7.2

Video Sampling and Interpolation

Eric Dubois
University of Ottawa

1 Introduction

This chapter is concerned with the sampled representation of time-varying imagery, often referred to as video. Time-varying imagery must be sampled in at least one dimension for the purposes of transmission, storage, processing or display. Examples are one-dimensional temporal sampling in motion-picture film, two-dimensional vertical-temporal scanning in the case of analog television, and three-dimensional horizontal-vertical-temporal sampling in digital video. In some cases a single sampling structure is used throughout an entire video processing or communication system. This is the case in standard analog television broadcasting where the signal is acquired, transmitted and, displayed using the same scanning standard from end to end. However, it is becoming increasingly more common to have different sampling structures used in the acquisition, processing, transmission, and display components of the system. In addition, the number of different sampling structures in use throughout the world is increasing. Thus, sampling structure conversion for video system is an important problem.

The initial acquisition and scanning is particularly critical because it determines what information is contained in the original data. The acquisition process can be modeled as a continuous space-time prefiltering followed by ideal sampling on a given sampling structure. The sampling structure determines the amount of spatio-temporal information that the sampled signal can carry, while the prefiltering serves to limit the amount of aliasing. At the final stage of the system, the desired display characteristics are closely related to the properties of the human visual system. The goal of the display is to convert the sampled signal to a continuous time-varying image that, when presented to the viewer, approximates the original continuous scene as closely as possible. In particular, the effects caused by sampling should be sufficiently attenuated so as to lie below the threshold of perceptibility.

This chapter has three main sections. First the sampling lattice, the basic tool in the analysis of spatiotemporal sampling, is introduced. The issues involved in the sampling and reconstruction of continuous time-varying imagery are then addressed. Finally, methods for the conversion of image sequences between different sampling structures are presented.

2 Spatiotemporal Sampling Structures

A continuous time-varying image $f_c(x, y, t)$ is a function of two spatial dimensions x and y and time t, usually observed in a rectangular spatial window $\mathcal{W}$ over some time interval $\mathcal{T}$. The spatiotemporal region $\mathcal{W} \times \mathcal{T}$ is denoted $\mathcal{W}_T$. The spatial window is of dimension $pw \times ph$ where pw is the picture width and ph is the picture height. Since the absolute physical size of an image depends on the display device used, and the sampling density for a particular video signal may be variable,

we choose to adopt the picture height *ph* as the basic unit of spatial distance, as is common in the broadcast video industry. The ratio of *pw/ph* is called the aspect ratio, the most common values being 4/3 for standard TV and 16/9 for HDTV. The image f_c can be sampled in one, two or three dimensions. It is almost always sampled in at least the temporal dimension, producing an *image sequence*. An example of an image sampled only in the temporal dimension is motion picture film. Analog video is typically sampled in the vertical and temporal dimensions while digital video is sampled in all three dimensions. The subset of $\mathbb{R}^3$ on which the sampled image is defined is called the *sampling structure* Ψ; it is contained in $\mathcal{W}_T$.

The mathematic structure most useful in describing sampling of time-varying images is the *lattice*. A discussion of lattices from the point of view of video sampling can be found in [1] and [2]. Some of the main properties are summarized here. A lattice Λ in D dimensions ia a discrete set of points that can be expressed as the set of all linear combinations with integer coefficients of D linearly independent vectors in $\mathbb{R}^D$ (called basis vectors),

$$\Lambda = \{n_1\boldsymbol{v}_1 + \cdots + n_D\boldsymbol{v}_D \mid n_i \in \mathbb{Z}\}, \tag{1}$$

where $\mathbb{Z}$ is the set of integers. For our purposes, D will be 1, 2 or 3 dimensions. The matrix $V = [\boldsymbol{v}_1 \mid \boldsymbol{v}_2 \mid \cdots \mid \boldsymbol{v}_D]$ whose columns are the basis vectors $\boldsymbol{v}_i$ is called a sampling matrix and we write $\Lambda = \mathrm{LAT}(V)$. The basis or sampling matrix for a given lattice is not unique however, since $\mathrm{LAT}(V) = \mathrm{LAT}(VE)$ where E is any unimodular ($|\det E| = 1$) integer matrix. Figure 1 shows an example of a lattice in two dimensions, with basis vectors $\boldsymbol{v}_1 = [2X, 0]^T$ and $\boldsymbol{v}_2 = [X, Y]^T$. The sampling matrix in this case is

$$V_\Lambda = \begin{bmatrix} 2X & X \\ 0 & Y \end{bmatrix}.$$

A *unit cell* of a lattice Λ is a set $\mathcal{P} \subset \mathbb{R}^D$ such that copies of $\mathcal{P}$ centered on each lattice point tile the whole space without overlap: $(\mathcal{P} + \boldsymbol{s}_1) \cap (\mathcal{P} + \boldsymbol{s}_2) = \emptyset$ for $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \Lambda, \boldsymbol{s}_1 \neq \boldsymbol{s}_2$, and $\cup_{\boldsymbol{s}\in\Lambda}(\mathcal{P} + \boldsymbol{s}) = \mathbb{R}^D$. The volume of a unit cell is $d(\Lambda) = |\det V|$, which is independent of the particular choice of sampling matrix. We can imagine that there is a region congruent to $\mathcal{P}$ of volume $d(\Lambda)$ associated with each sample in Λ, so that $d(\Lambda)$ is the reciprocal of the sampling density. The unit cell of a lattice is not unique. In Fig. 1, the shaded hexagonal region centered at the origin is a unit cell, of area $d(\Lambda) = 2XY$. The shaded parallelogram in the upper right is also a possible unit cell.

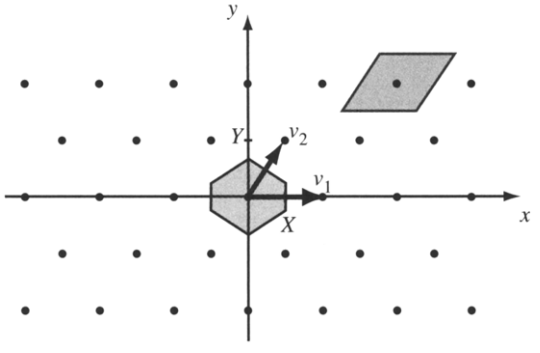

FIGURE 1 Example of a lattice in two dimensions with two possible unit cells.

