ . We must then have that  $F(\omega_x, \omega_t) = G(\omega_x) \cdot L(\omega_x, \omega_t)$  is also essentially bandlimited in the spatial direction to the range of frequencies  $\omega_x \in [-\Omega_x, \Omega_x]$ . Because of the butterfly structure in  $L(\omega_x, \omega_t)$ , however, this will also cause  $F(\omega_x, \omega_t)$  to be essentially bandlimited in the temporal direction to the range of frequencies

$$\omega_t \in [-(2\Omega_x\Gamma + 2\Omega_h), 2\Omega_x\Gamma + 2\Omega_h] \text{ rad/s.} \quad (3)$$

This fact, which is illustrated in Figure 1(b), exemplifies a central theme of our work: *filtering a video in the spatial direction can cause it to be essentially bandlimited both in space and in time.* Once again, we note that similar conclusions were reached in the analysis of plenoptic [3] and plenacoustic [4, 5] functions.

### 2.3 Sampling implications

Based on the temporal bandwidth predicted in (3), where we assume that  $G(\omega_x)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $\omega_x \in [-\Omega_x, \Omega_x]$ , the Nyquist theorem suggests that in order to avoid aliasing, the video should be sampled a minimum rate of

$$\frac{2\Omega_x\Gamma + 2\Omega_h}{\pi} \text{ samples/s.}$$

As a point of reference, during a sampling interval of  $\left(\frac{2\Omega_x\Gamma + 2\Omega_h}{\pi}\right)^{-1}$  seconds, an object moving with maximum speed  $\Gamma$  pix/s can traverse a maximum of

$$\frac{\pi\Gamma}{2\Omega_x\Gamma + 2\Omega_h} \text{ pix.} \quad (4)$$

Let us plug in some plausible numbers to illustrate the implications of these bounds. First, consider the spatial bandwidth  $\Omega_x$  of the prototype frame. In a reasonable imaging system, we might expect the pixel size to be balanced with the spatial bandwidth of the frame so that spatial aliasing is avoided. (This should occur naturally if we assume each pixel integrates spatially over a window of size approximately 1 pix.) Thus, one might anticipate that  $\Omega_x$  will be on the order of  $\pi$  rad/pix. (This corresponds to a spatial bandwidth of  $\frac{\pi}{2\pi} = \frac{1}{2}$  cycles/pix, which suggests a spatial Nyquist sample rate of one sample per pix.)

Under the assumption that  $\Omega_x = \pi$ , (3) suggests the video will have temporal bandwidth limited to approximately  $2\pi\Gamma + 2\Omega_h$ . We note that  $\Omega_h$ , the temporal bandwidth of  $h(t)$ , does not depend on the amplitude or slope of  $h(t)$ , but only on its shape and smoothness. The term  $\Gamma$ , in contrast, increases with the amplitude or slope of  $h(t)$ , which in turn could increase for objects closer to the camera. We conjecture that in practice, the  $2\pi\Gamma$  term will typically dominate the  $2\Omega_h$  term.<sup>1</sup> If

---

<sup>1</sup>A refined “Carson-type” analysis would permit the consideration of functions  $h(t)$  that are not strictly bandlimited.

this is indeed the case, (4) suggests that, in typical scenarios, to avoid temporal aliasing we should not allow a moving object to traverse more than  $\approx \frac{1}{2}$  pix between adjacent sampling times. While this of course makes strong intuitive sense, we have arrived at this conclusion through a principled analysis.

Let us make this even more concrete by further specifying some example parameter values. For an imaging system with, say, 1000 pixels per row, a distant object might move a maximum of  $\Gamma \approx 10$  pix/s, while a close object could move at up to  $\Gamma \approx 1000$  pix/s. These suggest minimum temporal bandwidths of  $20\pi$  rad/s and  $2000\pi$  rad/s, respectively, which correspond to minimum temporal Nyquist sampling rates of 20 samples/s and 2000 samples/s, respectively. As predicted above, these sampling rates would allow the object to move no more than about  $\frac{1}{2}$  pix per temporal sample.

Aside from exceptionally non-smooth motions  $h(t)$ , we do strongly suspect that the influence of the temporal bandwidth  $\Omega_h$  will be minor in comparison to the influence of the  $2\pi\Gamma$  term, and so *in general a temporal Nyquist sampling rate of  $2\Gamma$  samples/s will likely serve as a reasonable rule of thumb*. Certainly, this rule of thumb illustrates the direct relationship between the speed of object motion in the video and the video’s overall temporal bandwidth.

## 2.4 Experiments within our model assumptions

In this section, we analytically define continuous-space, continuous-time videos that allow us to test these behaviors and observe the tradeoffs between object speed, spatial filtering, and temporal bandwidth. In all experiments in this section, we let the prototype function

$$g(x) = \text{sinc}\left(\frac{\Omega_x x}{\pi}\right) \quad (5)$$

for some value of  $\Omega_x$  that we set as desired. This definition ensures that  $g(x)$  is bandlimited and that its bandwidth equals precisely  $\Omega_x$  rad/pix.

We oversample the videos compared to the predicted spatial and temporal bandwidths and show the approximate spectrum using the FFT. (All plots in fact show the magnitude of the FFT on a  $\log_{10}$  scale.) In some cases we apply a smooth Blackman-Harris window to the samples before computing the FFT; this helps remove artifacts from the borders of the sampling region.

### 2.4.1 Constant velocity model for $h(t)$

For our first experiment, we let the translation signal

$$h(t) = \Gamma t$$

for some value of  $\Gamma$  that we set as desired. As in (5), we let  $g(x) = \text{sinc}\left(\frac{\Omega_x x}{\pi}\right)$  for some value of  $\Omega_x$  that we set as desired.

Based on our discussion in Section 2.2.1, we anticipate that for this video,  $F(\omega_x, \omega_t)$  will be supported along a diagonal line in the 2D Fourier plane with slope (say,  $\Delta\omega_t$  over  $\Delta\omega_x$ ) equal to  $\Gamma$ . Figure 2 illustrates our estimated FFT spectrum for one experiment with  $\Omega_x = \pi$  rad/pix and  $\Gamma = 25$  pix/s; the solid blue box outlines the region  $(\omega_x, \omega_t) \in [-\Omega_x, \Omega_x] \times [-\Gamma\Omega_x, \Gamma\Omega_x]$ , and the diagonal, dashed blue line indicates the predicted support of the spectrum. We do indeed see that the estimated spectrum follows this diagonal line, and we see that this line ends at the spatial bandwidth of approximately  $\Omega_x$  rad/s, giving rise to a temporal bandwidth of  $\Gamma\Omega_x = 25\pi$  rad/s. Tests with other values of  $\Omega_x$  and  $\Gamma$  also behave as predicted.

Table 1: Video parameters used for experiments in several figures.

Test Video	Spatial BW $\Omega_x$	Motion BW $\Omega_h$	Max. Speed $\Gamma$	Max. Predicted Temporal BW $2\Omega_x\Gamma + 2\Omega_h$
Row 1	$\pi$ rad/pix	15 rad/s	25 pix/s	187 rad/s
Row 2	$3\pi$ rad/pix	15 rad/s	25 pix/s	501 rad/s
Row 3	$\pi$ rad/pix	45 rad/s	25 pix/s	247 rad/s
Row 4	$\pi$ rad/pix	15 rad/s	200 pix/s	1287 rad/s

### 2.4.2 Bounded velocity model for $h(t)$

In order to test our predictions in the case of bounded velocity, we set

$$h(t) = \sum_{i=1}^5 a_i \text{sinc} \left( \frac{\Omega_h(t - d_i)}{\pi} \right), \quad (6)$$

where  $\Omega_h$  controls the total bandwidth (and can be set as we desire), the delays  $d_i$  are chosen randomly, and the amplitudes  $a_i$  are chosen somewhat arbitrarily but ensure that the maximum value attained by  $|h(t)|$  equals some parameter  $\Gamma$ , which we can set as we desire. Note that we can thus independently articulate the bandwidth and the maximum slope of this signal. This “sum of sinc functions” model for  $h(t)$  was chosen to give this function an interesting shape while still allowing us to ensure that it is perfectly bandlimited. As in (5), we let  $g(x) = \text{sinc} \left( \frac{\Omega_x x}{\pi} \right)$  for some value of  $\Omega_x$  that we set as desired.

For several choices of our parameters  $(\Omega_x, \Omega_h, \Gamma)$ , Figure 3 shows the video  $f(x, t)$  along with its estimated spectrum. The blue lines in each case indicate the predicted “butterfly shape” which should bound the nonzero support of  $F(\omega_x, \omega_t)$ . Table 1 lists the values of our parameters  $(\Omega_x, \Omega_h, \Gamma)$  used in each row of the figure. In these experiments, we see that, as we vary the bandwidth and velocity parameters, the approximate support of the estimated spectrum stays within the “butterfly shape” predicted by our theory in Section 2.2.2.

As another way to support this theory, we have computed for each video the empirical temporal bandwidth based on our estimated spectrum. To do this, we determine the value of  $\Omega_t$  for which 99.99% of the energy in the FFT (or windowed FFT) falls within the range  $|\omega_t| \leq \Omega_t$ . For each of the four videos shown in Figure 3, we see that this empirical bandwidth  $\Omega_t$  equals roughly 48 to 52% of the bandwidth  $2\Omega_x\Gamma + 2\Omega_h$  predicted by our theory. (There are occasional exceptions where the unwindowed FFT gives a higher estimate, but this is likely due to sampling artifacts.) We note that this behavior is *consistent* with our theory, since our bandwidth prediction is merely an upper bound. Indeed, it is encouraging that such a high proportion of the empirical energy falls within our predicted region, since Carson’s bandwidth rule considers only up to about 98% of the signal power.

### 2.4.3 Sinusoidal model for $h(t)$

As one last interesting case that obeys all of our model assumptions, we let

$$h(t) = a \cos(\Omega_h t),$$

where  $\Omega_h$  controls the bandwidth (and can be set as we desire), and the amplitude  $a$  is chosen to ensure that the maximum value attained by  $|h(t)|$  equals some parameter  $\Gamma$ , which we can set as we desire. As in (5), we let  $g(x) = \text{sinc}\left(\frac{\Omega_x x}{\pi}\right)$  for some value of  $\Omega_x$  that we set as desired.

Figure 4 shows the corresponding videos and estimated spectra, where the four rows correspond to the same four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1. We see a distinct “striped” pattern in the estimated spectra; this is to be expected given the temporal periodicity of the video. (Sinusoidal functions  $h(t)$  are often chosen as a standard example in communications textbooks since the spectrum of  $e^{jh(t)}$  can be expressed analytically in terms of an impulse train with weights determined by Bessel functions [9].) Still, this signal obeys all of our model assumptions and its estimated spectrum generally does indeed stay within the “butterfly shape” predicted by our theory. Moreover, for each of the four videos shown in Figure 4, we see that the empirical bandwidth  $\Omega_t$  equals roughly 52 to 57% of the bandwidth  $2\Omega_x\Gamma + 2\Omega_h$  predicted by our theory.

## 2.5 Experiments beyond our model assumptions

Our formal analysis and the experiments in Section 2.4 pertain specifically to translational videos in which  $g(x)$  is bandlimited,  $h(t)$  has either constant velocity or bounded velocity and bandwidth, and the entire contents of the frame translate *en masse*. However, real world videos may contain objects whose appearance (neglecting translation) changes over time, objects that move in front of a stationary background, multiple moving objects, and so on. We suspect that as a general rule of thumb, the temporal bandwidth of real world videos will be dictated by the same tradeoffs of spatial resolution and object motion that our theory suggests. In particular, the prediction of  $2\Omega_x\Gamma + 2\Omega_h$  given by our theory may be approximately correct, if we let  $\Omega_x$  be the essential spatial bandwidth of the imaging system,  $\Gamma$  be the maximum speed of any object moving in the video, and  $\Omega_h$  be the essential bandwidth of any object motion. This last parameter is perhaps the most difficult to predict for a given video, but we suspect that in many cases its value will be small and thus its role minor in determining the overall temporal bandwidth.

To support these conjectures, we have identified the characteristic “butterfly” shape in the following experiments. However, it remains an open problem to back this up with theoretical analysis. We do note that we believe our results are largely consistent with the classical study by Dong and Atick concerning the statistics of real-world videos [10]. Although the videos in that study had relatively low temporal resolution, Dong and Atick did note a certain radial symmetry to the spectrum (with one term depending on  $\omega_t/\omega_x$ ) and observe that at low spatial frequencies the power spectrum will have a strong decay as a function of the temporal frequency.

### 2.5.1 Removing the bandlimitedness assumption from $g(x)$

We now consider a non-bandlimited profile function  $g(x)$  constructed by convolving a Gaussian bump having standard deviation  $\sigma$  (which we may set as desired) with the unit-step function  $u(x)$ . While  $g(x)$  is not strictly bandlimited, we may crudely approximate its essential bandwidth according to the formula  $\Omega_x \approx \frac{6}{\sigma}$ . For the moment, we will continue to use a bandlimited model for  $h(t)$ , choosing the “sum of sinc functions” as in (6), where we can specify the bandwidth as  $\Omega_h$  and the maximum value of  $|h(t)|$  as  $\Gamma$ .

Figure 5 shows the corresponding videos and estimated spectra, where the four rows correspond to the same four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1. Although the video is not strictly bandlimited spatially, we do see a reasonable decay of the spectrum in the vicinity of our predicted bandwidth  $\Omega_x$ . Aside from this distinction, the behavior follows our theory very closely,

and the estimated spectrum generally follows the predicted butterfly shape. For each of the four videos shown in Figure 5, we see that the empirical bandwidth  $\Omega_t$  equals roughly 5 to 20% of the bandwidth  $2\Omega_x\Gamma + 2\Omega_h$  predicted by our theory. These lower ratios compared to our experiment in Figure 3 likely result from the decay of the Gaussian spectrum (compared to the flat spectrum of a sinc function) and due to our conservative formula relating  $\Omega_x$  to  $\sigma$ .

As a second experiment, we consider a non-bandlimited profile function<sup>2</sup>

$$g(x) = \begin{cases} 0, & x \leq -\frac{\sigma}{2} \\ \frac{1}{2} + \frac{x}{\sigma}, & -\frac{\sigma}{2} < x < \frac{\sigma}{2} \\ 1, & x \geq \frac{\sigma}{2} \end{cases}$$

where in this case  $\sigma$  denotes the width of the linear transition region from 0 to 1 and can be set as we desire. This function has slower decay in the frequency domain, but we may crudely approximate its essential bandwidth according to the formula  $\Omega_x \approx \frac{12}{\sigma}$ . We again use the “sum of sinc functions” model for  $h(t)$ . Figure 6 shows the corresponding videos and estimated spectra, where the four rows correspond to the same four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1. We notice that while the basic behavior is the same as above, the slower decay of the spectrum in the spatial direction now gives rise to “longer wings” of the butterfly shape. Despite this, the estimated essential temporal bandwidths remain quite modest, typically with  $\Omega_t$  equal to roughly 9 to 26% of the bandwidth  $2\Omega_x\Gamma + 2\Omega_h$  predicted by our theory. However, one should take caution in interpreting this comparison given the crudeness of our relation between  $\Omega_x$  and  $\sigma$ .

### 2.5.2 Removing the bandlimitedness assumption from $h(t)$

We can also construct with videos for which the motion function  $h(t)$  is not bandlimited. For example, for each of the four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1, we construct videos of the form  $f(x, t) = g(x - h(t))$ , where  $g(x)$  is the non-bandlimited Gaussian-filtered step function with approximate bandwidth  $\Omega_x$  (as described in Section 2.5.1), and  $h(t)$  is a periodic triangle wave having fundamental frequency  $\Omega_h$  and maximum slope  $\Gamma$ . We emphasize that this triangle wave  $h(t)$  is *not* bandlimited, but for the sake of comparison with our other experiments, we set its fundamental frequency to be equal to the desired  $\Omega_h$ .

Figure 7 shows the corresponding videos and estimated spectra. We see a distinct “striped” pattern in the estimated spectra due to the temporal periodicity of the video (recall Section 2.4.3). Although the video is not strictly bandlimited spatially or temporally, we do see a reasonable decay of the spectrum in the vicinity of our predicted spatial and temporal bandwidths  $\Omega_x$  and  $2\Omega_x\Gamma + 2\Omega_h$ , respectively. (The decay is less strong in the temporal direction due to the stronger high-frequency content of  $h(t)$  compared to  $g(x)$ .) Overall, the behavior still generally follows our theory, and the estimated spectrum is somewhat concentrated within the predicted butterfly shape, though this is less true than in some of our other experiments due to the high-frequency content of  $h(t)$ . For each of the four videos shown in Figure 7, we see that the empirical bandwidth  $\Omega_t$  equals roughly 10 to 55% of the bandwidth  $2\Omega_x\Gamma + 2\Omega_h$  predicted by our theory.

### 2.5.3 Videos with multiple moving edges

We now extend our experiments to account for videos with multiple moving edges. We let  $g(x)$  be a Gaussian-filtered step function as described in Section 2.5.1, and we consider two moving edges

<sup>2</sup>It is debatable whether such a profile would be realistic for a physical imaging system.

with translation signals  $h(t)$  obeying the “sum of sinc functions” model. Thus, our setup is exactly as described in Section 2.5.1, except that we have two edges moving simultaneously, instead of one.

Figure 8 shows the corresponding videos and estimated spectra, where the four rows correspond to the same four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1. Despite the presence of multiple moving edges, the behavior of the estimated spectrum is virtually the same as that of the single-edge experiment in Figure 5, with much of the spectrum falling within the predicted butterfly-shaped region. Moreover, for each of the four videos shown in Figure 8, we see that the empirical bandwidth  $\Omega_t$  again equals roughly 5 to 20% of the prediction  $2\Omega_x\Gamma + 2\Omega_h$ .

#### 2.5.4 Videos with occlusions

To examine the effects of occlusions, we now consider videos constructed as follows. We construct  $h(t)$  using the “sum of sinc functions” as in (6), where we can specify the bandwidth as  $\Omega_h$  and the maximum value of  $|h(t)|$  as  $\Gamma$ . We then consider a translated unit-step function  $u(x - h(t))$  which moves in front of a stationary background consisting of a square wave taking values 0 and  $\frac{1}{2}$  and with period 8 pix. Wherever  $u(x - h(t)) = 1$ , the translated step occludes the background; wherever  $u(x - h(t)) = 0$ , the background is unoccluded. This “occluded background” signal is then filtered spatially with a Gaussian kernel having standard deviation  $\sigma$ . (We apply the filtering after occlusion, rather than vice versa, because this seems more plausible in a real imaging system.) For the purposes of our experiments, we estimate the spatial bandwidth of this filtered signal to be  $\Omega_x \approx \frac{6}{\sigma}$ .

Figure 9 shows the corresponding videos and estimated spectra, where the four rows correspond to the same four parameter sets for  $(\Omega_x, \Omega_h, \Gamma)$  specified in Table 1. Compared to Figure 5, we do see somewhat slower decay in the estimated spectrum along the temporal direction. Despite this, there does remain a distinct concentration of the spectrum in the predicted butterfly region, and for each of the four videos shown in Figure 9, the empirical bandwidth  $\Omega_t$  equals roughly 5 to 15% of the prediction  $2\Omega_x\Gamma + 2\Omega_h$ . It is interesting to note that the “temporal spreading” of the spectrum appears to be less pronounced than in the experiment of Section 2.5.2, in which  $h(t)$  was taken to be a triangle wave; one would expect both cases to suffer from the nondifferentiability of the video.