Most sampling structures of interest for time-varying imagery can be constructed using a lattice. In the case of 3D sampling, the sampling structure can be the intersection of $\mathcal{W}_T$ with a lattice, or in a few cases, with the union of two or more shifted lattices. The latter case occurs relatively infrequently (although there are several practical situations where it is used) and so the discussion here is limited to sampling on lattices. The theory of sampling on the union of shifted lattices (cosets) can be found in [1]. In the case of one or two-dimensional (partial) sampling ($D = 1$ or 2), the sampling structure can be constructed as the Cartesian product of a D-dimensional lattice and a continuous $(3 - D)$ dimensional space. For one-dimensional temporal sampling, the 1D lattice is $\Lambda_t = \{nT \mid n \in \mathbb{Z}\}$ where T is the frame period. The sampling structure is then $\mathcal{W} \times \Lambda_t = \{(\boldsymbol{x}, t) \mid \boldsymbol{x} \in \mathcal{W}, t \in \Lambda_t\}$. For two-dimensional vertical-temporal sampling (scanning) using a 2D lattice Λ_{yt}, the sampling structure is $\mathcal{W}_T \cap (\mathcal{H} \times \Lambda_{yt})$ where $\mathcal{H}$ is a one-dimensional subspace of $\mathbb{R}^3$ parallel to the scanning lines. In video systems, the scanning spot is moving down as it scans from left to right, and of course is moving forward in time. Thus $\mathcal{H}$ has both a vertical and temporal tilt, but this effect is minor and can usually be ignored; we assume that $\mathcal{H}$ is the line $y = 0$, $t = 0$. Many digital video signals are obtained by three-dimensional subsampling of signals that have initially been sampled with one or two-dimensional sampling as above. Although the sampling structure is space limited, the analysis is often simplified if the sampling structure is assumed to be of infinite spatial extent, with the image either set to zero outside of $\mathcal{W}_T$ or replicated periodically in some way.

Much insight into the effect of sampling time-varying images on a lattice can be achieved by studying the problem in the frequency domain. To do this, we introduce the Fourier transform for signals defined on different domains. For a continuous signal f_c the Fourier transform is given by

$$F_c(u, v, w) = \iiint f_c(x, y, t) \exp\left[-j2\pi(ux + vy + wt)\right] dx\, dy\, dt \tag{2}$$

or more compactly, setting $\boldsymbol{u} = (u, v, w)$ and $\boldsymbol{s} = (x, y, t)$,

$$F_c(\boldsymbol{u}) = \int_{\mathcal{W}_T} f_c(\boldsymbol{s}) \exp\left(-j2\pi\boldsymbol{u} \cdot \boldsymbol{s}\right) d\boldsymbol{s}, \quad \boldsymbol{u} \in \mathbb{R}^3. \tag{3}$$

The variables u and v are horizontal and vertical spatial frequencies in cycles/picture height (c/ph) and w is temporal frequency in Hz.

Similarly, a discrete signal $f(\boldsymbol{s})$, $\boldsymbol{s} \in \Lambda$ has a lattice Fourier transform (or discrete space-time Fourier transform)

$$F(\boldsymbol{u}) = \sum_{\boldsymbol{s}\in\Lambda} f(\boldsymbol{s}) \exp\bigl(-j2\pi \boldsymbol{u}\cdot\boldsymbol{s}\bigr), \quad \boldsymbol{u} \in \mathbb{R}^3. \tag{4}$$

With this non-normalized definition, both $\boldsymbol{s}$ and $\boldsymbol{u}$ have the same units as in Eq. (3). As with the 1D discrete-time Fourier transform, the lattice Fourier transform is periodic. If $\boldsymbol{k}$ is an element of $\mathbb{R}^3$ such that $\boldsymbol{k}\cdot\boldsymbol{s} \in \mathbb{Z}$ for all $\boldsymbol{s} \in \Lambda$, then $F(\boldsymbol{u}+\boldsymbol{k}) = F(\boldsymbol{u})$. It can be shown that $\{\boldsymbol{k} \mid \boldsymbol{k}\cdot\boldsymbol{s} \in \mathbb{Z} \text{ for all } \boldsymbol{s} \in \Lambda\}$ is a lattice called the *reciprocal lattice* Λ^*, and that if V is a sampling matrix for Λ, then $\Lambda^* = \mathrm{LAT}((V^T)^{-1})$. Thus $F(\boldsymbol{u})$ is completely specified by its values in a unit call of Λ^*.

For partially sampled signals, a mixed Fourier transform is required. For the examples of temporal and vertical-temporal sampling mentioned previously, these Fourier transforms are

$$F(\boldsymbol{u}, w) = \int_{\mathcal{W}} \sum_{n} f(\boldsymbol{x}, nT) \exp\bigl[-j2\pi(\boldsymbol{u}\cdot\boldsymbol{x} + wnT)\bigr] d\boldsymbol{x} \tag{5}$$

and

$$F(u, v, w) = \int_{\mathcal{H}} \sum_{(y,t)\in\Lambda_{yt}} f(x, y, t) \exp\bigl[-j2\pi(ux + vy + wt)\bigr] dx. \tag{6}$$

These Fourier transforms are periodic in the temporal frequency domain (with periodicity $1/T$) and in the vertical-temporal frequency domain (with periodicity lattice Λ_{yt}^*) respectively.

The terminology is illustrated with two examples that will be discussed in more detail later in this chapter. Figure 2 shows two vertical-temporal sampling lattices: a rectangular lattice Λ_R in Fig. 2(a) and a hexagonal lattice Λ_H in Fig. 2(b). These correspond to progressive scanning and interlaced scanning respectively in video systems. Possible sampling matrices for the two lattices are

$$V_R = \begin{bmatrix} Y & 0 \\ 0 & T \end{bmatrix} \quad \text{and} \quad V_H = \begin{bmatrix} Y & 0 \\ T/2 & T \end{bmatrix}. \tag{7}$$

Both lattices have the same sampling density, with $d(\Lambda_R) = d(\Lambda_H) = YT$. Figure 3 shows the reciprocal lattices Λ_R^* and Λ_H^* with several possible unit cells.