#### 2.5.5 Real world videos

We conclude this examination with a series of experiments on real-world videos. These videos (courtesy of MERL) were collected in a laboratory setting using a high-speed video camera, but the scenes being imaged contained natural (not particularly high-speed) motions. For each video, we select a 2D “slice” of the 3D video cube, extracting one spatial dimension and one temporal dimension.

We begin with the *Candle* video which features two candle flames in front of a dark background; the video was acquired at a rate of 1000 frames per second. We extract 512 pixels from row 300 of each video frame. Figure 10 shows the videos and estimated spectra, where from top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. In the estimated spectra, we recognize the approximate butterfly shape and an approximate limitation to the video bandwidth, both spatially and temporally. More specifically, we typically see a collection of lines with various slopes, with the lines passing roughly through the origin. However, there is also a bit of “thickness” near the origin due to a possible  $\Omega_h$  term. The slopes of these lines match what might be expected based on an empirical examination of the video itself. For example, during the first 512 frames

of the video, the candle flame moves relatively slowly. From the windowed FFT, we see that the slope of the butterfly wings equals approximately 150 pix/s, whereas from the video itself, we see that the maximum translational speed of the candle flame appears to equal roughly 100 pix/s. The ratio of 1.5 between these estimates is within the maximum ratio of 2.0 predicted by our theory. As we consider more frames of the candle video (2048 or more), we begin to see faster motion of the candle flames. This increases the slope of the butterfly wings as expected. For example, at roughly 2.6 seconds into the video, the candle flame appears to translate to the right with a speed of roughly 1500 pix/s, and consequently we see a portion of the estimated spectrum oriented along a line with slope of approximately 4000 pix/s. Overall, for the four videos shown in Figure 10, the empirical temporal bandwidth (the value of  $\Omega_t$  for which 99% of the energy in the windowed FFT falls within the range  $|\omega_t| \leq \Omega_t$ ) equals 50 rad/s, 117 rad/s, 348 rad/s, and 468 rad/s, respectively. This suggests that the video’s temporal sampling rate (1000 frames/sec) may have been higher than necessary.

Next, we consider the *Pendulum + Cars* video, featuring two translating cars and an occlusion as one car passes in front of the other; the video was acquired at a rate of 250 frames per second. We extract 640 pixels from row 390 of each video frame. Figure 11 shows the videos and estimated spectra, where from top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. Once again, in the estimated spectra, we recognize the approximate butterfly shape and an approximate limitation to the video bandwidth, both spatially and temporally, and once again, we typically see a collection of lines with various slopes, but now with a bit more “thickness” near the origin due to a possible  $\Omega_h$  term. The slopes of these lines match what might be expected based on an empirical examination of the video itself. For example, the maximum slope appears to be on the order of 140 pix/s, while the maximum translational speed of the cars appears to be on the order of 70 pix/s; however, a closer inspection of the video would be warranted to make this number more precise. The overall empirical temporal bandwidth is typically on the order of 25 to 35 rad/s, and consequently, this video’s temporal sampling rate (250 frames/sec) may also have been higher than necessary.

Finally, we consider the *Card + Monster* video, featuring a playing card translating in front of a dark background; the video was acquired at a rate of 250 frames per second. We extract 640 pixels from row 40 of each video frame. Figure 12 shows the videos and estimated spectra, where from top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. Our conclusions are similar to the experiments above, though we do see a more significant “thickness” in the butterfly shape for this video.

### 3 Videos with two spatial dimensions

#### 3.1 Problem setup

##### 3.1.1 Signal model

Our analysis is easily generalized to the more conventional case of videos having two spatial dimensions. We will again use the variable  $t \in \mathbb{R}$  to index time in seconds, and we will use the variables  $x, y \in \mathbb{R}$  to index spatial position in pix.

Let  $g(x, y)$  denote a 2D function of space (think of this as a continuous-space “still image”), and consider a continuous-space, continuous-time video  $f(x, y, t)$  in which each “frame” of the video



consists of a shifted version of this prototype frame. More formally, suppose that

$$f(x, y, t) = g(x - h_x(t), y - h_y(t))$$

where  $h(t) = (h_x(t), h_y(t))$  is a function that controls how much (in pix) the prototype frame is shifted in the  $x$ - and  $y$ -directions at each time step.

### 3.1.2 Fourier setup

Let  $F(\omega_x, \omega_y, \omega_t)$  denote the 3D Fourier transform of  $f(x, y, t)$ , and let  $G(\omega_x, \omega_y)$  denote the 2D Fourier transform of  $g(x, y)$ . Keeping track of units,  $\omega_x$  and  $\omega_y$  are measured in terms of rad/pix, and  $\omega_t$  is measured in terms of rad/s. The following relationships will be of use to us.

First, let  $\mathcal{F}_{x,y}\{\cdot\}$  denote an operator on  $L_2(\mathbb{R}^3)$  that performs the 2D Fourier transform in the spatial directions, and let  $\mathcal{F}_t\{\cdot\}$  denote an operator on  $L_2(\mathbb{R}^3)$  that performs the 1D Fourier transform in the temporal direction. Due to the separability of the 3D Fourier transform, we know that

$$F(\omega_x, \omega_y, \omega_t) = \mathcal{F}_{x,y}\{\mathcal{F}_t\{f\}\}(\omega_x, \omega_y, \omega_t) = \mathcal{F}_t\{\mathcal{F}_{x,y}\{f\}\}(\omega_x, \omega_y, \omega_t). \quad (7)$$

Now, due to the shift property of the 1D Fourier transform, we have

$$\mathcal{F}_{x,y}\{f\}(\omega_x, \omega_y, t) = \mathcal{F}_{x,y}\{g(x - h_x(t), y - h_y(t))\}(\omega_x, \omega_y, t) = G(\omega_x, \omega_y)e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}.$$

Again, this function is complex-valued and has constant magnitude in the temporal direction; therefore it has only phase changes in the temporal direction.

Following (7),  $F$  will be the result of taking the 1D Fourier transform of this function  $\mathcal{F}_{x,y}\{f\}$  in the temporal direction. We have

$$F(\omega_x, \omega_y, \omega_t) = \mathcal{F}_t\{G(\omega_x, \omega_y)e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}\}(\omega_x, \omega_y, \omega_t) = G(\omega_x, \omega_y) \cdot L(\omega_x, \omega_y, \omega_t),$$

where

$$L(\omega_x, \omega_y, \omega_t) := \mathcal{F}_t\{e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}\}(\omega_x, \omega_y, \omega_t).$$

For fixed  $\omega_x, \omega_y$ ,  $L(\omega_x, \omega_y, \omega_t)$  equals the 1D Fourier transform of  $e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}$  with respect to time, evaluated at the frequency  $\omega_t$ .

## 3.2 Temporal bandwidth analysis

### 3.2.1 Constant velocity model for $h_x(t)$ and $h_y(t)$

Suppose the translation has a constant velocity, i.e., that  $h_x(t) = \Gamma_x t$  and that  $h_y(t) = \Gamma_y t$  for some constants  $\Gamma_x, \Gamma_y$  (having units of pix/s). In this case we have  $L(\omega_x, \omega_y, \omega_t) = \delta(\omega_t - \omega_x \Gamma_x - \omega_y \Gamma_y)$ , and so

$$F(\omega_x, \omega_y, \omega_t) = G(\omega_x, \omega_y) \cdot \delta(\omega_t - \omega_x \Gamma_x - \omega_y \Gamma_y),$$

which corresponds to a diagonal line in the 3D Fourier plane.

To see the implications of this in terms of bandwidth, let us suppose that  $G(\omega_x, \omega_y)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $(\omega_x, \omega_y) \in [-\Omega_x, \Omega_x] \times [-\Omega_y, \Omega_y]$  rad/pix. In this case, it follows that  $F(\omega_x, \omega_y, \omega_t)$  must be bandlimited (or essentially bandlimited) to the range of frequencies  $(\omega_x, \omega_y, \omega_t) \in [-\Omega_x, \Omega_x] \times [-\Omega_y, \Omega_y] \times [-(\Gamma_x \Omega_x + \Gamma_y \Omega_y), \Gamma_x \Omega_x + \Gamma_y \Omega_y]$ . In other words, the temporal bandwidth of the video is no greater than  $\Gamma_x \Omega_x + \Gamma_y \Omega_y$  rad/s.

### 3.2.2 Bounded velocity model for $h_x(t)$ and $h_y(t)$

In slightly more generality, we could assume that the position functions  $h_x(t)$  and  $h_y(t)$  have bounded slope, i.e., that for some  $\Gamma_x, \Gamma_y > 0$ ,

$$\left| \frac{dh_x(t)}{dt} \right| \leq \Gamma_x \text{ pix/s} \quad \text{and} \quad \left| \frac{dh_y(t)}{dt} \right| \leq \Gamma_y \text{ pix/s}$$

for all  $t$ . This corresponds to a bound on the “speed” at which the object can move in the video. We also assume that both translation signals  $h_x(t)$  and  $h_y(t)$  have bandwidths bounded by  $\Omega_h$  rad/s. (This guarantees that for any fixed  $\omega_x$  and  $\omega_y$ , the bandwidth of  $\phi(t)$  below is also bounded by  $\Omega_h$  rad/s.)

For any fixed  $\omega_x$  and  $\omega_y$ , we can recognize  $e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}$  as a phase-modulated (PM) sinusoid having carrier frequency 0 and phase deviation (in radians)

$$\phi(t) = -\omega_x h_x(t) - \omega_y h_y(t),$$

or equivalently, as a frequency-modulated (FM) sinusoid having carrier frequency 0 and instantaneous frequency (in rad/s)

$$\omega_i(t) = \frac{d\phi(t)}{dt} = -\omega_x \frac{dh_x(t)}{dt} - \omega_y \frac{dh_y(t)}{dt}.$$

For an FM signal, we have

$$\frac{d\phi(t)}{dt} = \omega_d m(t),$$

where  $m(t)$  is known as the modulating signal and  $\omega_d$  is the frequency deviation constant. Thus, we have

$$m(t) = \frac{1}{\omega_d} \frac{d\phi(t)}{dt} = -\frac{\omega_x}{\omega_d} \frac{dh_x(t)}{dt} - \frac{\omega_y}{\omega_d} \frac{dh_y(t)}{dt}.$$

Let us also define the deviation term

$$D := \frac{\omega_d}{\Omega_h} \max |m(t)| \leq \frac{|\omega_x|}{\Omega_h} \max \left| \frac{dh_x(t)}{dt} \right| + \frac{|\omega_y|}{\Omega_h} \max \left| \frac{dh_y(t)}{dt} \right| \leq \frac{|\omega_x| \Gamma_x + |\omega_y| \Gamma_y}{\Omega_h}.$$

From Carson’s bandwidth rule for frequency modulation, we have that for fixed  $\omega_x$  and  $\omega_y$ , at least 98% of the total power of  $e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}$  must be concentrated in the frequency range  $\omega_t \in [-2(D+1)\Omega_h, 2(D+1)\Omega_h]$  rad/s. Since  $D \leq \frac{|\omega_x| \Gamma_x + |\omega_y| \Gamma_y}{\Omega_h}$ , we conclude that at least 98% of the total power of  $e^{-j\omega_x h_x(t) - j\omega_y h_y(t)}$  must be concentrated in the frequency range  $\omega_t \in [-(2|\omega_x| \Gamma_x + 2|\omega_y| \Gamma_y + 2\Omega_h), 2|\omega_x| \Gamma_x + 2|\omega_y| \Gamma_y + 2\Omega_h]$  rad/s. We note that the dependence of this bandwidth on  $\omega_x$  is essentially linear.

We conclude that  $L(\omega_x, \omega_y, \omega_t)$  will have a characteristic “polytope hourglass shape”. Considering the first octant of the 3D frequency space (in which  $\omega_x, \omega_y, \omega_t \geq 0$ ), most of the total power of  $L(\omega_x, \omega_y, \omega_t)$  will fall below (in the temporal direction) a plane passing through the points  $(0, 0, 2\Omega_h)$ ,  $(1, 0, 2\Gamma_x + 2\Omega_h)$ , and  $(0, 1, 2\Gamma_y + 2\Omega_h)$ . The other seven octants follow symmetrically.

To see the implications of this in terms of bandwidth, let us again suppose that  $G(\omega_x, \omega_y)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $(\omega_x, \omega_y) \in [-\Omega_x, \Omega_x] \times [-\Omega_y, \Omega_y]$ . We must then have that  $F(\omega_x, \omega_y, \omega_t) = G(\omega_x, \omega_y) \cdot L(\omega_x, \omega_y, \omega_t)$  is also essentially bandlimited in the spatial direction to the range of frequencies  $(\omega_x, \omega_y) \in [-\Omega_x, \Omega_x] \times [-\Omega_y, \Omega_y]$ . Because of the

hourglass structure in  $L(\omega_x, \omega_y, \omega_t)$ , however, this will also cause  $F(\omega_x, \omega_y, \omega_t)$  to be essentially bandlimited in the temporal direction to the range of frequencies

$$\omega_t \in [-(2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h), 2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h]. \quad (8)$$

Therefore, we see that filtering such a video in the spatial directions can cause it to be essentially bandlimited both in space and in time.

### 3.3 Sampling implications

Based on the temporal bandwidth predicted in (8), where we assume that  $G(\omega_x, \omega_y)$  is bandlimited (or essentially bandlimited) to the range of frequencies  $(\omega_x, \omega_y) \in [-\Omega_x, \Omega_x] \times [-\Omega_y, \Omega_y]$ , the Nyquist theorem suggests that in order to avoid aliasing, the video should be sampled at a minimum rate of

$$\frac{2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h}{\pi} \text{ samples/s.}$$

As a point of reference, during a sampling interval of  $\left(\frac{2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h}{\pi}\right)^{-1}$  seconds, an object moving with maximum speed of  $\Gamma_x$  pix/s in the  $x$ -direction can traverse a maximum of

$$\frac{\pi\Gamma_x}{2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h} \text{ pix}$$

in the  $x$ -direction. Similarly, an object moving with maximum speed of  $\Gamma_y$  pix/s in the  $y$ -direction can traverse a maximum of

$$\frac{\pi\Gamma_y}{2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h} \text{ pix.}$$

Using the triangle inequality for the sake of simplicity, we conclude that the object can move no more than

$$\frac{\pi(\Gamma_x + \Gamma_y)}{2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y + 2\Omega_h} \text{ pix} \quad (9)$$

in any direction.

Let us plug in some plausible numbers to illustrate the implications of these bounds. If we expect that both  $\Omega_x$  and  $\Omega_y$  will be on the order of  $\pi$  rad/pix, and if we assume that the  $2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y$  term will typically dominate the  $2\Omega_h$  term, then (9) suggests that, in typical scenarios, to avoid temporal aliasing we should not allow a moving object to traverse more than  $\approx \frac{1}{2}$  pix in any direction between adjacent sampling times.

Aside from exceptionally non-smooth motions  $h(t)$ , we do strongly suspect that the influence of the temporal bandwidth  $\Omega_h$  will be minor in comparison to the influence of the  $2\Omega_x\Gamma_x + 2\Omega_y\Gamma_y$  term, and so *in general a temporal Nyquist sampling rate of  $2(\Gamma_x + \Gamma_y)$  samples/s will likely serve as a reasonable rule of thumb.*<sup>3</sup> Again, this rule of thumb illustrates the direct relationship between the speed of object motion in the video and the video's overall temporal bandwidth.

### 3.4 Experiments

It would not be difficult to conduct experiments analogous to those presented in Sections 2.4 and 2.5. Due to time limitations, however, we have not included such experiments.

---

<sup>3</sup>It is possible that this could be reduced to  $2\sqrt{\Gamma_x^2 + \Gamma_y^2}$  through the natural modifications to our analysis.

## 4 Sampling and interpolation principles

The insight we have developed in the past two sections suggests that many videos of interest may indeed be exactly or approximately bandlimited in the temporal direction. For problems involving CS of such videos, this implies that there may be a limit to the “complexity” of the information collected by compressive measurement devices with high temporal sampling rates. One way of exploiting this limited complexity is in the context of classical interpolation identities for bandlimited signals. We briefly review these identities in this section before exploring their applications for CS reconstruction in Section 5.

### 4.1 Interpolation theory

Before considering 2D or 3D video signals, let us first review the basic principles involved in sampling, interpolation, and reconstruction of a bandlimited 1D signal. Suppose that  $f(t)$  is a signal with temporal bandwidth bounded by  $\Omega_t$  rad/s. The Nyquist theorem states that this signal can be reconstructed from a discrete set of samples  $\{f(nT_s)\}_{n \in \mathbb{Z}}$ , where the sampling interval  $T_s \leq \frac{\pi}{\Omega_t}$  seconds. In particular, it holds that

$$f(t) = \sum_{n \in \mathbb{Z}} f(nT_s) \operatorname{sinc} \left( \frac{t - nT_s}{T_s} \right), \quad (10)$$

where  $\operatorname{sinc}(t) = \frac{\sin(\pi t)}{\pi t}$ . Instead of actually *reconstructing* the continuous-time signal  $f(t)$ , a more important consequence of (10) for us will be the fact that, for any  $t_0 \in \mathbb{R}$ ,  $f(t_0)$  can be *represented* as a linear combination of the discrete samples  $\{f(nT_s)\}_{n \in \mathbb{Z}}$ .

With varying degrees of approximation, it is possible to replace the sinc interpolation kernel in (10) with other, more localized kernels. We will write

$$\tilde{f}(t) = \sum_{n \in \mathbb{Z}} f(nT_s) \gamma \left( \frac{t - nT_s}{T_s} \right), \quad (11)$$

where  $\gamma(t)$  is a prototype interpolation kernel. In addition to the sinc kernel, for which

$$\gamma(t) = \operatorname{sinc}(t),$$

other possible choices include the zero-order hold (rectangular) kernel, for which

$$\gamma(t) = \operatorname{rect}(t) = \begin{cases} 1, & |t| \leq \frac{1}{2} \\ 0, & \text{otherwise,} \end{cases}$$

the first-order hold (triangular, or “linear interpolation”) kernel, for which

$$\gamma(t) = \operatorname{tri}(t) = \begin{cases} 1 - |t|, & |t| \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

and a variety of cubic interpolation kernels [11].

In general, a smoother choice for  $\gamma(t)$  will better approximate the ideal sinc kernel.<sup>4</sup> However, smoother kernels tend to have wider temporal supports, and it can be desirable in some applications

---

<sup>4</sup>The implications of this can be appreciated in the frequency domain. Since one can view the reconstruction process as convolution of an impulse train with the scaled interpolation kernel  $\gamma\left(\frac{t}{T_s}\right)$ , in the frequency domain we are windowing the periodic spectrum of  $f(t)$  with the frequency response of the interpolation kernel. Smoother interpolation kernels better approximate the ideal lowpass filter.

(such as the CS recovery problem discussed below) to limit the temporal support of the kernel. One way to improve the performance of the lower-order, more narrow interpolation kernels is to decrease the sampling interval  $T_s$ .<sup>5</sup> However, for our CS recovery problem discussed below, this too will increase the complexity of the recovery algorithm by increasing the number of “anchor frames”.

For a 2D or 3D video with limited temporal bandwidth, the separability of the Fourier transform implies that the interpolation formulas presented above should hold for each spatial location (i.e., for each pixel). For several of the approximately bandlimited test videos considered in Sections 2.4 and 2.5, we explore below the quality of these interpolation formulas as a function of the temporal sampling rate and the interpolation kernel type.

## 4.2 Experiments within our model assumptions

First, we recall Section 2.4.2 and consider videos  $f(x, t) = g(x - h(t))$  with bandlimited sinc profile  $g(x)$  and bandlimited “sum of sinc functions” model for the translation signal  $h(t)$ . In the right side Figure 13(a), we show a video with parameters  $(\Omega_x, \Omega_h, \Gamma)$  drawn from the fourth row of Table 1; the time axis runs vertically. We set  $T_s$  equal to 0.5 times the Nyquist sampling interval corresponding to the predicted temporal bandwidth of  $2\Omega_x\Gamma + 2\Omega_h$  rad/s. In the left side of Figure 13(b), using a linear interpolation kernel  $\gamma(t)$ , we plot as a function of time  $t$  the maximum interpolation error

$$\max_x |f(x, t) - \tilde{f}(x, t)|$$

over all  $x$  in our sampling grid; the times  $t$  are selected from a fine grid, and again the time axis runs vertically and is presented for comparison with the video on the right. We see that interpolation errors are greatest for times in the video at which  $h(t)$  has large slope.

For each of the four parameter sets  $(\Omega_x, \Omega_h, \Gamma)$  listed in Table 1, we also compute the maximum interpolation error

$$\max_{x,t} |f(x, t) - \tilde{f}(x, t)|$$

as a function of the temporal sampling interval  $T_s$ . Using a linear interpolation kernel for  $\gamma(t)$ , Figure 13(b) plots the maximum interpolation error for each type of video, where the colored lines represent each of the four video types (blue = row 1 in Table 1, green = row 2, red = row 3, and cyan = row 4), and the horizontal axis represents the ratio of  $T_s$  to the Nyquist sampling interval corresponding to the predicted temporal bandwidth of  $2\Omega_x\Gamma + 2\Omega_h$  for each video. Figure 13(c) repeats this experiment using a cubic interpolation kernel  $\gamma(t)$ . In each case, we see that the sampling near the predicted Nyquist rate allows interpolation of the missing video frames with reasonable accuracy. Also, the similar nature of the four curves suggests that the actual values of  $(\Omega_x, \Omega_h, \Gamma)$  are not critical in determining interpolation accuracy; what seems to be important is how these values combine into a predicted Nyquist sample rate, and how fast the video is sampled compared to this rate.

---

<sup>5</sup>Doing so provides more space in the frequency domain between the replicated copies of the spectrum of  $f(t)$ , permitting nonideal lowpass filters a better opportunity to select only the baseband copy and to do so with minimal distortion.

### 4.3 Experiments beyond our model assumptions

#### 4.3.1 Removing the bandlimitedness assumption from $g(x)$

In Figure 14 we repeat all of the above experiments for the type of videos examined in Section 2.5.1, with bandlimited sinc profile  $g(x)$  and bandlimited “sum of sinc functions” model for the translation signal  $h(t)$ . From panel (a), we see that interpolation errors are again typically highest at times for which  $h(t)$  has large slope. From panels (b) and (c), however, we see that compared to Figure 13, interpolation errors for these videos are relatively lower for a given sampling rate (as a fraction of the predicted Nyquist rate). As discussed in Section 2.5.1, this is likely due to the decay of the Gaussian spectrum (compared to the flat spectrum of a sinc function) and due to our conservative formula relating  $\Omega_x$  to  $\sigma$ .

#### 4.3.2 Removing the bandlimitedness assumption from $h(t)$

In Figure 15 we repeat all of these experiments for the type of videos examined in Section 2.5.2, with non-bandlimited Gaussian-filtered step function  $g(x)$  and non-bandlimited triangle wave model for  $h(t)$ . From panel (a), we see that interpolation errors are typically highest at times where  $h(t)$  has an abrupt change of slope (at the peaks of the triangles). Despite the non-bandlimited nature of  $h(t)$  we do see from panels (b) and (c) that reasonable interpolation quality is achievable as long as the video is sampled near the predicted (approximate) Nyquist rate.

#### 4.3.3 Videos with occlusions

In Figure 16 we repeat all of these experiments for the type of videos examined in Section 2.5.4, with  $f(x, t)$  containing a moving edge occluding a stationary background pattern. Similar conclusions hold.

#### 4.3.4 Real world videos

Finally, Figure 17, Figure 18, and Figure 19 present similar experiments conducted on the *Candle*, *Pendulum + Cars*, and *Card + Monster* videos, respectively. The experiments are conducted by subsampling the original high-rate videos and attempting to interpolate the values of the omitted frames. Panel (a) of each figure illustrates, for each  $t$ , the maximum interpolation error over all  $x$ . We see that moments of high speed motion present the most difficulty for accurate interpolation. Panel (b) of each figure plots the maximum interpolation error over all  $x$  and  $t$  for both linear interpolation (solid blue line) and cubic interpolation (dashed red line) as a function of the absolute number of samples retained per second. It is clear that these videos contain some features beyond our model assumptions, which prevents the maximum interpolation error from achieving a small value. However, we do see characteristic improvements as we increase our sample rate.

### 4.4 A manifold-based interpretation

As a brief note, we mention that one could reinterpret these tradeoffs of bandwidth vs. interpolation quality in the context of manifolds. In past work [12], we have shown that typical manifolds of images containing sharp edges are non-differentiable. However, these manifolds contain a multiscale structure that can be accessed by regularizing the images. The more one smoothes the images, the smoother the manifold will become, and local tangent approximations will remain accurate over longer distances across the manifold.

In the context of this document, the spatial profile function  $g(x)$  is often assumed to be bandlimited due to an inherent resolution limit in the imaging system. Thus, our translational video model  $f(x, t) = g(x - h(t))$  corresponds to a “walk” along a manifold consisting of blurred images of the same object, but translated to different positions. From the manifold perspective, a decrease in the bandwidth of  $g(x)$  corresponds to an increase in the amount of blurring, which smooths the manifold. From the perspective of this document, a decrease in the bandwidth of  $g(x)$  decreases the anticipated temporal bandwidth of the video, which improves the quality of linear (and other) interpolation schemes. The reason for the improved interpolation accuracy can be appreciated by considering the increased smoothness of the manifold; in particular, linear interpolation improves as the tangent spaces twist more slowly.

These intuitive connections could be explored more formally in future work. For example, the performance of higher order interpolation methods (such as cubic interpolation) may be tied to the rate at which higher order derivatives change along the manifold.

## 5 Reconstruction of temporally bandlimited videos from streaming measurements

In this section, we will consider how the Compressive Sensing (CS) reconstruction problem can be formulated in streaming scenarios, where we acquire one measurement of a video per time instant.

### 5.1 Measurement process

Consider a continuous-space, continuous-time video  $f(x, t)$  that has temporal bandwidth bounded by  $\Omega_t$  rad/s. Let  $f_d : \{1, 2, \dots, N_x\} \times \mathbb{R} \rightarrow \mathbb{R}$  denote a sampled discrete-space, continuous-time version of this video, where for  $p = 1, 2, \dots, N_x$ ,

$$f_d(p, t) = f(p\Delta_x, t). \quad (12)$$

In the expression above,  $\Delta_x$  represents the spatial sampling resolution (in pix). We note that the temporal bandwidth of  $f_d$  will still be bounded by  $\Omega_t$  rad/s.

Now, let  $T_1$  denote a measurement interval (in seconds); typically  $T_1$  will be much smaller than the Nyquist sampling interval of  $\frac{\pi}{\Omega_t}$  suggested by the video’s bandwidth. Suppose that one linear measurement is collected from  $f_d$  every  $T_1$  seconds. Letting  $y(m)$  denote the measurement collected at time  $mT_1$ , we can write

$$y(m) = \sum_{p=1}^{N_x} \phi_m(p) f_d(p, mT_1) = \langle \phi_m, f_d(:, mT_1) \rangle, \quad (13)$$

where  $\phi_m \in \mathbb{R}^{N_x}$  is a vector of random numbers, and we use “Matlab notation” to refer to a vector  $f_d(:, t) \in \mathbb{R}^{N_x}$ .

Stacking all of the measurements, we have

$$y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix} = \begin{bmatrix} \langle \phi_1, f_d(:, T_1) \rangle \\ \langle \phi_2, f_d(:, 2T_1) \rangle \\ \vdots \\ \langle \phi_M, f_d(:, MT_1) \rangle \end{bmatrix} = \underbrace{\begin{bmatrix} \phi_1^T & & & \\ & \phi_2^T & & \\ & & \ddots & \\ & & & \phi_M^T \end{bmatrix}}_{M \times MN_x} \underbrace{\begin{bmatrix} f_d(:, T_1) \\ f_d(:, 2T_1) \\ \vdots \\ f_d(:, MT_1) \end{bmatrix}}_{MN_x \times 1} \quad (14)$$

Unfortunately, there are two difficulties with attempting to use this formulation for a CS recovery: first, the recovery problem is very highly underdetermined, with the number of measurements representing only  $\frac{1}{N_x}$  times the number of unknowns; and second, the size of the recovery problem, with  $MN_x$  unknowns, can be immense. (Both of these difficulties will be even more significant for videos with two spatial dimensions.)

## 5.2 Simplifying the linear equations

Fortunately, the bandlimitedness of the video allows us to simplify this recovery process somewhat. Let  $T_s$  denote a sampling interval no greater than the Nyquist limit ( $\frac{\pi}{\Omega_t}$  seconds) suggested by the video's bandwidth, and assume that  $T_s = VT_1$  for some integer  $V \geq 1$ . Then for any integer  $j$ , we can write

$$\begin{aligned}
f_d(:, jT_1) &= \sum_{n \in \mathbb{Z}} f_d(:, nT_s) \gamma \left( \frac{jT_1 - nT_s}{T_s} \right) \\
&= \sum_{n \in \mathbb{Z}} f_d(:, nT_s) \gamma \left( \frac{j\frac{T_s}{V} - nT_s}{T_s} \right) \\
&= \sum_{n \in \mathbb{Z}} f_d(:, nT_s) \gamma \left( \frac{j}{V} - n \right) \\
&= \left[ \cdots \gamma \left( \frac{j}{V} - 1 \right) I_{N_x} \quad \gamma \left( \frac{j}{V} - 2 \right) I_{N_x} \quad \gamma \left( \frac{j}{V} - 3 \right) I_{N_x} \quad \cdots \right] \begin{bmatrix} \vdots \\ f_d(:, T_s) \\ f_d(:, 2T_s) \\ f_d(:, 3T_s) \\ \vdots \end{bmatrix},
\end{aligned}$$

where  $I_{N_x}$  denotes the  $N_x \times N_x$  identity matrix. Therefore,

$$\begin{bmatrix} f_d(:, T_1) \\ f_d(:, 2T_1) \\ \vdots \\ f_d(:, MT_1) \end{bmatrix} = \begin{bmatrix} \cdots & \gamma \left( \frac{1}{V} - 1 \right) I_{N_x} & \gamma \left( \frac{1}{V} - 2 \right) I_{N_x} & \gamma \left( \frac{1}{V} - 3 \right) I_{N_x} & \cdots \\ \cdots & \gamma \left( \frac{2}{V} - 1 \right) I_{N_x} & \gamma \left( \frac{2}{V} - 2 \right) I_{N_x} & \gamma \left( \frac{2}{V} - 3 \right) I_{N_x} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \gamma \left( \frac{M}{V} - 1 \right) I_{N_x} & \gamma \left( \frac{M}{V} - 2 \right) I_{N_x} & \gamma \left( \frac{M}{V} - 3 \right) I_{N_x} & \cdots \end{bmatrix} \begin{bmatrix} \vdots \\ f_d(:, T_s) \\ f_d(:, 2T_s) \\ f_d(:, 3T_s) \\ \vdots \end{bmatrix} \quad (15)$$

Assuming  $\gamma(t)$  has temporal support within some reasonable bound, the matrix above will have size  $MN_x \times (\frac{MN_x}{V} + O(1))$  and so this allows a dimensionality reduction by a factor of  $V$ .

Putting all of this together, we have

$$y = \underbrace{\begin{bmatrix} \phi_1^T & & & \\ & \phi_2^T & & \\ & & \ddots & \\ & & & \phi_M^T \end{bmatrix}}_{M \times MN_x} \underbrace{\begin{bmatrix} f_d(:, T_1) \\ f_d(:, 2T_1) \\ \vdots \\ f_d(:, MT_1) \end{bmatrix}}_{MN_x \times 1}$$



$$\begin{aligned}
&= \underbrace{\begin{bmatrix} \phi_1^T & & & \\ & \phi_2^T & & \\ & & \ddots & \\ & & & \phi_M^T \end{bmatrix}}_{M \times MN_x} \underbrace{\begin{bmatrix} \cdots & \gamma \left(\frac{1}{V} - 1\right) I_{N_x} & \gamma \left(\frac{1}{V} - 2\right) I_{N_x} & \gamma \left(\frac{1}{V} - 3\right) I_{N_x} & \cdots \\ \cdots & \gamma \left(\frac{2}{V} - 1\right) I_{N_x} & \gamma \left(\frac{2}{V} - 2\right) I_{N_x} & \gamma \left(\frac{2}{V} - 3\right) I_{N_x} & \cdots \\ & & \vdots & \vdots & \\ \cdots & \gamma \left(\frac{M}{V} - 1\right) I_{N_x} & \gamma \left(\frac{M}{V} - 2\right) I_{N_x} & \gamma \left(\frac{M}{V} - 3\right) I_{N_x} & \cdots \end{bmatrix}}_{MN_x \times \left(\frac{MN_x}{V} + O(1)\right)} \underbrace{\begin{bmatrix} \vdots \\ f_d(:, T_s) \\ f_d(:, 2T_s) \\ f_d(:, 3T_s) \\ \vdots \end{bmatrix}}_{\left(\frac{MN_x}{V} + O(1)\right) \times 1} \\
&= \underbrace{\begin{bmatrix} \cdots & \gamma \left(\frac{1}{V} - 1\right) \phi_1^T & \gamma \left(\frac{1}{V} - 2\right) \phi_1^T & \gamma \left(\frac{1}{V} - 3\right) \phi_1^T & \cdots \\ \cdots & \gamma \left(\frac{2}{V} - 1\right) \phi_2^T & \gamma \left(\frac{2}{V} - 2\right) \phi_2^T & \gamma \left(\frac{2}{V} - 3\right) \phi_2^T & \cdots \\ & & \vdots & \vdots & \\ \cdots & \gamma \left(\frac{M}{V} - 1\right) \phi_M^T & \gamma \left(\frac{M}{V} - 2\right) \phi_M^T & \gamma \left(\frac{M}{V} - 3\right) \phi_M^T & \cdots \end{bmatrix}}_{M \times \left(\frac{MN_x}{V} + O(1)\right)} \underbrace{\begin{bmatrix} \vdots \\ f_d(:, T_s) \\ f_d(:, 2T_s) \\ f_d(:, 3T_s) \\ \vdots \end{bmatrix}}_{\left(\frac{MN_x}{V} + O(1)\right) \times 1} \quad (16)
\end{aligned}$$

Now we see that we have reduced the total number of unknowns, from  $MN_x$  to  $\frac{MN_x}{V} + O(1)$ . Indeed, the largest dimension of any matrix or vector in this formulation is now limited to  $\frac{MN_x}{V} + O(1)$  instead of  $MN_x$ . Moreover, due to decay in  $\gamma$ , the matrix above will be banded, with zeros in all positions sufficiently far from the diagonal. This facilitates storage of the matrix and possible streaming reconstruction.

From the formulation above, we see that it is possible to focus the reconstruction process on recovery of a relatively low-rate stream of “anchor frames”, rather than the high rate stream of frames measured by the imaging system. If the anchor frames are defined at the video’s temporal Nyquist rate, and if there are no additional assumptions made about the video, then one should not expect any temporal correlations to remain among the anchor frames. In many real world settings, however, there will be objects moving within the scene, and the smoothness of the object motion can lead to temporal correlations, e.g., that can be captured via motion compensated transforms. Thus, in order to impose the strongest possible model on the vector of anchor frames, it may be necessary to look for sparsity in a motion compensated wavelet transform, to invoke a dynamical system model for the changing anchor frames, or to find some other model for exploiting temporal correlations. In many applications, this may indeed be the only way to reduce the number of unknowns to some quantity smaller than  $M$  and thus to truly be able to solve the CS inverse problem.

## 5.3 Experiments

### 5.3.1 Single moving pulse

Let us illustrate the basic idea through the following experiment. We consider a video  $f(x, t) = g(x - h(t))$  having one spatial dimension. The profile  $g(x)$  is taken to be a Gaussian pulse having standard deviation  $\sigma = 0.5$ . Unlike earlier experiments in this document, we do *not* convolve the Gaussian with a unit step function for this experiment. We may crudely approximate the essential bandwidth of  $g(x)$  according to the formula  $\Omega_x \approx \frac{6}{\sigma} = 12$  rad/pix. For the translational signal  $h(t)$  we use the “sum of sinc functions” model (6), with  $\Omega_h = 8$  rad/s and  $\Gamma = 50$  pix/s. This video has a predicted temporal bandwidth of  $2\Omega_x\Gamma + 2\Omega_h = 1216$  rad/s, and so Nyquist suggests a minimum temporal sampling rate of approximately 387 samples/sec. We discretize this video in space, taking

$N_x = 50$  pixels/frame at a spatial sampling resolution of  $\Delta_x = 1$  pix/pixel. Following (12), we let  $f_d(p, t)$  denote the discretized video; the video is shown in Figure 20(a).

We then set  $T_1 = \frac{1}{2 \cdot 387} = \frac{1}{774}$  seconds, and letting  $m$  range from 1 to 774 (i.e.,  $t$  ranges from 0 to 1 sec), we collect a total of 774 random projections of this video, one at a time in equispaced intervals, as dictated by (13). To be clear, for each  $m$ , the measurement we collect at time  $mT_1$  represents the inner product of  $f_d(:, mT_1)$  against a length-50 random vector. In Figure 20(b), we follow (14) and employ a standard CS reconstruction algorithm<sup>6</sup> to reconstruct  $f_d$  at all of the time instants  $\{mT_1\}_{m=1}^{774}$ . In total, we are solving for  $50 \cdot 774 = 38700$  unknowns using just 774 random measurements, and since the sparsity level of this video in the canonical basis is clearly at least 774 (at least one pixel is nonzero at every time instant), there is no hope for a decent reconstruction.

For various values of  $V$ , then, we consider the formulation (16), using a linear interpolation kernel  $\gamma(t)$ , and reconstruct only the anchor frames. Since the anchor frames are spaced approximately  $\frac{V}{774}$  seconds apart, the number of unknowns we are solving for reduces to approximately  $\frac{38700}{V}$ ; we are attempting solve for these unknowns using the same 774 measurements. After solving for the anchor frames, we employ (15) to recover an estimate for  $f_d$  at all of the time instants  $\{mT_1\}_{m=1}^{774}$ . Figures 20(c) through 20(g) show the resulting reconstructions for  $V = 8$ ,  $V = 12$ ,  $V = 16$ ,  $V = 20$ , and  $V = 24$ , respectively. Figure 20(h) plots the MSE of the discretized frames  $f_d(:, t)$  as a function of time, for each of the various values of  $V$ .

Across a wide range of  $V$ , we see in these experiments that the reconstruction error is significantly improved compared to the full-scale reconstruction suggested by (14). However, we do see that by choosing  $V$  too low, there is a degradation in performance, which is presumably due to the fact that we must solve for too many unknowns given the available measurements. We also see that by choosing  $V$  too large, there is a degradation in performance, which is presumably due to the fact that the interpolation equation (15) on which we base our technique breaks down. For a fixed  $V$ , we also see that the reconstruction error increases as the slope of  $h(t)$  increases, which is to be expected and is again due to the fact that (15) is breaking down.