3 Sampling and Reconstruction of Continuous Time-Varying Imagery

The process for sampling a time-varying image can be approximated by the system shown in Fig. 4. The light arriving on the sensor is collected and weighted in space and time by the sensor aperture $a(\boldsymbol{s})$ to give the output

$$f_{ca}(\boldsymbol{s}) = \int_{\mathbb{R}^3} f_c(\boldsymbol{s}+\boldsymbol{s}')a(\boldsymbol{s}')d\boldsymbol{s}' \tag{8}$$

where it is assumed here that the sensor aperture is space and time invariant. The resulting signal $f_{ca}(\boldsymbol{s})$ is then sampled in an ideal fashion on the sampling structure Ψ,

$$f(\boldsymbol{s}) = f_{ca}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Psi. \tag{9}$$

By defining $h_a(\boldsymbol{s}) = a(-\boldsymbol{s})$, it is seen that the aperture weighting is a linear shift-invariant filtering operation, i.e., the convolution of $f_c(\boldsymbol{s})$ with $h_a(\boldsymbol{s})$

$$f_{ca}(\boldsymbol{s}) = \int_{\mathbb{R}^3} f_c(\boldsymbol{s}-\boldsymbol{s}')h_a(\boldsymbol{s}')d\boldsymbol{s}'. \tag{10}$$

Thus, if $f_c(\boldsymbol{s})$ has a Fourier transform $F_c(\boldsymbol{u})$, then $F_{ca}(\boldsymbol{u}) = F_c(\boldsymbol{u})H_a(\boldsymbol{u})$, where $H_a(\boldsymbol{u})$ is the Fourier transform of the

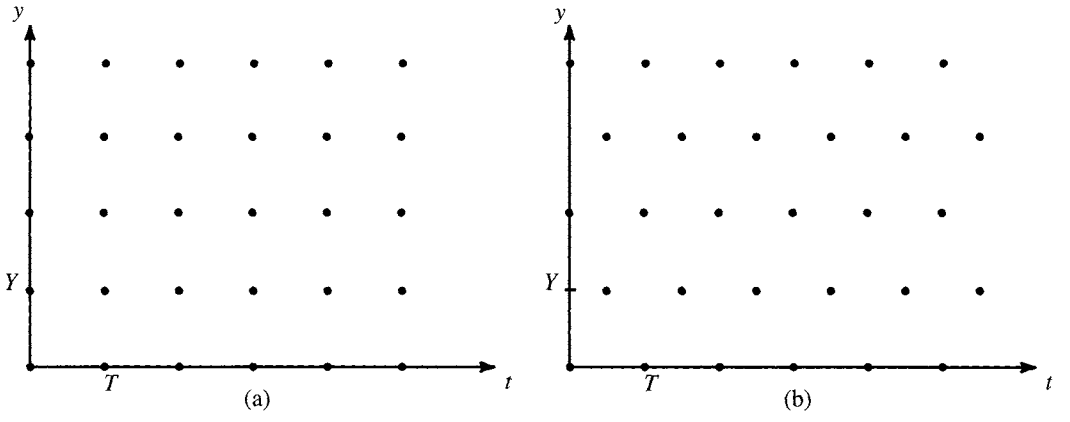

FIGURE 2 Two-dimensional vertical-temporal lattices. (a) Rectangular lattice Λ_R. (b) Hexagonal lattice Λ_H.

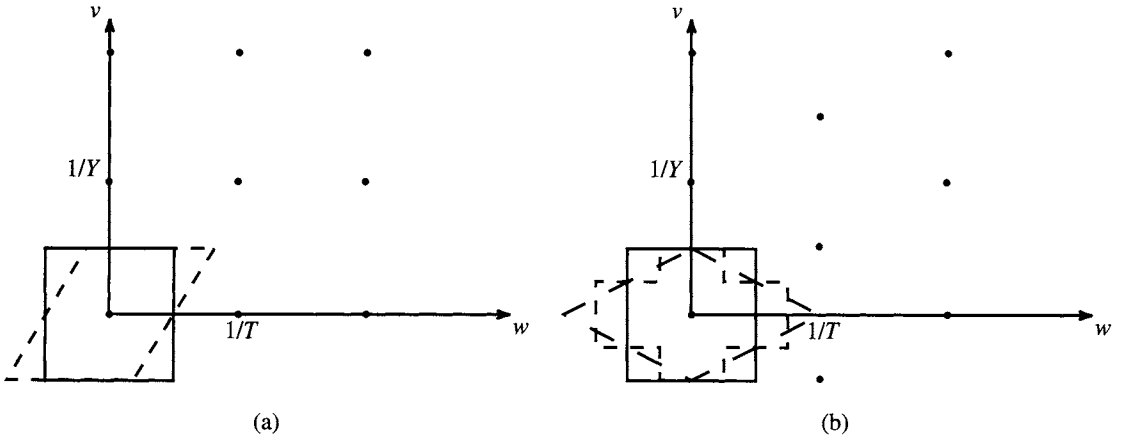

FIGURE 3 Reciprocal lattices of the two-dimensional vertical-temporal lattices of Fig. 2 with several possible unit cells. (a) Rectangular lattice Λ_R^* (b) Hexagonal lattice Λ_H^*.

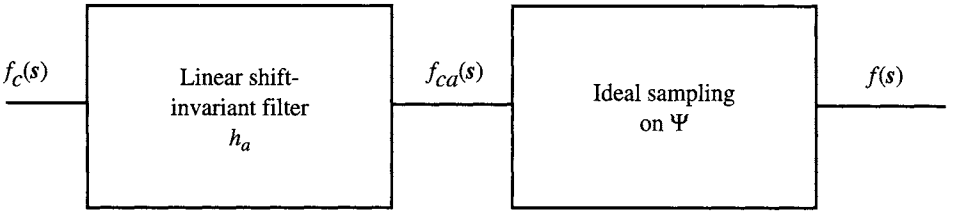

FIGURE 4 System for sampling a time-varying image.

aperture impulse response. In typical acquisition systems, the sampling aperture can be modelled as a rectangular or Gaussian function.

If the sampling structure is a lattice Λ, then the effect in the frequency domain of the sampling operation is given by [1]

$$F(\boldsymbol{u}) = \frac{1}{d(\Lambda)} \sum_{\boldsymbol{k} \in \Lambda^*} F_{ca}(\boldsymbol{u} + \boldsymbol{k}), \tag{11}$$

in other words, the continuous signal spectrum $F_{ca}(\boldsymbol{u})$ is replicated on the points of the reciprocal lattice. The terms in the sum of Eq. (11) other than for $\boldsymbol{k} = \boldsymbol{0}$ are referred to as *spectral repeats*. There are two main consequences of the sampling process. The first is that these spectral repeats, if not removed by the display/viewer system, may be visible in the form of flicker, line structure or dot patterns. The second is that if the regions of support of $F_{ca}(\boldsymbol{u})$ and $F_{ca}(\boldsymbol{u} + \boldsymbol{k})$ have non-zero intersection for some values $\boldsymbol{k} \in \Lambda^*$, we have aliasing; a frequency $\boldsymbol{u}_a$ in this intersection can represent both the frequencies $\boldsymbol{u}_a$ and $\boldsymbol{u}_a - \boldsymbol{k}$ in the original signal. Thus, to avoid aliasing, the spectrum $F_{ca}(\boldsymbol{u})$ should be confined to a unit cell of Λ^*; this can accomplished to some extent by the sampling aperture h_a. Aliasing is particularly problematic because once introduced it is difficult to remove, since there is more than one acceptable interpretation of the observed data. Aliasing is a familiar effect that tends to be localized to those regions of the image with high frequency details. It can be seen as moiré patterns in such periodic-like patterns as fishnets and venetian blinds, and as staircase-like effects on high-contrast oblique edges. The aliasing is particularly visible and annoying when these patterns are moving. Aliasing is controlled by selecting a sufficiently dense sampling structure and through the prefiltering effect of the sampling aperture.

If the support of $F_{ca}(\boldsymbol{u})$ is confined to a unit cell $\mathcal{P}^*$ of Λ^*, then it is possible to reconstruct f_{ca} exactly from the samples. In this case, we have

$$F_{ca}(\boldsymbol{u}) = \begin{cases} d(\Lambda) F(\boldsymbol{u}) & \text{if } \boldsymbol{u} \in \mathcal{P}^* \\ 0 & \text{if } \boldsymbol{u} \notin \mathcal{P}^* \end{cases} \tag{12}$$

and it follows that

$$f_{ca}(\boldsymbol{s}) = \sum_{\boldsymbol{s}' \in \Lambda} f(\boldsymbol{s}') t(\boldsymbol{s} - \boldsymbol{s}') \tag{13}$$

where

$$t(\boldsymbol{s}) = d(\Lambda) \int_{\mathcal{P}^*} \exp(j2\pi \boldsymbol{u} \cdot \boldsymbol{s}) d\boldsymbol{u} \tag{14}$$

is the impulse response of an ideal lowpass filter (with sampled input and continuous output) having passband $\mathcal{P}^*$. This is the multidimensional version of the familiar sampling theorem.

In practical systems, the reconstruction is achieved by

$$\hat{f}_{ca}(\boldsymbol{s}) = \sum_{\boldsymbol{s}' \in \Lambda} f(\boldsymbol{s}')d(\boldsymbol{s} - \boldsymbol{s}') \tag{15}$$

where $d(\boldsymbol{s})$ is the display aperture, which generally bears little resemblance to the ideal $t(\boldsymbol{s})$ of Eq. (14). The display aperture is usually separable in space and time, $d(\boldsymbol{s}) = d_s(x, y)d_t(t)$, where $d_s(x, y)$ may be Gaussian or rectangular, and $d_t(t)$ may be exponential or rectangular, depending on the type of display system. In fact, a large part of the reconstruction filtering is often left to the spatiotemporal response of the human visual system. The main requirement is that the first temporal frequency repeat at zero spatial frequency (at $1/T$ for progressive scanning and $2/T$ for interlaced scanning, Fig. 2) be at least 50 Hz for large area flicker to be acceptably low.

If the display aperture is the ideal lowpass filter specified by Eq. (14), then the optimal sampling aperture is also an ideal lowpass filter with passband $\mathcal{P}^*$; neither of these are realizable in practice. If the actual aperture of a given display device operating on a lattice Λ is given, it is possible to determine the optimal sampling aperture according to a weighted-squared-error criterion [3]. This optimal sampling aperture, which will not be an ideal lowpass filter, is similarly not physically realizable, but it could at least form the design objective rather than the inappropriate ideal lowpass filter.

If sampling is performed in only one or two dimensions, the spectrum is replicated in the corresponding frequency dimensions. For the two cases of temporal and vertical-temporal sampling respectively, we obtain

$$F(\boldsymbol{u}, w) = \frac{1}{T} \sum_{l=-\infty}^{\infty} F_{ca}\left(\boldsymbol{u}, w + \frac{l}{T}\right) \tag{16}$$

and

$$F(u, v, w) = \frac{1}{d(\Lambda_{yt})} \sum_{\boldsymbol{k} \in \Lambda_{yt}^*} F_{ca}(u, (v, w) + \boldsymbol{k}). \tag{17}$$

Consider first the case of pure temporal sampling, as in motion-picture film. The main parameters in this case are the sampling period T and the temporal aperture. As shown in Eq. (16), the signal spectrum is replicated in temporal frequency at multiples if $1/T$. In analogy with one-dimensional signals, one might think that the time-varying image should be bandlimited in temporal frequency to $1/2T$ before sampling. However, this is not the case. To illustrate, consider the spectrum of an image undergoing translation with constant velocity $\boldsymbol{v}$. This can model the local behavior in a large class of time-varying imagery. The assumption implies that $f_c(\boldsymbol{x}, t) = f_{c0}(\boldsymbol{x} - \boldsymbol{v}t)$, where $f_{c0}(\boldsymbol{x}) = f_c(\boldsymbol{x}, 0)$. A straightforward analysis [4] shows that $F_c(\boldsymbol{u}, w) = F_{c0}(\boldsymbol{u})\delta(\boldsymbol{u} \cdot \boldsymbol{v} + w)$, where $\delta(\cdot)$ is the Dirac delta function. Thus, the spectrum of the time-varying image is not spread throughout spatiotemporal frequency space but rather it is concentrated around the plane $\boldsymbol{u} \cdot \boldsymbol{v} + w = 0$. When this translating image is sampled in the temporal dimension, these planes are parallel to each other and do not intersect, i.e., there is no aliasing, even if the temporal bandwidth far exceeds $1/2T$. This is most easily illustrated in two dimensions. Consider the case of vertical motion only. Figure 5 shows the