We can also examine the role played by the choice of interpolation kernel. In Figure 21 we repeat the above experiment, using a rectangular nearest neighbor interpolation kernel  $\gamma(t)$  instead of a linear interpolation kernel. With this choice of kernel, the accuracy of (15) is diminished, and the reconstruction performance suffers as one would expect. By comparing Figure 21 to Figure 20 we see the true potential value in our kernel-based approach, as Figure 21 could be interpreted as a simple “clustering” of the measurements in time, which is a natural approach that comes to mind when one considers streaming measurements from an architecture such as the Rice single pixel camera.

### 5.3.2 Two moving pulses

To test the performance of these techniques when the complexity of the video is increased, we repeat the above experiment (using a linear interpolation kernel) for a video containing two moving pulses. The results are shown in Figure 22. It is clear that the sparsity level of this video in the canonical basis is roughly twice that of the video from our initial experiment above. Consequently, the reconstruction performance is diminished in all cases. All of our relative statements about the role of  $V$  and the implications of the slope of  $h(t)$  carry over, however.

---

<sup>6</sup>This video was chosen due to the fact that it is sparse in the canonical basis; in practice other sparse bases could be employed for reconstruction.

### 5.3.3 Real world video

We have also experimented with the *Candle* video discussed previously; see Figure 23(a). In this case, we select  $N_x = 225$  pixels from row 300 of the video. We consider 400 adjacent frames of the video, which were collected at a sample rate of 1000 frames/sec. To construct streaming random measurements of this video, we actually compute 5 random measurements of each recorded frame;<sup>7</sup> thus the total number of random measurements collected is  $5 \cdot 400 = 2000$ .

In Figure 23(b), we adapt (14) to account for the fact that 5 (rather than 1) measurements are collected at each time instant and employ a standard CS reconstruction algorithm<sup>8</sup> to reconstruct the 400 frames. In total, we are solving for  $225 \cdot 400 = 90000$  unknowns using just 2000 random measurements. Since the sparsity level of this video exceeds 10000, there is no hope for a decent reconstruction.

For various values of  $V$ , then, we adapt (16) as appropriate and, using a linear interpolation kernel  $\gamma(t)$ , reconstruct only the anchor frames. In each case, the number of unknowns we are solving for reduces to approximately  $\frac{90000}{V}$ ; we are attempting solve for these unknowns using the same 2000 measurements. After solving for the anchor frames, we employ (15) to recover an estimate for all 400 original frames. Figures 23(c) through 23(g) show the resulting reconstructions for  $V = 8$ ,  $V = 12$ ,  $V = 20$ ,  $V = 30$ , and  $V = 40$ , respectively. Figure 23(h) plots the MSE of the reconstructed frames as a function of time, for each of the various values of  $V$ . Overall, these experiments illustrate the same issues concerning the role of  $V$  and speed of the motion in the video.

## 5.4 Toward a general purpose reconstruction algorithm

The preceding experiments provide a promising proof of concept for interpolation-based simplifications of the CS recovery problem. There are many directions in which this basic idea may be extended in developing a more general purpose CS reconstruction algorithm. Several such directions are briefly surveyed below.

### 5.4.1 Streaming reconstruction

In practice, we may be presented with a long stream of measurements and wish to reconstruct a video over a long time span. While it would be possible in theory to exploit the relationship (16) over the whole video at once, it would also be possible to window the measurements into smaller segments and reconstruct the segments individually using (16). If the segments are chosen to overlap slightly, then the final anchor frames in one segment could be used as a prior to inform the reconstruction of the initial anchor frames in the next segment. Or, when a start/end anchor frame is not available for properly initializing a reconstruction based on (16), certain interference cancellation schemes [13] could be exploited to account for the fact that those anchor frames may not be properly recoverable.

---

<sup>7</sup>Roughly speaking, this would correspond to the operation of a “five pixel” camera. Though such an architecture is perhaps not realistic in practice, we needed some way to construct more than 400 measurements for the 400 frames of interest.

<sup>8</sup>Again, for simplicity this video was chosen due to its sparsity in the canonical basis. However, reconstruction of this and other videos could likely be improved through the use of sparse transforms that exploit spatial correlations.

### 5.4.2 Optimizing anchor spacing

In streaming or non-streaming reconstructions, one would have the option of changing the anchor spacing (parameterized by  $V$ ) adaptively throughout the video, perhaps based on real-time estimates of the speed of object motion in the video. Our plots in Figures 20(h), 21(h), 22(h), and 23(h), however, suggest that the favorable choices of  $V$  do not change significantly over time. A deeper study of real world videos would be required to characterize the potential gains of changing  $V$  on the fly.

### 5.4.3 Optimizing the spatial resolution of the measurements

One could also imagine a compressive imaging architecture that, in times of fast object motion,<sup>9</sup> adaptively reduces the spatial resolution of the measurement vectors in order to artificially reduce the spatial (and thus also the temporal) bandwidth of the video. The goal of throttling back the measurement resolution in such a manner would be to better ensure that the identity (15) would approximately hold. To test the viability of such an approach, we repeat in Figure 24 the same experiment presented in Figure 20, but where each measurement vector has been lowpass filtered and downsampled by a factor of 2. Similarly, Figure 25 shows the results for a video with two moving pulses; it should be compared with Figure 22.

In general, we see that reducing the measurement resolution reduces the spatial resolution of the reconstructed frames. Although we rarely see an overall gain in terms of mean square error, there can occasionally be clear visual improvements in the lower resolution reconstructions. A deeper study of real world videos would be required to better characterize the potential gains of changing the measurement resolution on the fly. It would also be useful to consider whether it is possible to adapt just the reconstruction resolution without changing the actual measurement resolution.

### 5.4.4 Exploiting residual correlations among anchor frames

In the hypothetical case where no residual correlations remain among the anchor frames, there would be little room for improvement upon this reconstruction approach. In real-world videos, however, we do expect that significant correlations may remain. The question of how to exploit these correlations has been addressed in the more conventional CS recovery literature. For example, one could employ a linear dynamical system model to capture the restricted degrees of freedom of the anchor frames [6]. Alternatively, we are currently developing an algorithm that will involve estimating the motion of the video and then formulating the recovery of the anchor frames in a motion-compensated sparse basis [7]; the details of this algorithm will be reported in a future publication. As better solutions continue to be developed for exploiting such correlations, we expect that they could be combined with the methods discussed in this paper to permit better reconstruction from streaming or “single-pixel” measurements. It may also be possible to combine the insight gained from this paper with techniques for “manifold lifting” [14], building on the connections discussed briefly in Section 4.4.

---

<sup>9</sup>The video could perhaps be probed periodically with repeated pairs of measurement functions to detect object speeds.

## Acknowledgments

Thanks to Jae Young Park for input on the technical analysis in the early stages of this research and for discussions concerning the broader applicability of these ideas in a motion-compensated reconstruction algorithm. Thanks also to Aswin Sankaranarayanan and Richard Baraniuk for helpful discussions during the collaborative ONR project that supported this research. Thanks to Amit Ashok for providing a helpful reference, and thanks to MERL for providing the real-world videos used in these experiments.

## References

- [1] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, “Compressive imaging for video representation and coding,” (in *Picture Coding Symposium - PCS 2006*, Beijing, China), April 2006.
- [2] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, “Single-pixel imaging via compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 83–91, 2008.
- [3] M. Do, D. Marchand-Maillet, and M. Vetterli, “On the bandwidth of the plenoptic function,” *IEEE Transactions on Image Processing*, vol. 21, Feb. 2012.
- [4] T. Ajdler, L. Sbaiz, and M. Vetterli, “The plenacoustic function and its sampling,” *IEEE Transactions on Signal Processing*, vol. 54, no. 10, pp. 3790–3804, 2006.
- [5] T. Ajdler, L. Sbaiz, and M. Vetterli, “Dynamic measurement of room impulse responses using a moving microphone,” *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1636–1645, 2007.
- [6] A. C. Sankaranarayanan, P. K. Turaga, R. G. Baraniuk, and R. Chellappa, “Compressive acquisition of dynamic scenes,” in *European Conf. on Computer Vision*, Sept. 2010.
- [7] J. Y. Park and M. B. Wakin, “A multiscale framework for compressive sensing of video,” in *Proc. Picture Coding Symp. (PCS)*, 2009.
- [8] Y. Wang, J. Ostermann, and Y. Q. Zhang, *Video processing and communications*. Prentice Hall, 2002.
- [9] R. E. Ziemer and W. H. Tranter, *Principles of Communication: Systems, Modulation and Noise*.
- [10] D. W. Dong and J. J. Atick, “Statistics of natural time-varying images,” *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 345–358, 1995.
- [11] E. Meijering and M. Unser, “A note on cubic convolution interpolation,” *IEEE Transactions on Image Processing*, vol. 12, pp. 477–479, April 2003.
- [12] M. B. Wakin, D. Donoho, H. Choi, and R. G. Baraniuk, “The multiscale structure of non-differentiable image manifolds,” in *Proc. Wavelets XI at SPIE Optics and Photonics*, August 2005.
- [13] M. Davenport, P. Boufounos, and R. Baraniuk, “Compressive domain interference cancellation,” in *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, (Saint-Malo, France), April 2009.
- [14] M. B. Wakin, “A manifold lifting algorithm for multi-view compressive imaging,” in *Proc. Picture Coding Symp. (PCS)*, 2009.

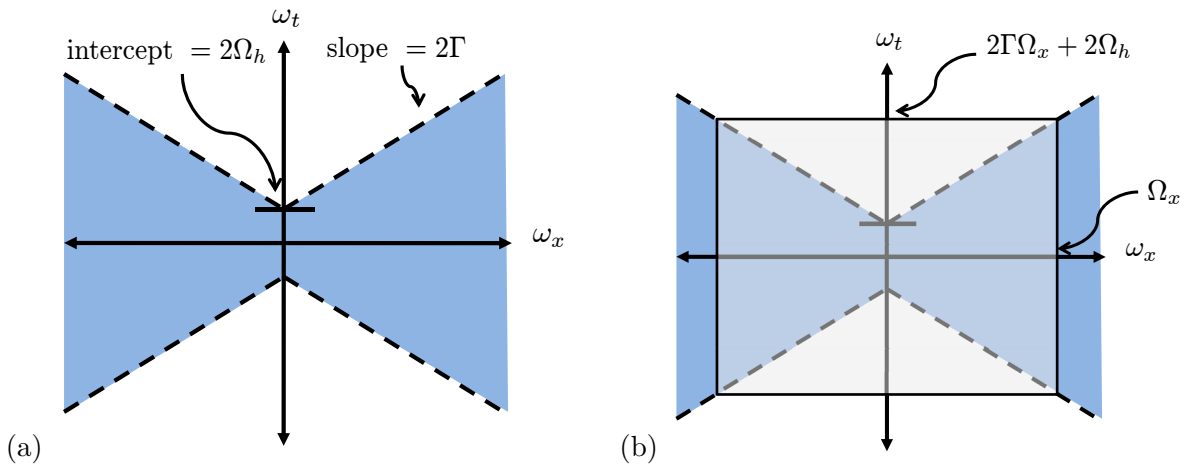


Figure 1: (a) Butterfly structure in the two-dimensional Fourier transform of a simple translational 1D video. The slope of the lines is proportional to the maximum speed of the translation. (b) Bandlimiting the video in space (e.g., by spatial lowpass filtering) will also bandlimit the video in time. The resulting temporal bandwidth will be proportional to the spatial bandwidth times the maximum translational speed.

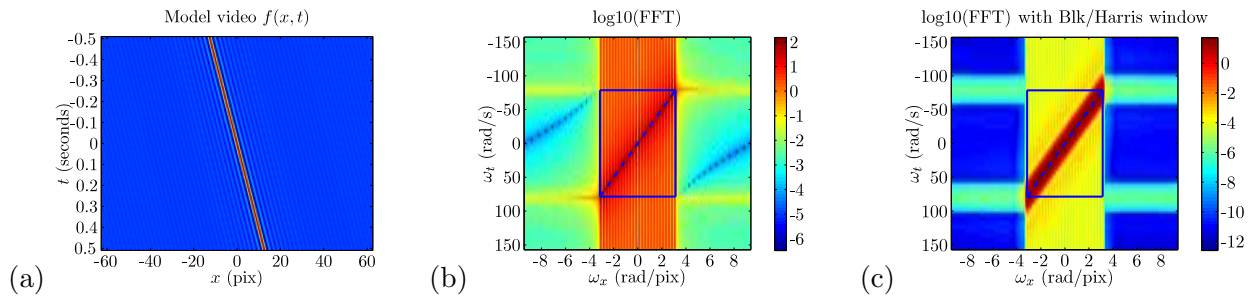


Figure 2: (a) Video  $f(x, t) = g(x - h(t))$  with bandlimited sinc profile  $g(x)$  and constant velocity translation signal  $h(t)$ ; see Section 2.4.1. (b) Estimated spectrum. (c) Estimated spectrum with windowing to alleviate border artifacts.

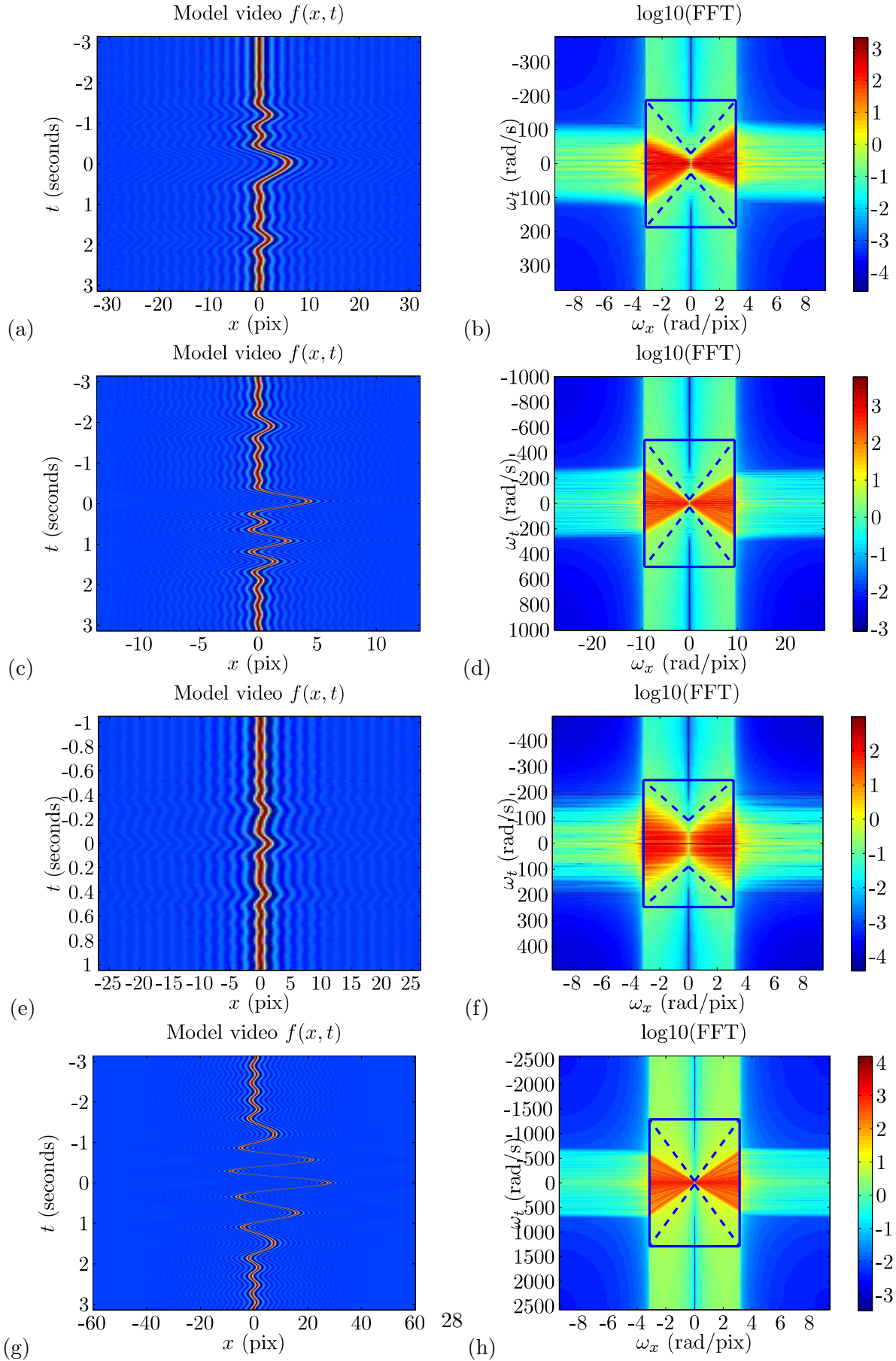


Figure 3: Left column: Videos  $f(x, t) = g(x - h(t))$  with bandlimited sinc profile  $g(x)$  and bandlimited “sum of sinc functions” model for the translation signal  $h(t)$ ; see Section 2.4.2. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra.



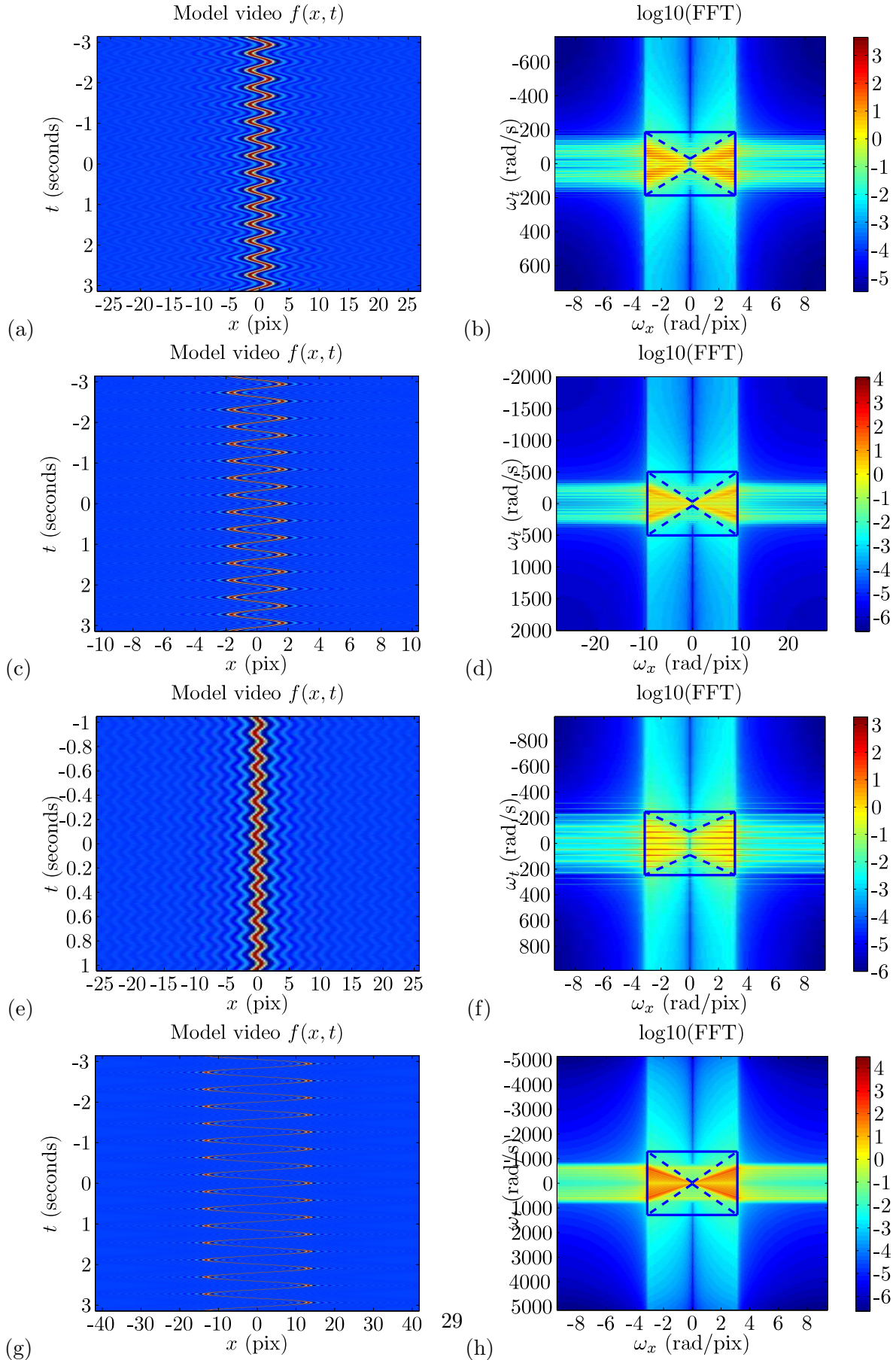


Figure 4: Left column: Videos  $f(x, t) = g(x - h(t))$  with bandlimited profile  $g(x)$  and bandlimited sinusoidal model for  $h(t)$ ; see Section 2.4.3. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra.