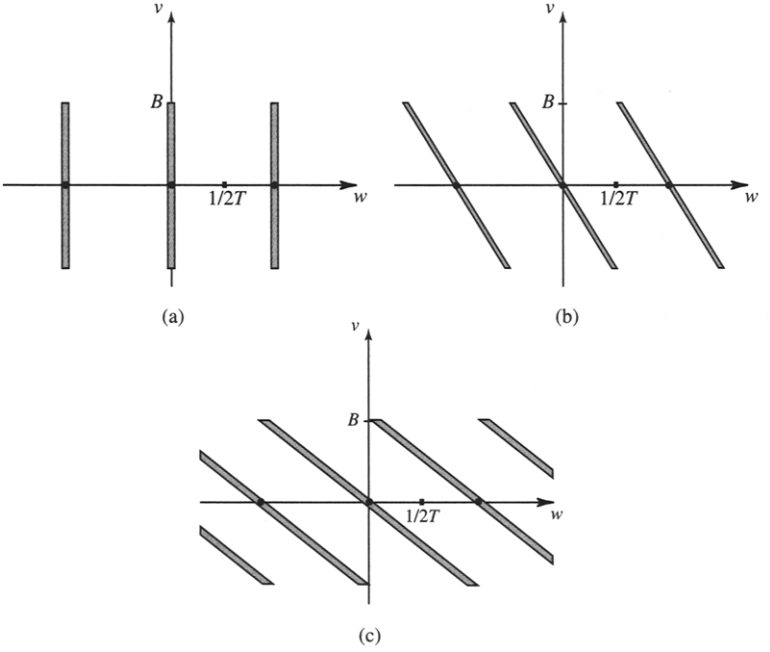

FIGURE 5 Vertical-temporal projection of the spectrum of temporally sample time-varying image with vertical motion of velocity v. (a) $v = 0$. (b) $v = 1/2TB$. (c) $v = 1/TB$.

vertical-temporal projection of the spectrum of the sampled image for different velocities v. Assume that the image is vertically bandlimited to B c/ph. It follows that when the vertical velocity reaches $1/2TB$ picture heights per second (ph/s), the spectrum will extend out to the temporal frequency of $1/2T$ as shown in Fig. 5(b). At twice that velocity ($1/TB$), it would extend to a temporal frequency of $1/T$ which might suggest severe aliasing. However, as seen in Fig. 5(c), there is no spectral overlap. To reconstruct the continuous signal correctly however, a vertical-temporal filtering adapted to the velocity is required. Bandlimiting the signal to a temporal frequency of $1/2T$ before sampling would effectively cut the vertical resolution in half for this velocity. Note that the velocities mentioned above are not really very high. To consider some typical numbers, if $T = 1/24$ s, as in film, and $B = 500$ c/ph (corresponding to 1,000 scanning lines) the velocity $1/2TB$ is about 1/42 ph/s. It should be noted that if the viewer is tracking the vertical movement, the spectrum of the image on the retina will be far less tilted, again arguing against sharp temporal bandlimiting. (This is in fact a kind of motion-compensated filtering by the visual system.) The temporal camera aperture can roughly be modeled as the integration of f_c for a period $T_a \leq T$. The choice of the value of the parameter T_a is a compromise between motion blur and signal-to-noise ratio.

Similar arguments can be made in the case of the two most popular vertical-temporal scanning structures, progressive scanning and interlaced scanning. Referring to Fig. 6, the vertical temporal spectrum of a vertically translating image at the same three velocities (assuming that $1/Y = 2B$) is shown for these two scanning structures. For progressive scanning there continues to be no spectral overlap, while for interlaced scanning the spectral overlap can be severe at certain velocities (e.g., $1/TB$ as in Fig. 6(f)). This is a strong advantage for progressive scanning. Another disadvantage of interlaced scanning is that each field is spatially undersampled and pure spatial processing or interpolation is very difficult. An illustration in three dimensions of some of these ideas can be found in [5].

4 Sampling Structure Conversion

There are numerous spatiotemporal sampling structures used for the digital representation of time-varying imagery. However, the vast majority of those in use fall into one of two categories corresponding to progressive or interlaced scanning with aligned horizontal sampling. This corresponds to sampling matrices of the form

$$\begin{bmatrix} X & 0 & 0 \\ 0 & Y & 0 \\ 0 & 0 & T \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} X & 0 & 0 \\ 0 & Y & 0 \\ 0 & T/2 & T \end{bmatrix}$$

respectively. A three-dimensional view of these two sampling lattices is shown in Fig. 7. It can be observed how the odd numbered horizontal lines in each frame from the progressive lattice (y/Y odd) in Fig. 7(a) have been delayed temporally by $T/2$ for the interlaced lattice of Fig. 7(b).

Table 1 shows the parameters for a number of commonly used sampling structures covering a broad range of applications from low-resolution QCIF used in videophone to HDTV and digitized IMAX film (the popular large-format film, about 70 mm by 52 mm, used by Imax Corporation). Note that of these, only HDTV and IMAX formats have $X = Y$ (i.e., square pixels). It is frequently required to convert a time-varying image sampled on one such structure to another. An input image sequence $f(\boldsymbol{x})$ sampled on lattice Λ_1 is to be converted to the output sequence $f_o(\boldsymbol{x})$ sampled on the lattice Λ_2. This is illustrated in Fig. 8. The continuous signal $f_c(\boldsymbol{x})$ is acquired on the lattice Λ_1 using a physical camera modelled as in Fig. 4 with impulse response $h_a(\boldsymbol{x})$ to yield $f(\boldsymbol{x})$. It is desired to estimate the signal $f_o(\boldsymbol{x})$ that would have been obtained if $f_c(\boldsymbol{x})$ was sampled on the lattice Λ_2 with an ideal or theoretical camera having impulse response $h_{oa}(\boldsymbol{x})$. Note that since this camera is *theoretical*, the impulse response $h_{oa}(\boldsymbol{x})$ does not have to be realizable with any particular technology. It can be optimized to give the best displayed image on Λ_2 [3]. A system $\mathcal{H}$, which can be linear or nonlinear, is then required to estimate $f_o(\boldsymbol{x})$ from $f(\boldsymbol{x})$.

Besides converting between different standards, sampling structure conversion can also be incorporated into the acquisition or display portions of an imaging system to compensate for the difficulty in performing adequate prefiltering with the camera aperture, or adequate postfiltering with the display aperture. Specifically, the time-varying image can initially be sampled at a higher density than required, using the camera aperture as prefilter, and then downsampled to the desired structure using digital prefiltering, which offers much more flexibility. Similarly, the image can be upsampled for the display device using digital filtering, so that the subsequent display aperture has a less critical task to perform.

4.1 Frame-Rate Conversion

Consider first the case of pure frame-rate conversion. This applies when both the input and the output sampling structures are separable in space and time with the same spatial sampling structure, and where spatial aliasing is assumed to be negligible. The temporal sampling period is to be changed from T_1 to T_2. This situation corresponds to input and output sampling lattices

$$\Lambda_1 = \begin{bmatrix} v_{11} & v_{12} & 0 \\ 0 & v_{22} & 0 \\ 0 & 0 & T_1 \end{bmatrix}, \quad \Lambda_2 = \begin{bmatrix} v_{11} & v_{12} & 0 \\ 0 & v_{22} & 0 \\ 0 & 0 & T_2 \end{bmatrix}. \tag{18}$$

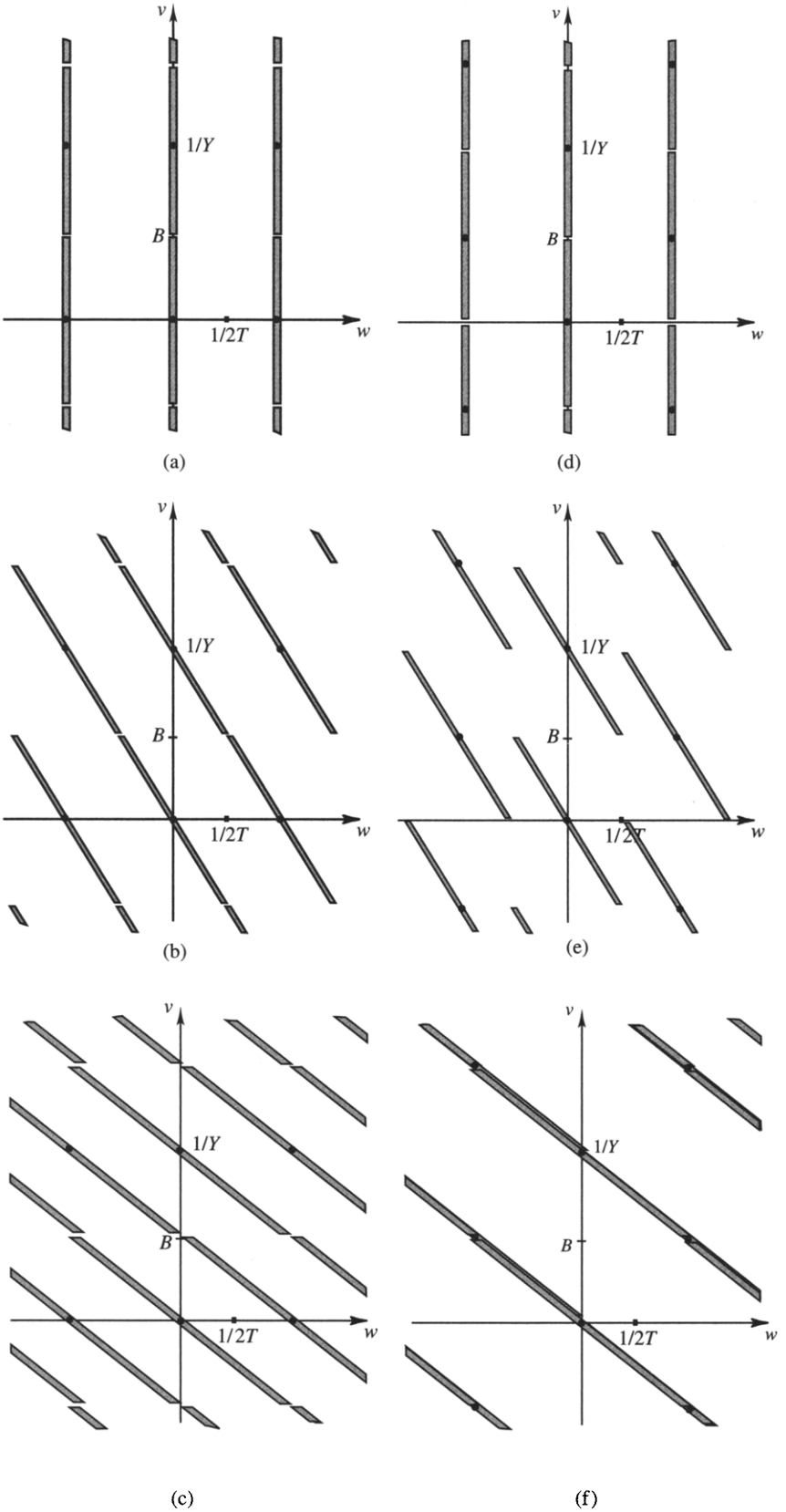

FIGURE 6 Vertical-temporal projection of spectrum of vertical-temporal sampled time-varying image with progressive and interlaced scanning. Progressive: (a) $\nu = 0$. (b) $\nu = 1/2TB$. (c) $\nu = 1/TB$. Interlaced: (d) $\nu = 0$. (e) $\nu = 1/2TB$. (f) $\nu = 1/TB$.

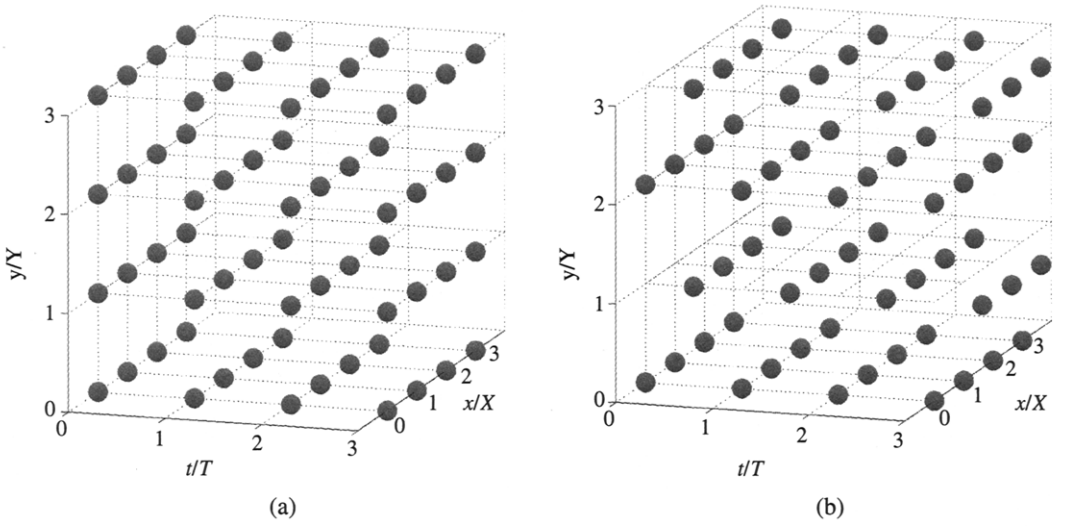

FIGURE 7 Three-dimensional view of spatiotemporal sampling lattices. (a) Progressive. (b) Interlaced.