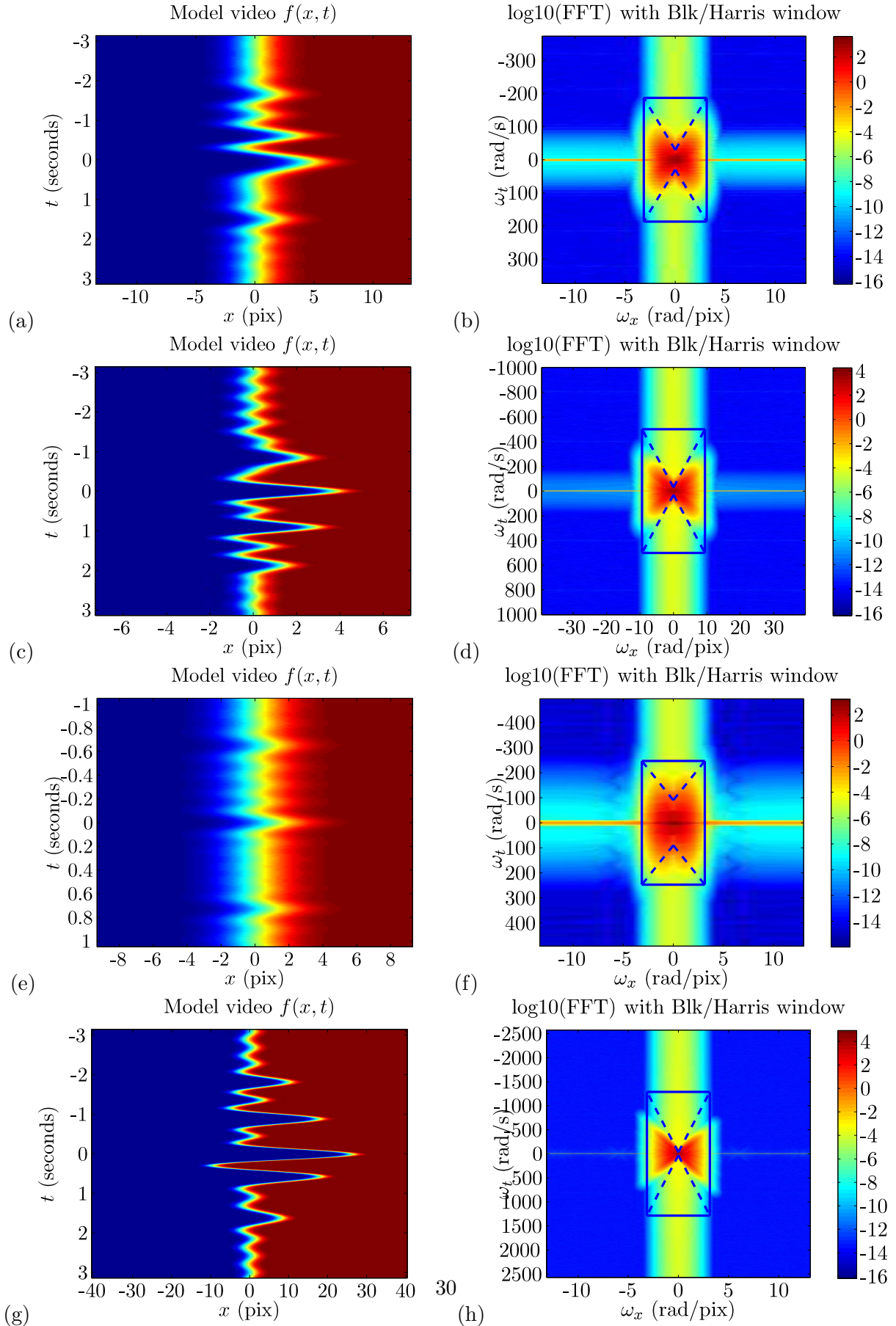


Figure 5: Left column: Videos  $f(x, t) = g(x - h(t))$  with non-bandlimited Gaussian-filtered step function  $g(x)$  and bandlimited “sum of sinc functions” model for  $h(t)$ ; see Section 2.5.1. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra with windowing to alleviate border artifacts.

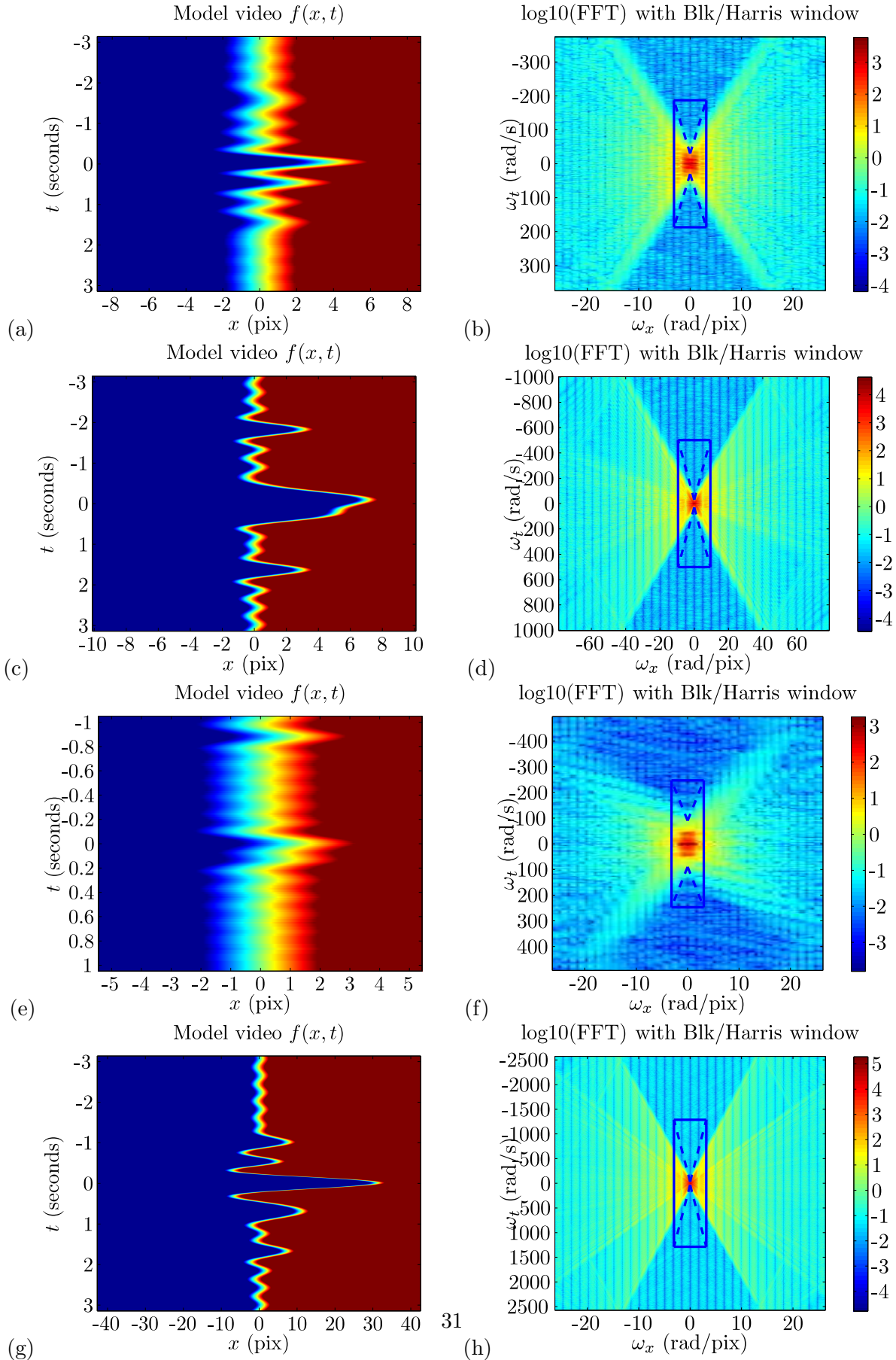


Figure 6: Left column: Videos  $f(x, t) = g(x - h(t))$  with non-bandlimited “triangular step” function  $g(x)$  and bandlimited “sum of sinc functions” model for  $h(t)$ ; see Section 2.5.1. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra with windowing to alleviate border artifacts.



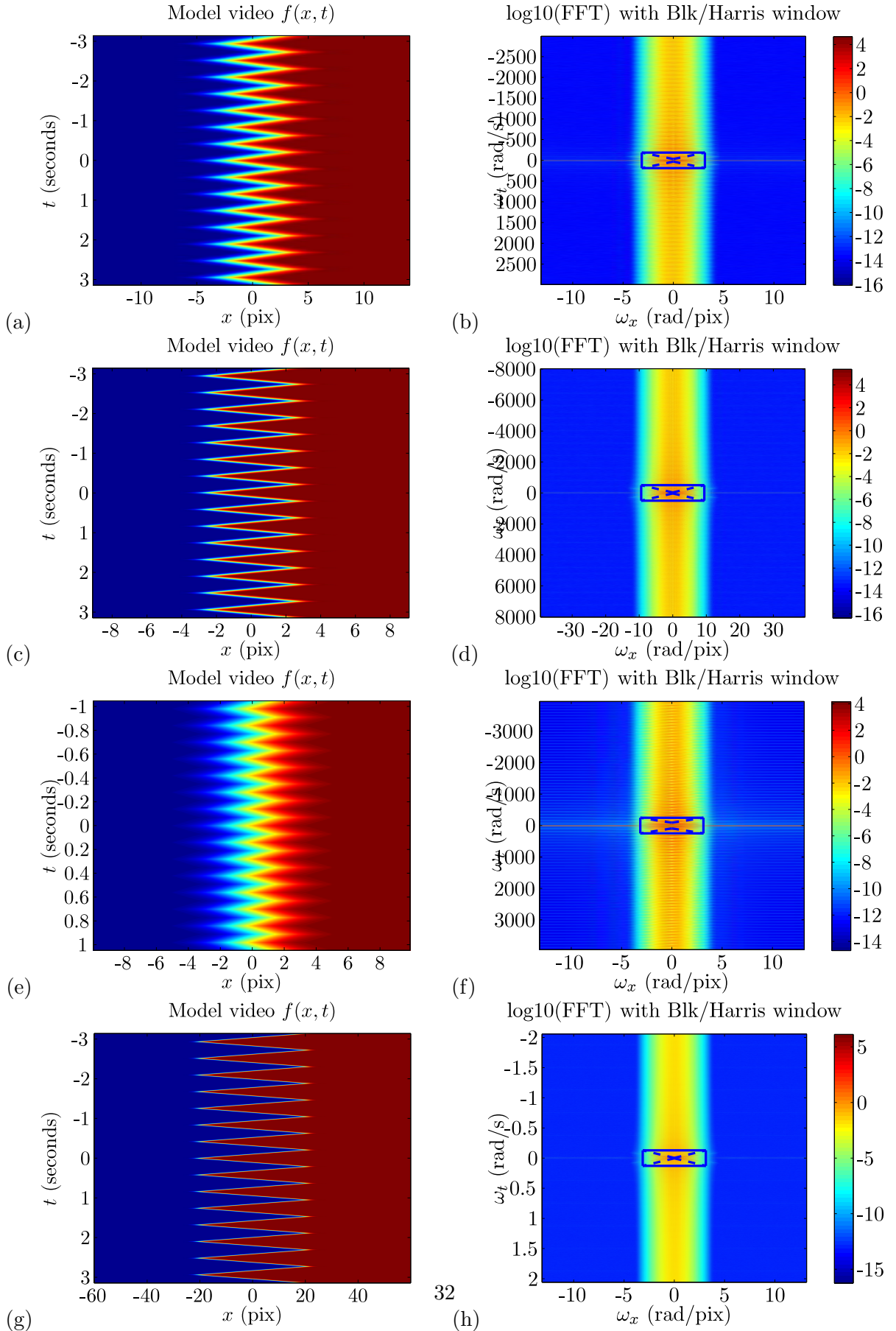


Figure 7: Left column: Videos  $f(x,t) = g(x - h(t))$  with non-bandlimited Gaussian-filtered step function  $g(x)$  and non-bandlimited triangle wave model for  $h(t)$ ; see Section 2.5.2. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra with windowing to alleviate border artifacts.

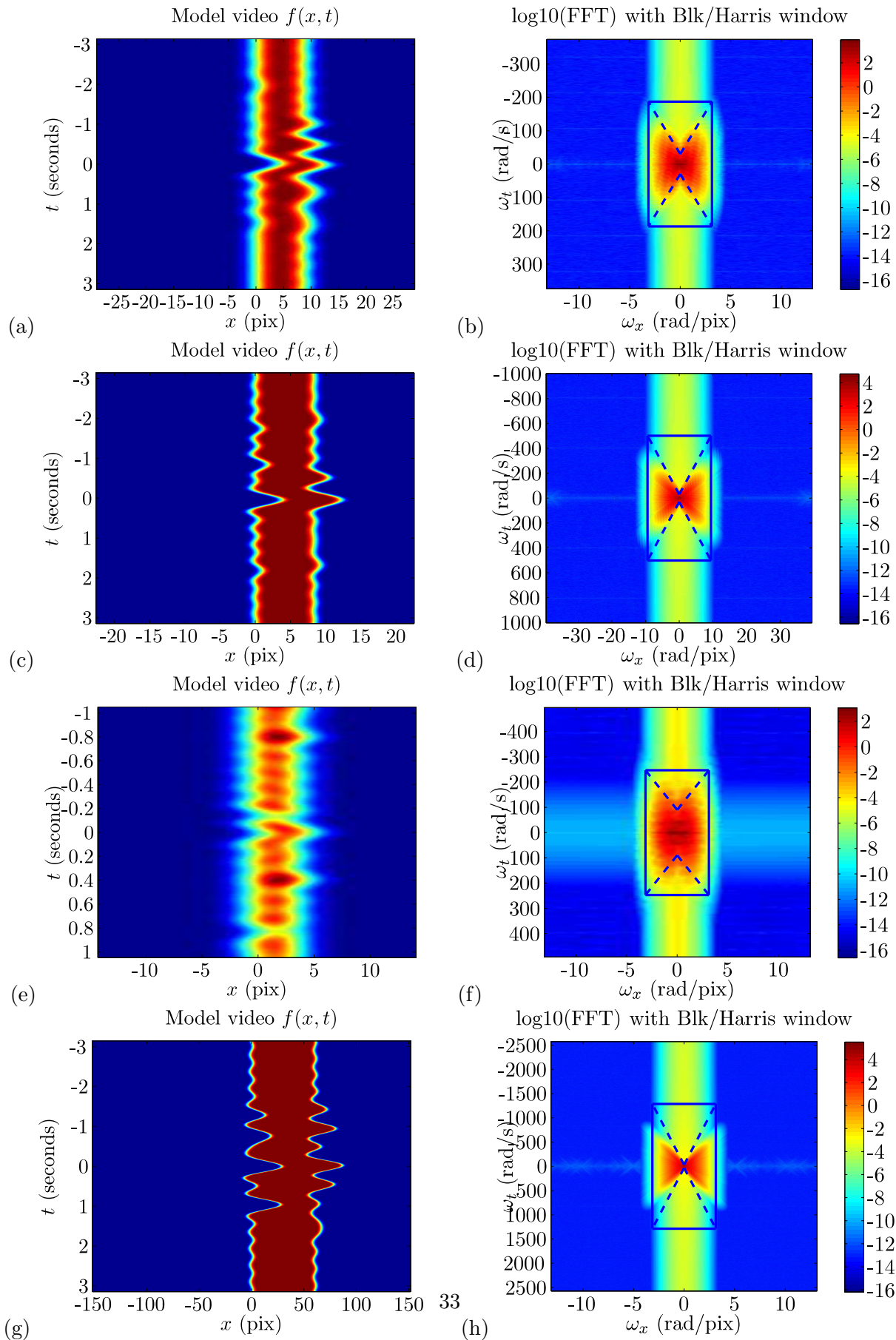


Figure 8: Left column: Videos  $f(x,t)$  containing multiple moving edges with non-bandlimited Gaussian-filtered step functions  $g(x)$  and bandlimited “sum of sinc functions” models for  $h(t)$ ; see Section 2.5.3. Each row corresponds to one set of parameter values ( $\Omega_x, \Omega_h, \Gamma$ ) as specified in Table 1. Right column: Estimated spectra with windowing to alleviate border artifacts.