TABLE 1 Parameters of several common spatiotemporal sampling structures. P indicates progressive scanning and I indicates interlaced scanning

System	X	Y	T	Structure	Aspect ratio
QCIF	$\frac{1}{176}$ pw$=\frac{1}{132}$ ph	$\frac{1}{144}$ ph	$\frac{1}{10}$ s	P	4:3
CIF	$\frac{1}{352}$ pw$=\frac{1}{264}$ ph	$\frac{1}{288}$ ph	$\frac{1}{15}$ s	P	4:3
ITU-R-601 (30)	$\frac{1}{720}$ pw$=\frac{1}{540}$ ph	$\frac{1}{480}$ ph	$\frac{1}{29.97}$ s	I	4:3
ITU-R-601 (25)	$\frac{1}{720}$ pw$=\frac{1}{540}$ ph	$\frac{1}{576}$ ph	$\frac{1}{25}$ s	I	4:3
HDTV-P	$\frac{1}{1280}$ pw$=\frac{1}{720}$ ph	$\frac{1}{720}$ ph	$\frac{1}{60}$ s	P	16:9
HDTV-I	$\frac{1}{1920}$ pw$=\frac{1}{1080}$ ph	$\frac{1}{1080}$ ph	$\frac{1}{30}$ s	I	16:9
IMAX	$\frac{1}{4096}$ pw$=\frac{1}{3002}$ ph	$\frac{1}{3002}$ ph	$\frac{1}{24}$ s	P	1.364

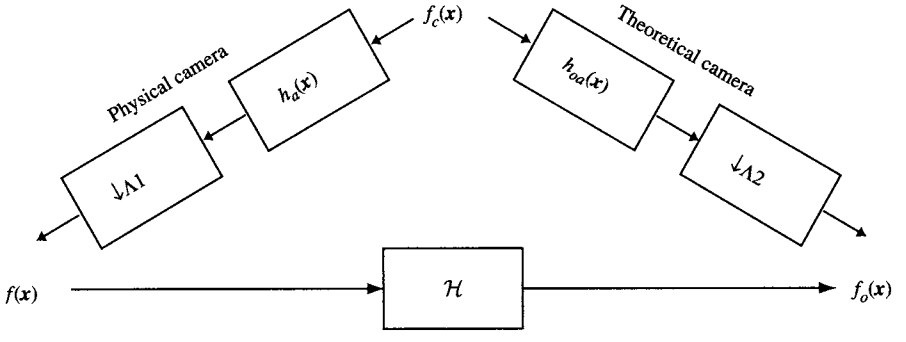

FIGURE 8 Acquisition models for the observed signal $f(\boldsymbol{x})$ on Λ_1 and the desired output signal $f_0(\boldsymbol{x})$ on Λ_2.

Pure Temporal Interpolation

The most straightforward approach is pure temporal interpolation, where a temporal resampling is performed independently at each spatial location $\boldsymbol{x}$. A typical application for this is increasing the frame rate in motion picture film from 24 frames/s to 48 or 60 frames/s, giving significantly better motion rendition. Using linear filtering, the interpolated image sequence is given by

$$f_o(\boldsymbol{x}, nT_2) = \sum_m f(\boldsymbol{x}, mT_1)h(nT_2 - mT_1). \tag{19}$$

If the temporal spectrum of the underlying continuous time-varying image satisfies the Nyquist criterion, the output points can be computed by ideal *sinc* interpolation:

$$h(t) = \frac{\sin(\pi t/T_1)}{\pi t/T_1}. \tag{20}$$

However, aside from the fact that this filter is unrealizable, it is unlikely, and in fact undesirable according to the discussion of Section 3, for the temporal spectrum to satisfy the Nyquist criterion. Thus high order interpolation kernels that approximate Eq. (20) are not found to be useful and are rarely used. Instead, simple low-order interpolation kernels are frequently applied. Examples are zero-order and linear (straight-line) interpolation kernels given by

$$h(t) = \begin{cases} 1 & \text{if } 0 \le t \le T_1 \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

and

$$h(t) = \begin{cases} 1 - |t|/T_1 & \text{if } 0 \le |t| \le T_1 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

respectively. Note that Eq. (22) defines a non-casual filter and that in practice a delay of T_1 must be introduced. Zero-order hold is also called frame repeat and is the method used in film projection to go from 24 to 48 frames/s. These simple interpolators work well if there is little or no motion, but as the amount of motion increases they will not adequately remove spectral repeats causing effects such as jerkiness, and they may also remove useful information, introducing blurring. The problems with pure temporal interpolation can easily be illustrated for the image corresponding to Fig. 5(c) for the case of doubling the frame rate, i.e., $T_2 = T_1/2$. Using a one-dimensional temporal lowpass filter with cutoff at about $1/2T_1$ removes the desired high vertical frequencies in the baseband signal above $B/2$ (motion blur) and leaves undesirable aliasing at high vertical frequencies, as shown in Fig. 9(a).

Motion-Compensated Interpolation

It is clear that to correctly deal with situation such as in Fig. 4(c), it is necessary to adapt the interpolation to the local orientation of the spectrum, and thus to the velocity, as suggested in Fig. 9(b). This is called motion-compensated interpolation. An auxiliary motion analysis process determines information about local motion in the image and attempts to track the trajectory of scene points over time. Specifically, suppose we wish to estimate the signal value at position $\boldsymbol{x}$ at time nT_2 from neighboring frames at times mT_1. We can assume that the scene point imaged at position $\boldsymbol{x}$ at time nT_2 was imaged at position $\boldsymbol{c}(mT_1; \boldsymbol{x}, nT_2)$ at time mT_1 [6]. If we know $\boldsymbol{c}$ exactly, we can compute

$$f_o(\boldsymbol{x}, nT_2) = \sum_m f(\boldsymbol{c}(mT_1;\ \boldsymbol{x},\ nT_2),\ mT_1)h(nT_2 - mT_1). \tag{23}$$

Since we assume that $f(\boldsymbol{x}, t)$ is very slowly varying along the motion trajectory, a simple filter such as the linear interpolator of Eq. (22) would probably do very well. Of course, we do not know $\boldsymbol{c}(mT_1; \boldsymbol{x}, nT_2)$ so we must estimate it. Furthermore, since the position $(\boldsymbol{c}(mT_1;\ \boldsymbol{x},\ nT_2),\ mT_1)$ probably does not lie on the input lattice Λ_1, $f(\boldsymbol{c}(mT_1;\ \boldsymbol{x},\ nT_2),\ mT_1)$ must be spatially interpolated from its neighbors.