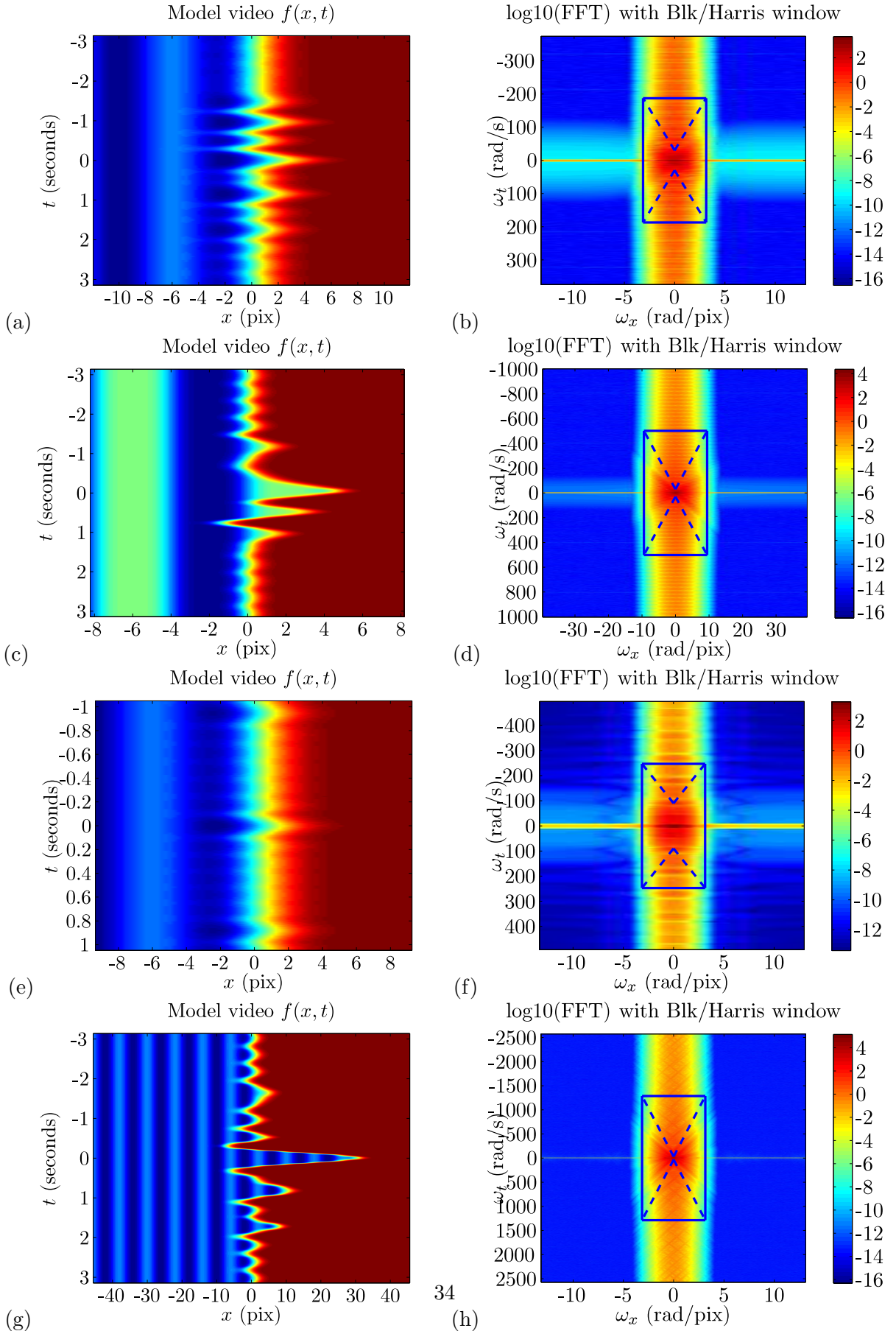


Figure 9: *Left column: Videos  $f(x, t)$  with a moving edge occluding a stationary background pattern; see Section 2.5.4. Each row corresponds to one set of parameter values  $(\Omega_x, \Omega_h, \Gamma)$  as specified in Table 1. Right column: Estimated spectra with windowing to alleviate border artifacts.*



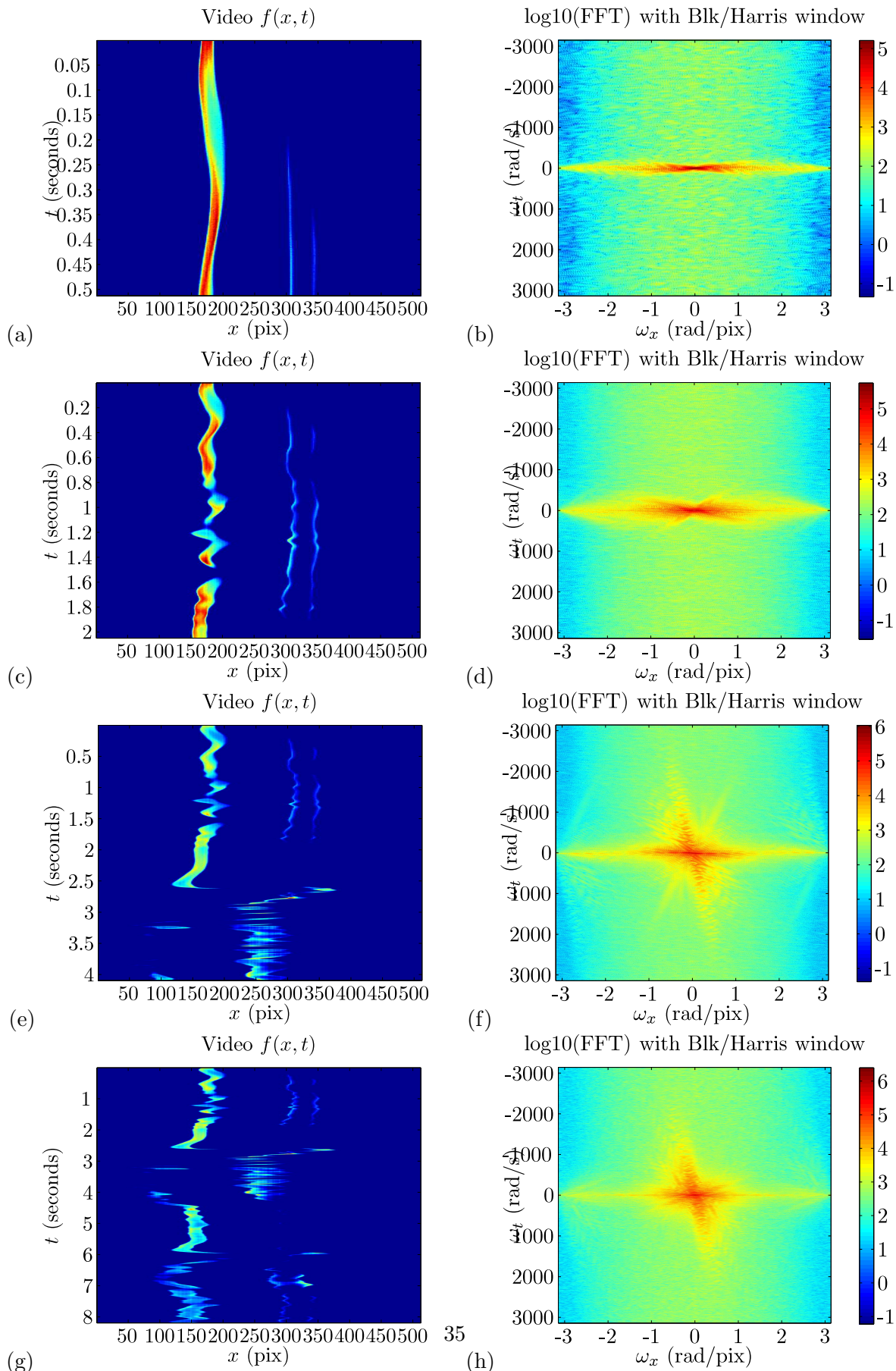


Figure 10: *Left column: Sampled 1D rows from Candle video; see Section 2.5.5. From top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. Right column: Estimated spectra with windowing to alleviate border artifacts.*

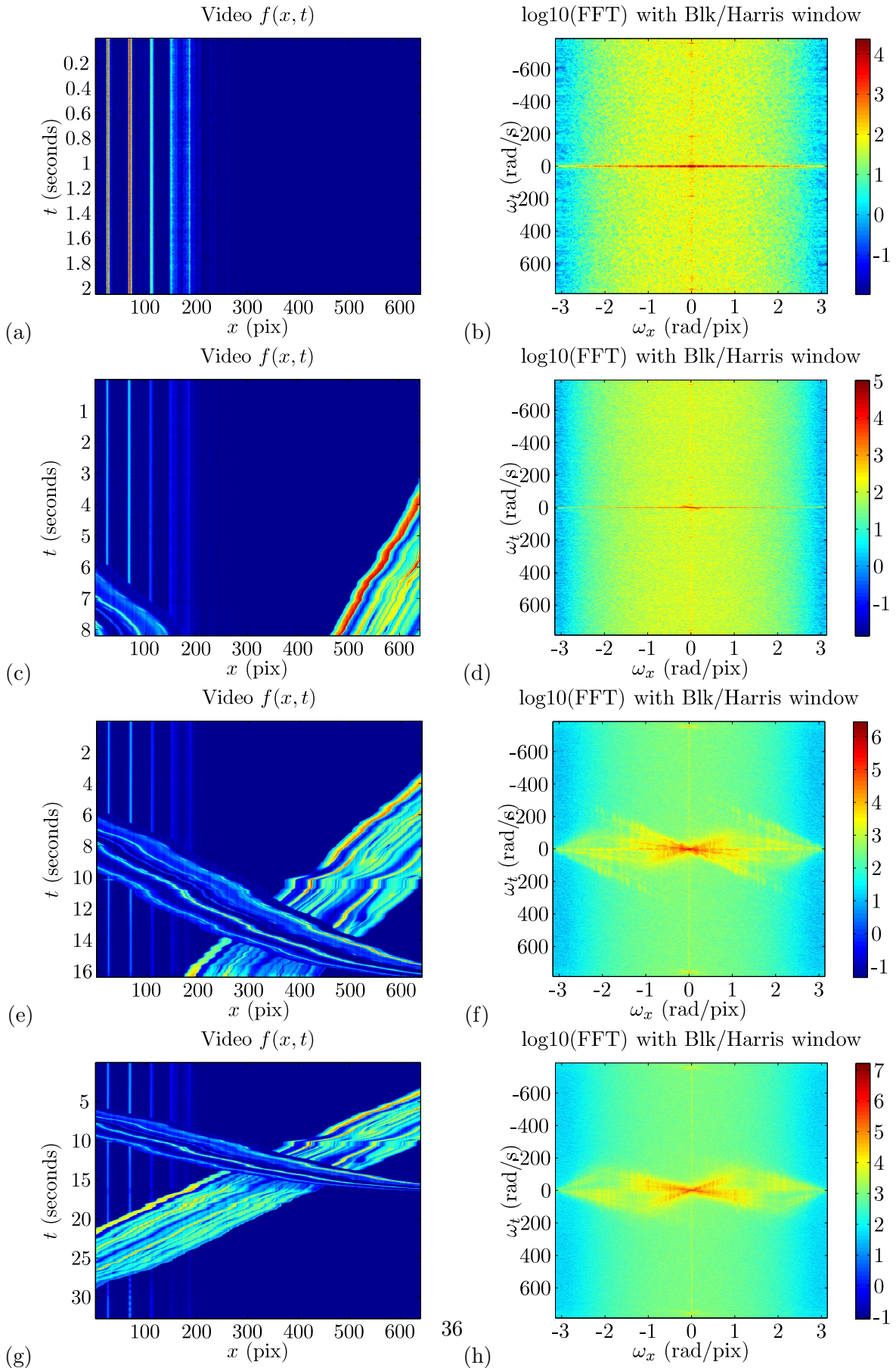


Figure 11: *Left column: Sampled 1D rows from Pendulum + Cars video; see Section 2.5.5. From top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. Right column: Estimated spectra with windowing to alleviate border artifacts.*



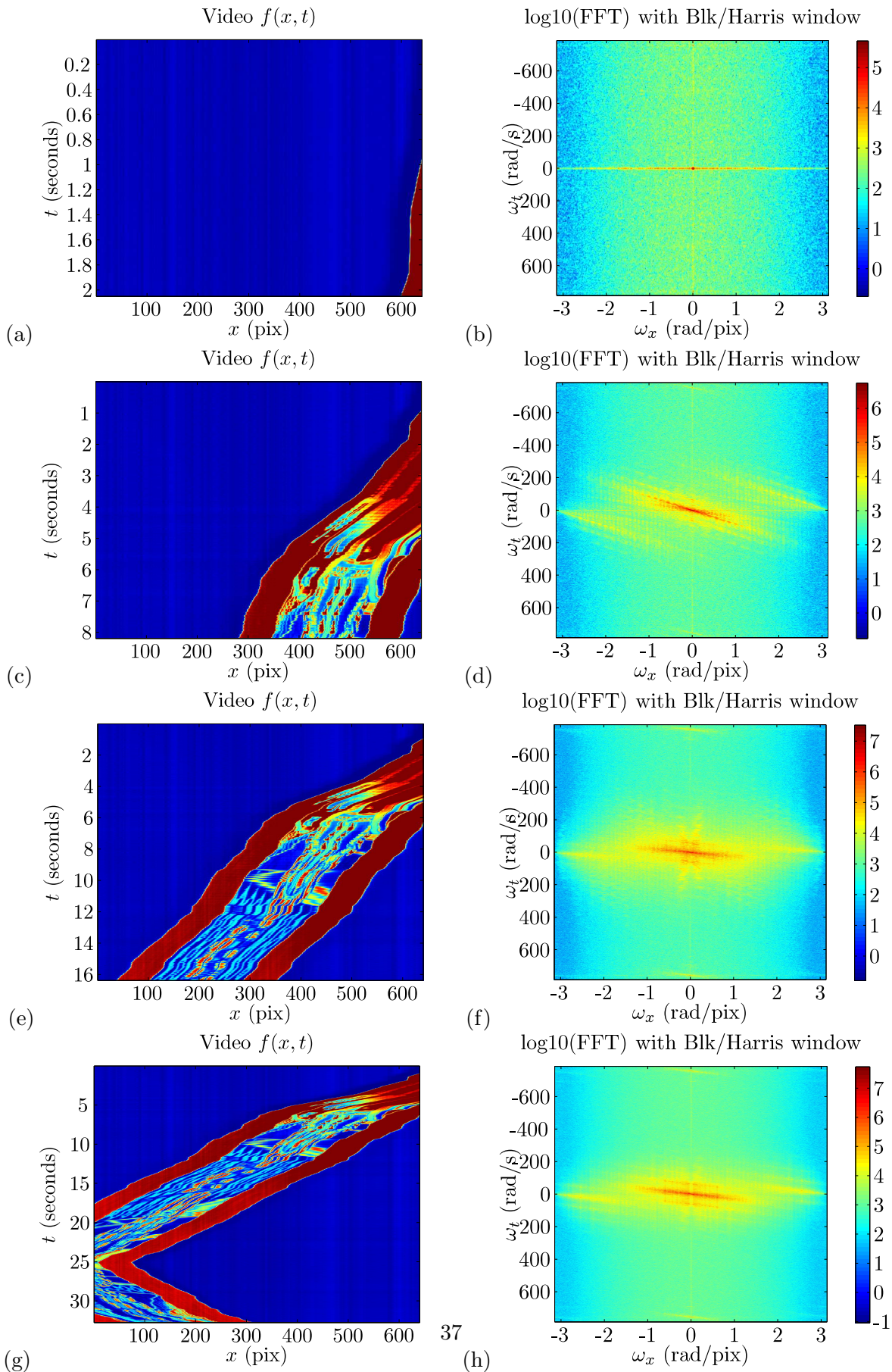
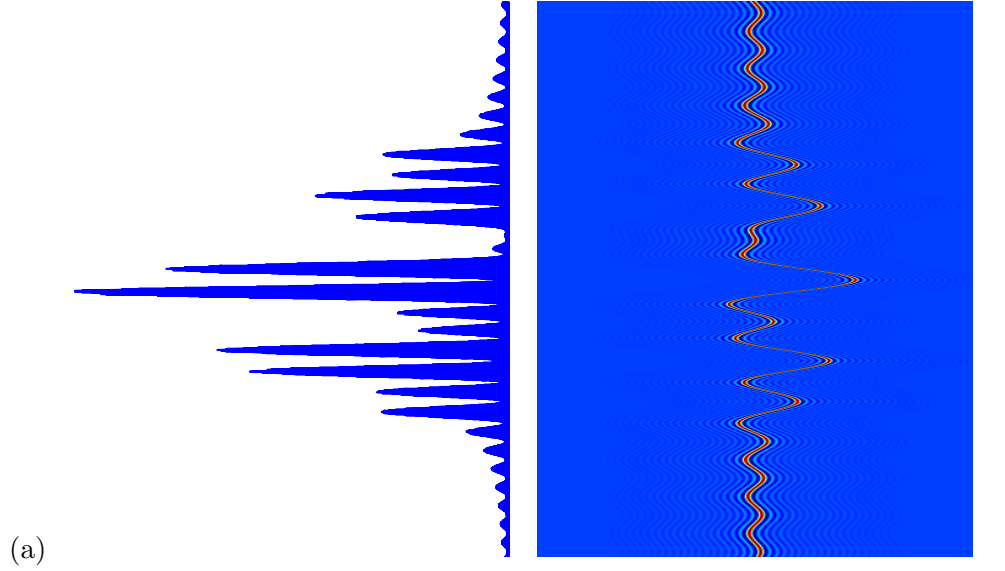
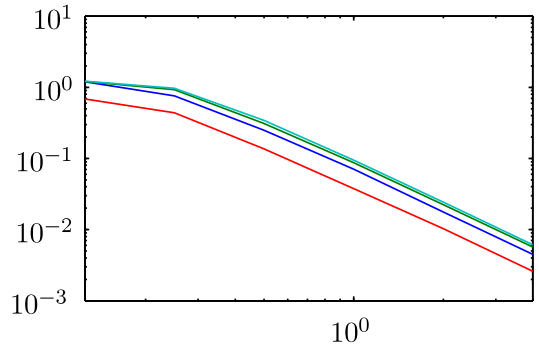


Figure 12: Left column: Sampled 1D rows from Card + Monster video; see Section 2.5.5. From top to bottom, we keep the first 512, 2048, 4096, and 8192 time samples of the video. Right column: Estimated spectra with windowing to alleviate border artifacts.



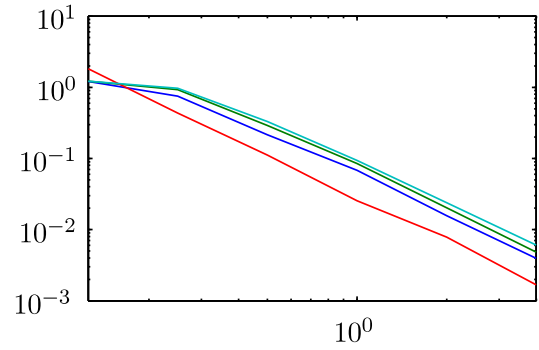
(a)

Max lin. interp. error



(b) Fraction of pred. Nyquist sample rate

Max cub. interp. error



(c) Fraction of pred. Nyquist sample rate

Figure 13: Interpolation experiments for videos  $f(x, t) = g(x - h(t))$  with bandlimited sinc profile  $g(x)$  and bandlimited “sum of sinc functions” model for the translation signal  $h(t)$ ; see Section 4.2 for details.

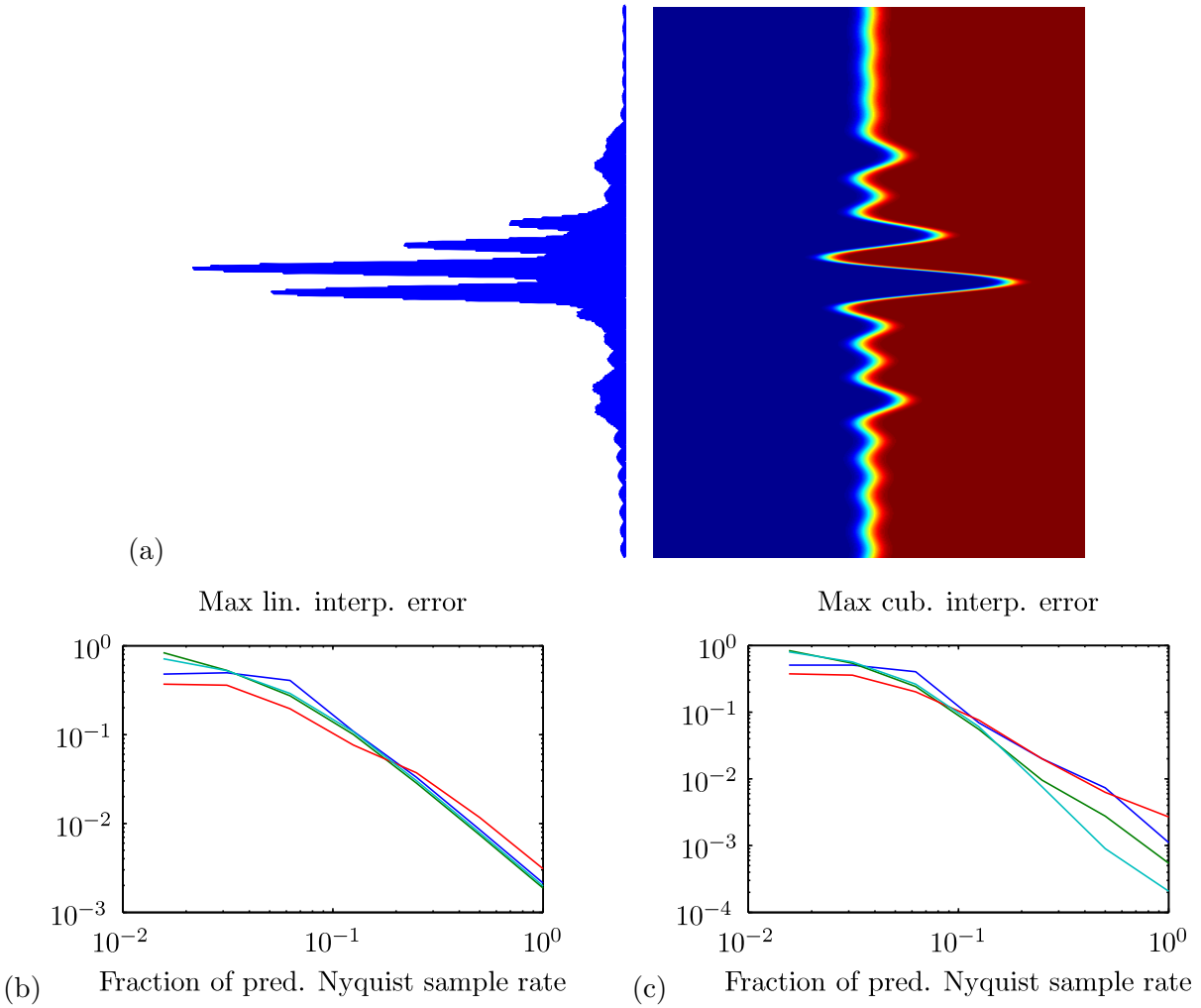


Figure 14: Interpolation experiments for videos  $f(x, t) = g(x - h(t))$  with non-bandlimited Gaussian-filtered step function  $g(x)$  and bandlimited “sum of sinc functions” model for  $h(t)$ ; see Section 4.3.1 for details.

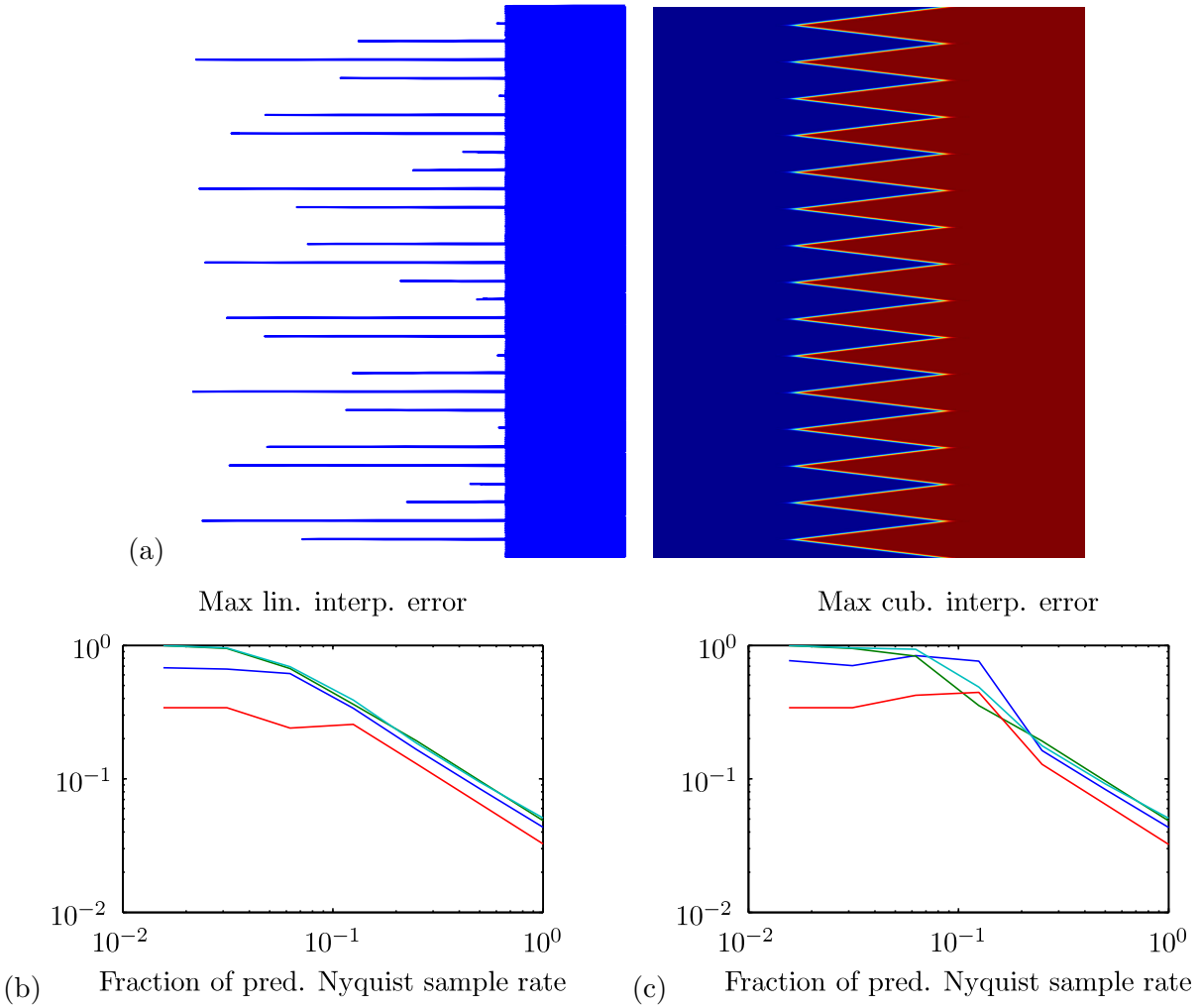


Figure 15: Interpolation experiments for videos  $f(x, t) = g(x - h(t))$  with non-bandlimited Gaussian-filtered step function  $g(x)$  and non-bandlimited triangle wave model for  $h(t)$ ; see Section 4.3.2 for details.

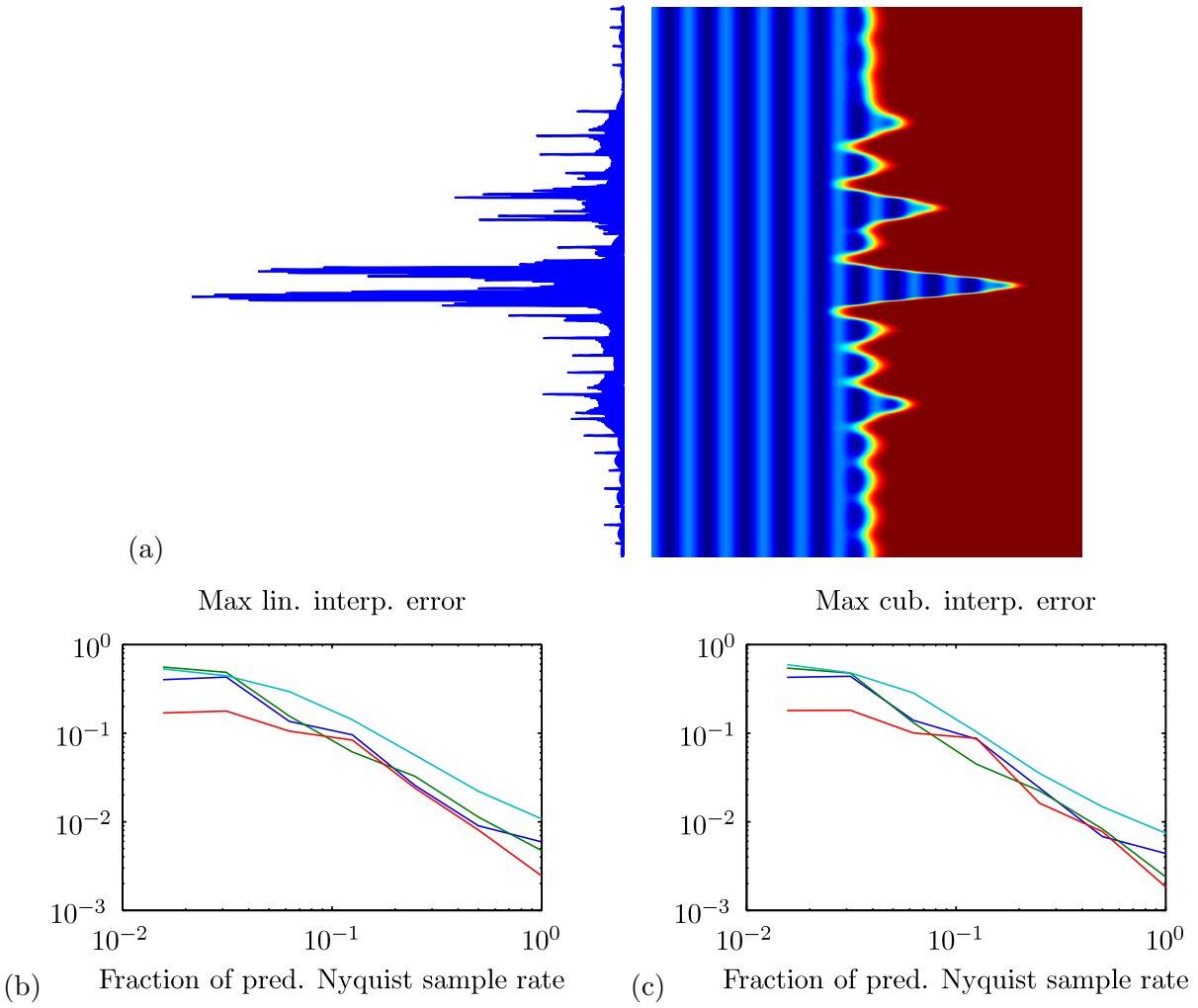


Figure 16: *Interpolation experiments for videos  $f(x,t)$  containing a moving edge occluding a stationary background pattern; see Section 4.3.3 for details.*

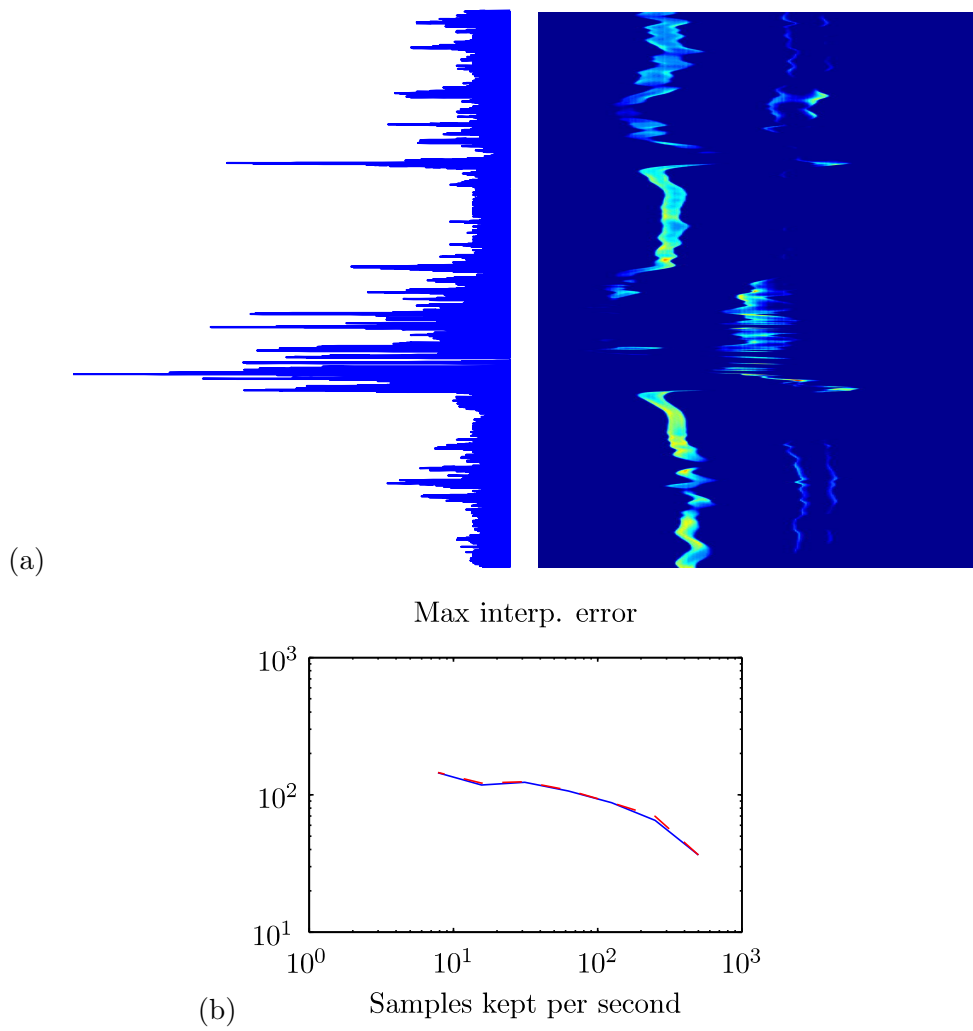


Figure 17: *Interpolation experiments for Candle video; see Section 4.3.4 for details.*

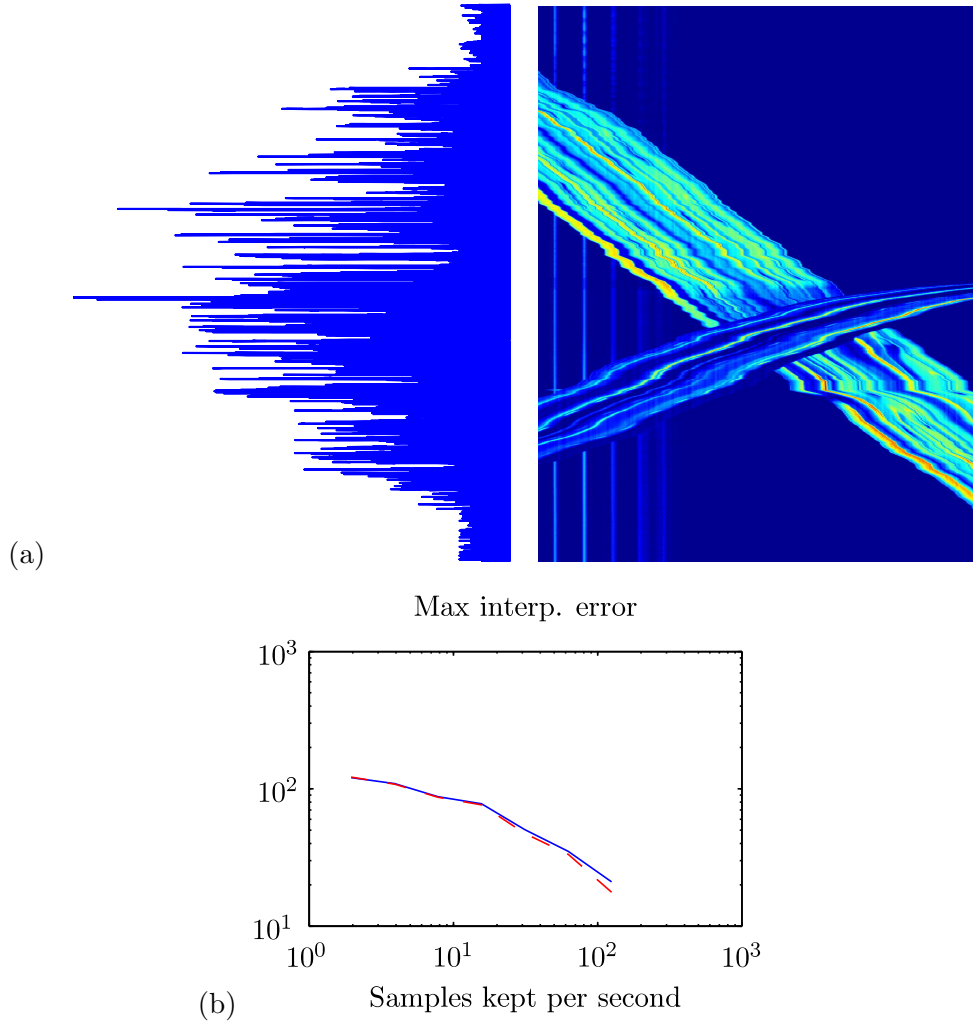


Figure 18: *Interpolation experiments for Pendulum + Cars video; see Section 4.3.4 for details.*

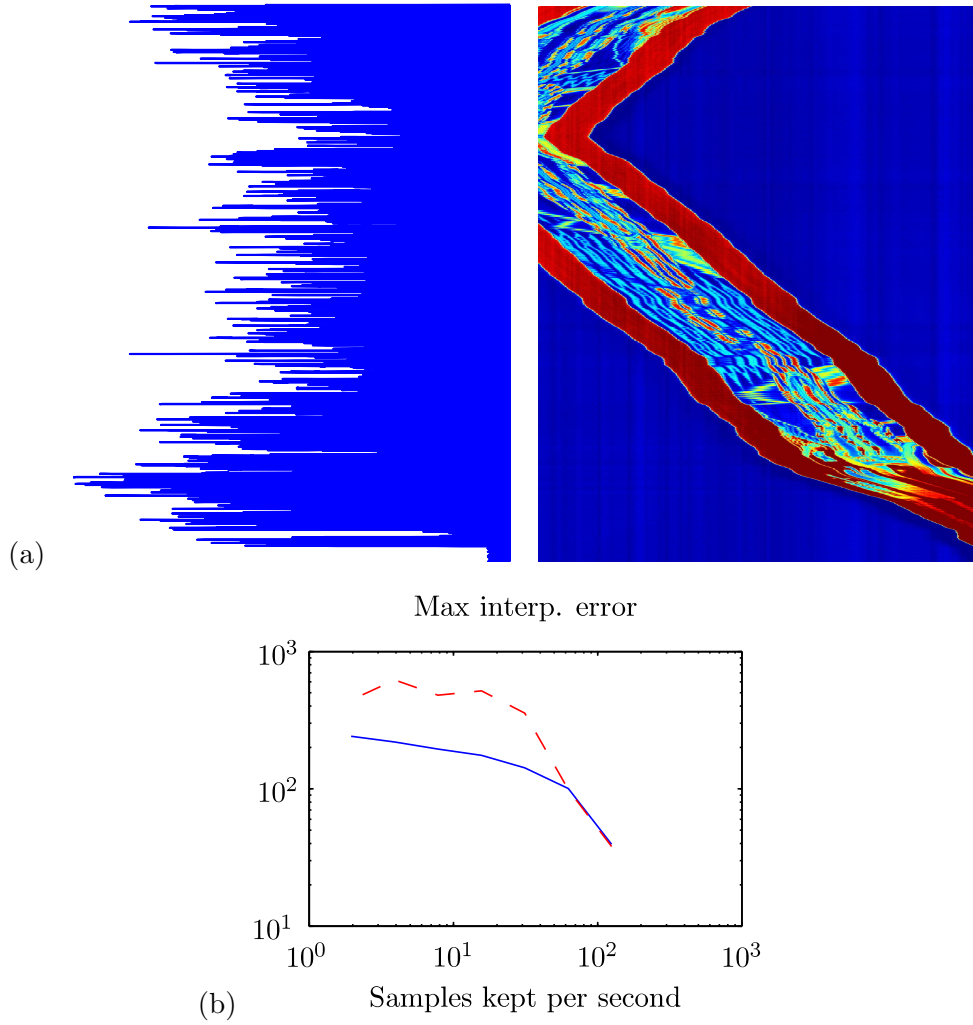


Figure 19: *Interpolation experiments for Card + Monster video; see Section 4.3.4 for details.*



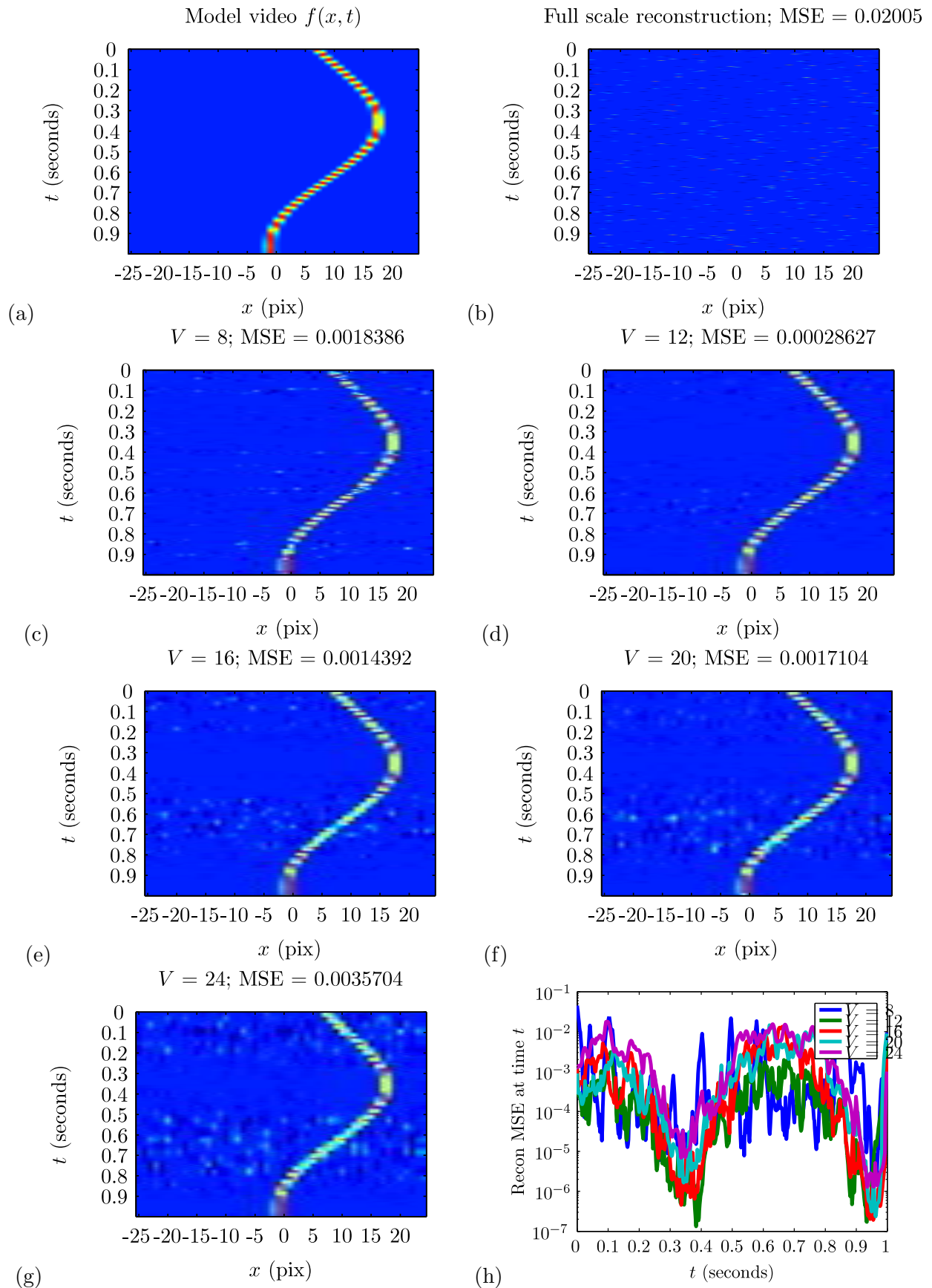


Figure 20: CS reconstruction experiments for a translational video with a single moving pulse using a linear interpolation kernel; see Section 5.3.1 for details.

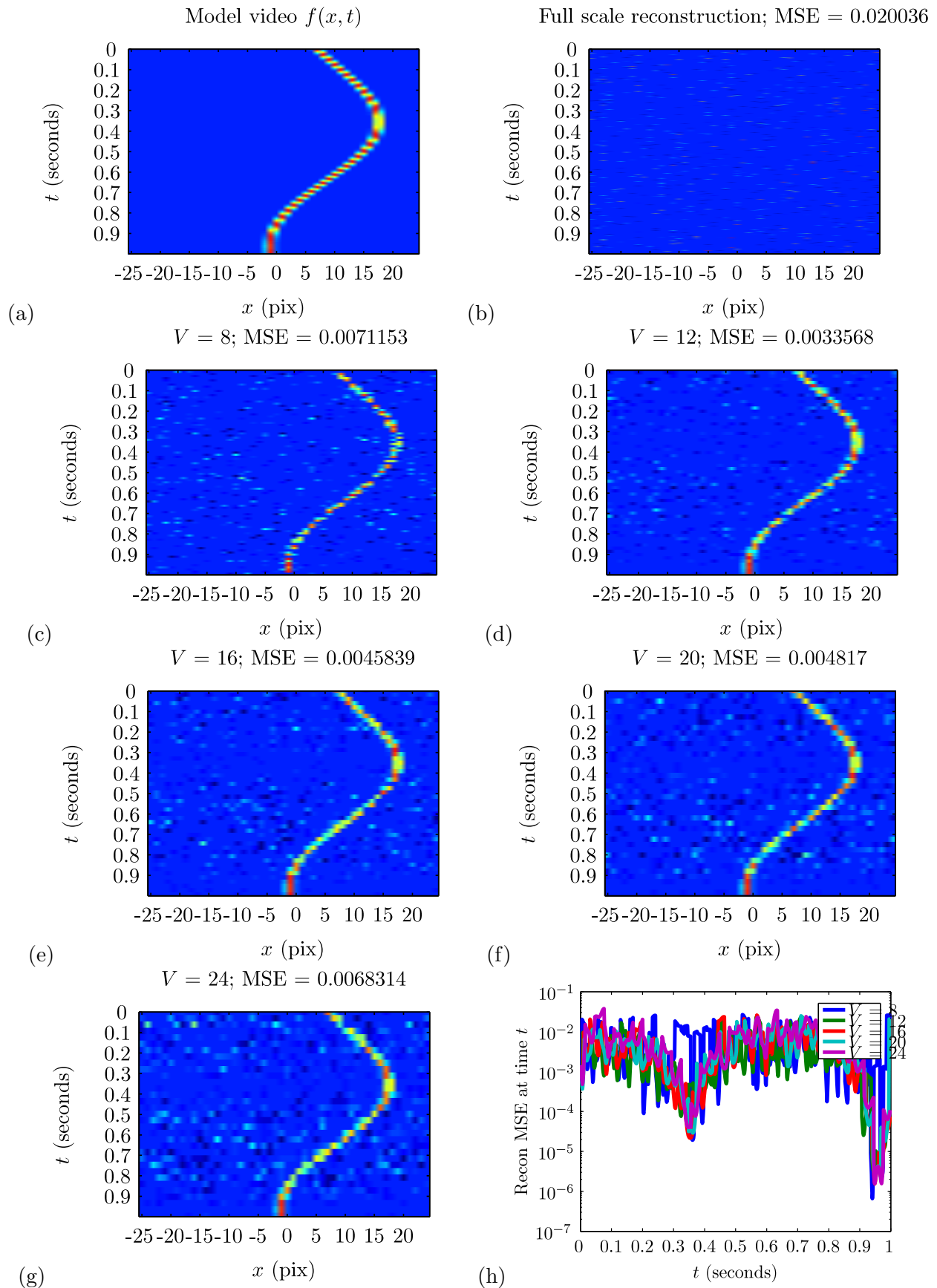


Figure 21: CS reconstruction experiments for a translational video with a single moving pulse using a nearest neighbor (rectangular) interpolation kernel; see Section 5.3.1 for details.

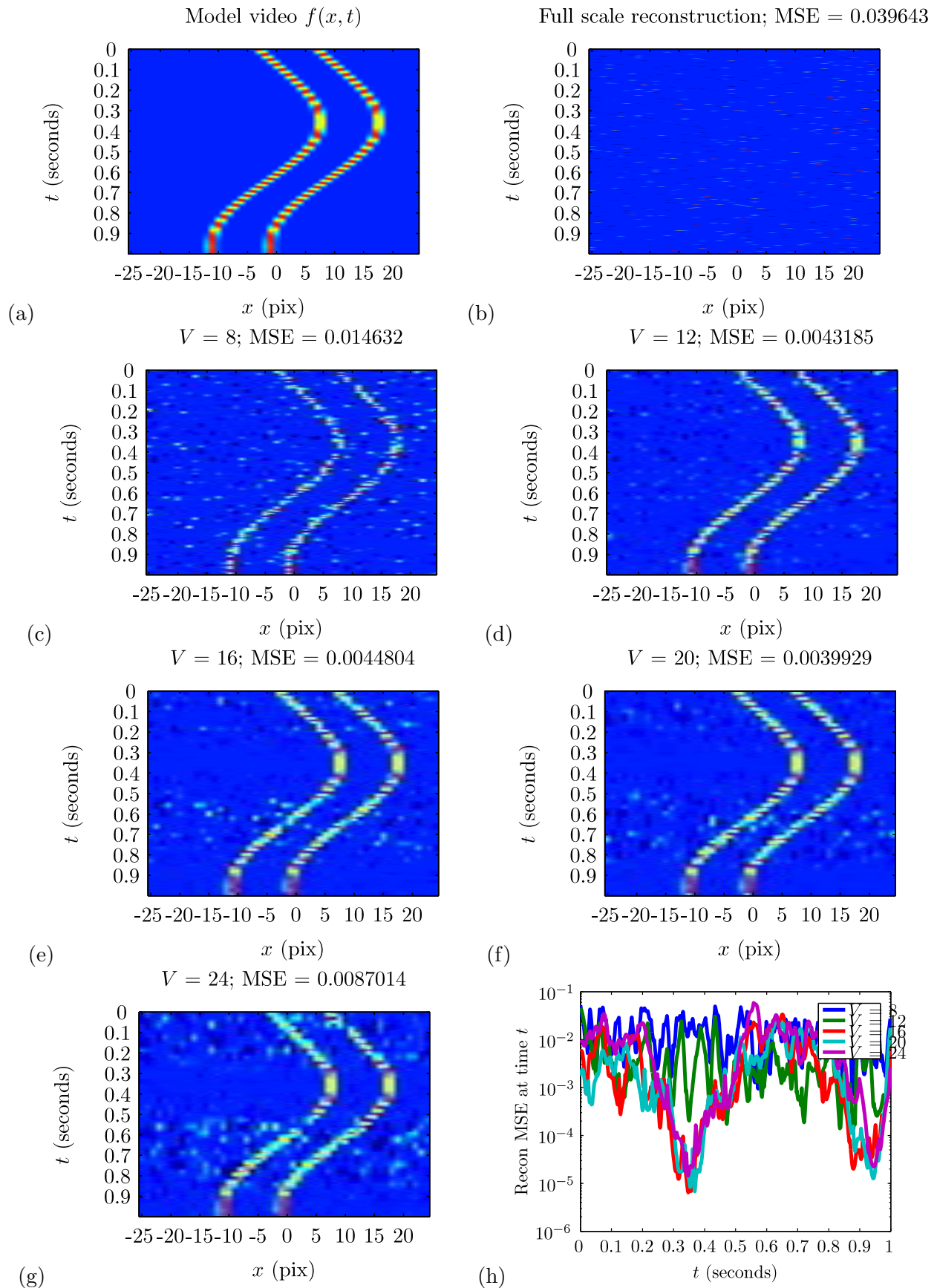


Figure 22: CS reconstruction experiments for a translational video with a two moving pulses using a linear interpolation kernel; see Section 5.3.2 for details.

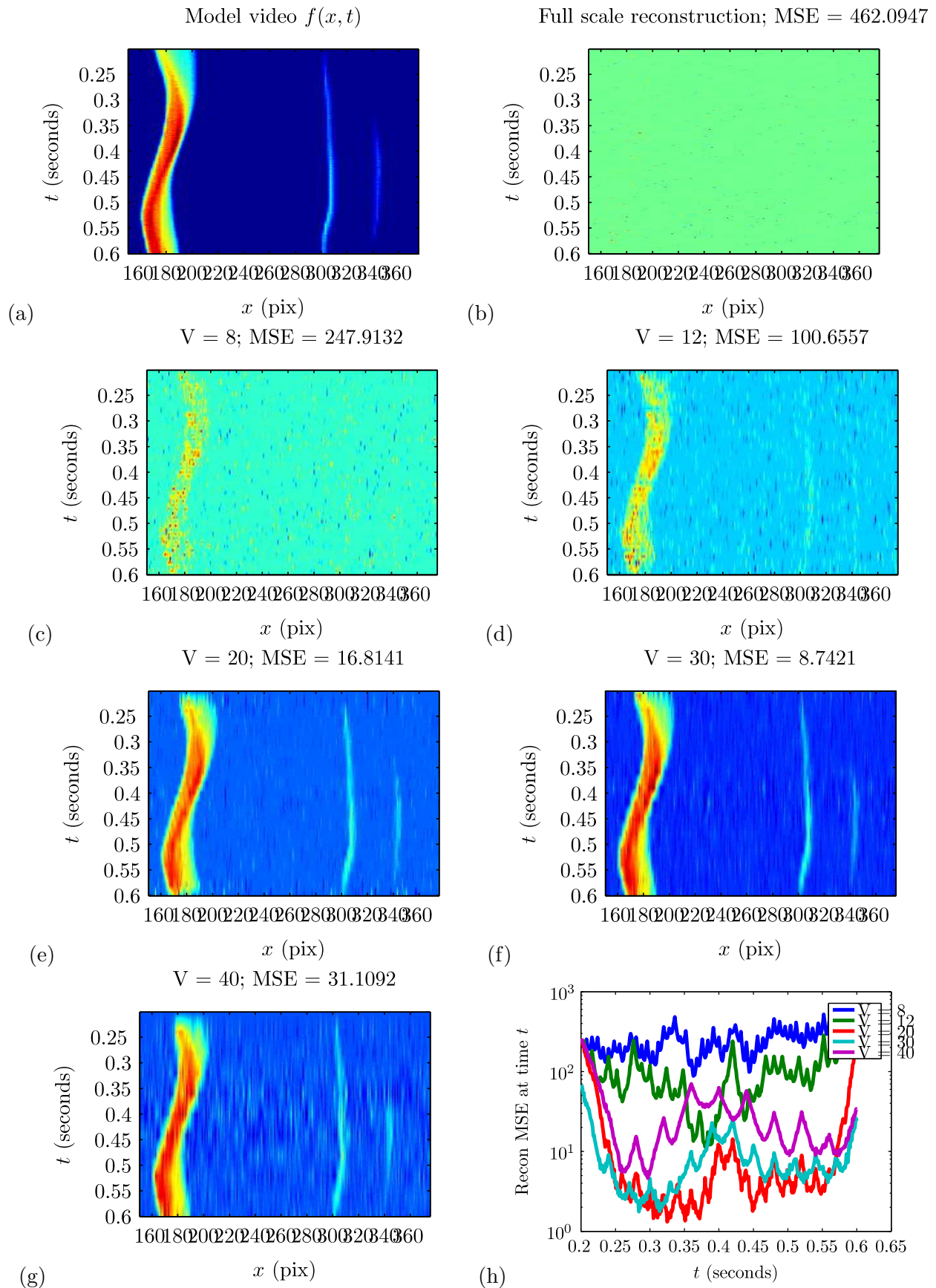


Figure 23: CS reconstruction experiments for Candle video using a linear interpolation kernel; see Section 5.3.3 for details.

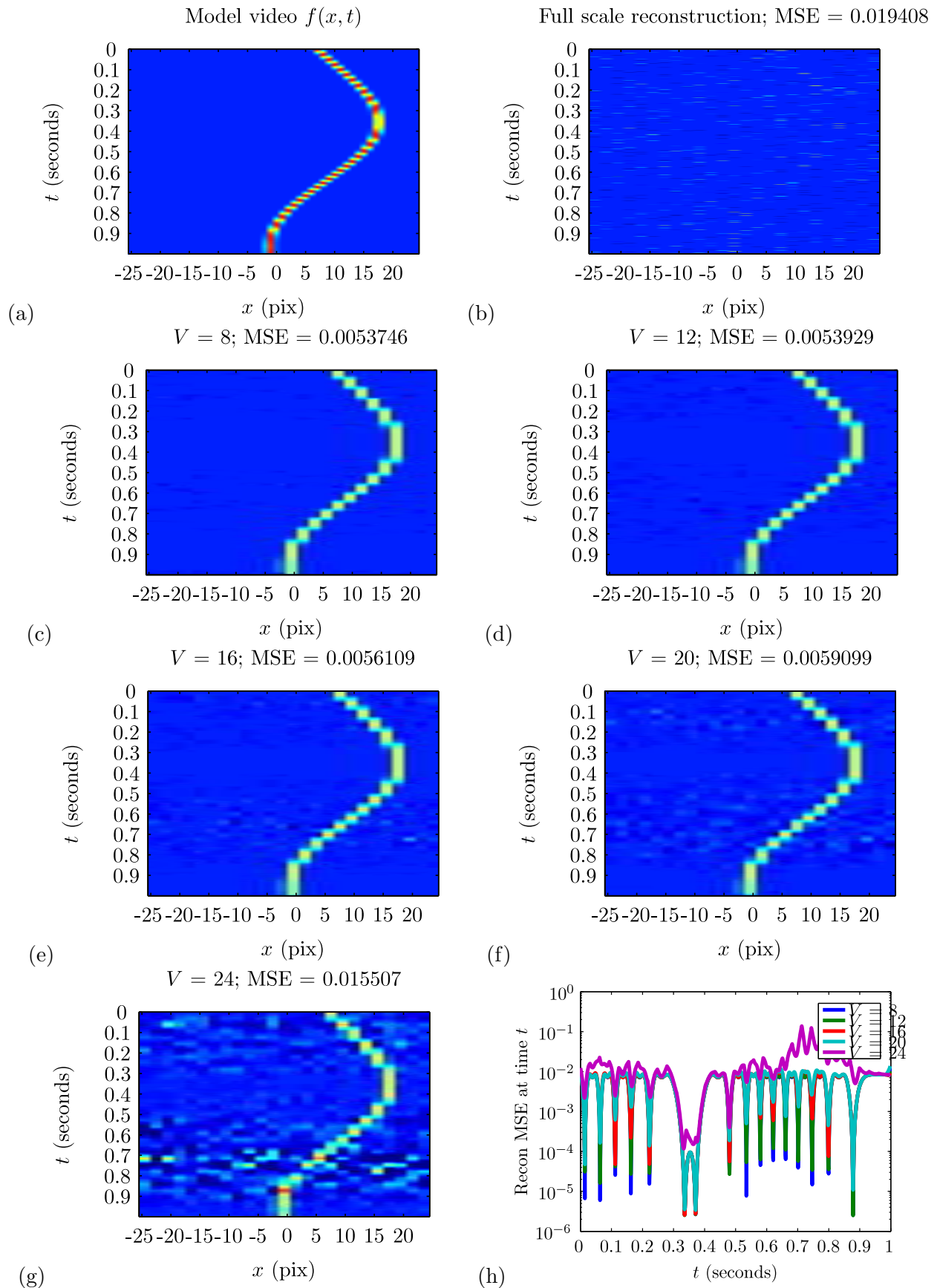


Figure 24: CS reconstruction experiments for a translational video with a single moving pulse using a linear interpolation kernel, and with lowpass filtered measurement vectors; see Section 5.4.3 for details.