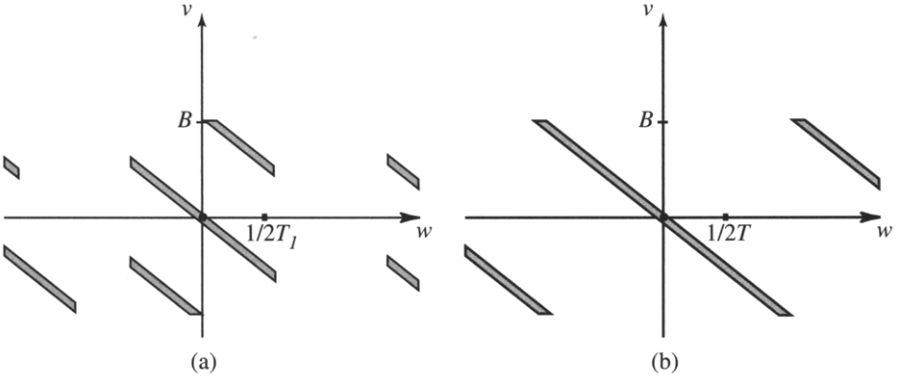

FIGURE 9 Frequency domain interpretation of 2:1 temporal interpolation of an image with vertical velocity $1/TB$. (a) Pure temporal interpolation. (b) Motion-compensated interpolation.

If spatial aliasing is low as we have assumed, this interpolation can be done well (see chapter 7.1).

If a two-point temporal interpolation is used, we only need to find the correspondence between the point at $(\boldsymbol{x}, nT_2)$ and points in the frames at times lT_1 and $(l+1)T_1$ where $lT_1 \leq nT_2$ and $(l+1)T_1 > nT_2$. This is specified by the backward and forward displacements

$$\boldsymbol{d}_b(\boldsymbol{x}, nT_2) = \boldsymbol{x} - \boldsymbol{c}(lT_1; \boldsymbol{x}, nT_2) \tag{24}$$

$$\boldsymbol{d}_f(\boldsymbol{x}, nT_2) = \boldsymbol{c}((l+1)T_1; \boldsymbol{x}, nT_2) - \boldsymbol{x} \tag{25}$$

respectively. The interpolated value is then given by

$$\begin{aligned} f_o(\boldsymbol{x}, nT_2) = {} & f(\boldsymbol{x} - \boldsymbol{d}_b(\boldsymbol{x}, nT_2), lT_1)h(nT_2 - lT_1) \\ & + f(\boldsymbol{x} + \boldsymbol{d}_f(\boldsymbol{x}, nT_2), (l+1)T_1)h(nT_2 - (l+1)T_1). \end{aligned} \tag{26}$$

There are a number of key design issues in this process. The main one relates to the complexity and precision of the motion estimator. Since the image at time nT_2 is not available, the trajectory must be estimated from the existing frames at times mT_1, and often just from lT_1 and $(l+1)T_1$ as defined above. In the latter case, the forward and backward displacements will be collinear. We can assume that better motion estimators will lead to better motion-compensated interpolation. However, the tradeoff between complexity and performance must be optimized for each particular application. For example, block-based motion estimation (say one motion vector per 16×16 block) with accuracy rounded to the nearest pixel location will give very good results in large moving areas with moderate detail, giving significant overall improvement for most sequences. However, areas with complex motion and higher detail may continue to show quite visible artifacts, and more accurate motion estimates would be required to get good performance in these areas. Better motion estimates could be achieved with smaller blocks, parametric motion models, or dense motion estimates, for example. Motion estimation is treated in detail in Chapter 3.10. Some specific considerations related to estimating trajectories passing through points in between frames in the input sequence can be found in [6].

If the motion estimation method used sometimes yields unreliable motion vectors, it may be advantageous to be able to fall back to pure temporal interpolation. A test can be performed to determine whether pure temporal interpolation or motion-compensated interpolation is expected to yield better results, for example by comparing $|f(\boldsymbol{x}, (l+1)T_1) - f(\boldsymbol{x}, lT_1)|$ with $f(\boldsymbol{x} + \boldsymbol{d}_f(\boldsymbol{x}, nT_2), (l+1)T_1) - f(\boldsymbol{x} - \boldsymbol{d}_b(\boldsymbol{x}, nT_2), lT_1)|$ either at a single point or over a small window. Then the interpolated value can either be computed by the method suspected to be better, or by an appropriate weighted combination of the two.

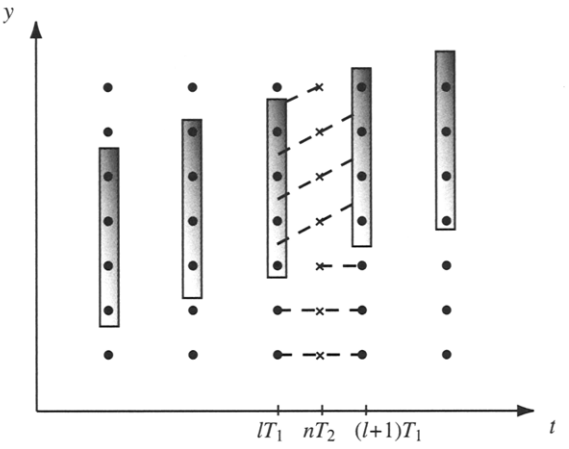

FIGURE 10 Example of motion-compensated temporal interpolation including occlusion handling.

Occlusions pose a particular problem, since the pixel to be interpolated may be visible only in the previous frame (newly covered area) or in the subsequent frame (newly exposed area). In particular, if $|f(\boldsymbol{x}+\boldsymbol{d}_f(\boldsymbol{x}, nT_2), (l+1)T_1) - f(\boldsymbol{x} - \boldsymbol{d}_b)(\boldsymbol{x}, nT_2), lT_1)|$ is relatively large, this may signal that $\boldsymbol{x}$ lies in an occlusion area. In this case, we may wish to use zero-order hold interpolation based on either the frame at lT_1 or at $(l+1)T_1$, according to some local analysis. Figure 10 depicts the motion-compensated interpolation of a frame midway between lT_1 and $(l+1)T_1$ including occlusion processing, where we assume that a single object is moving upward.

4.2 Spatiotemporal Sampling Structure Conversion

In this section, we consider the case where both the spatial and the temporal sampling structures are changed, and when one or both of the input and output sampling structures is not separable in space and time (usually because of interlace). If the input sampling structure Λ_1 is separable in space and time (as in Eq. (18)) and spatial aliasing is minimal, then the methods of the previous section can be combined with pure spatial interpolation. If we want to interpolate a sample at a time mT_1, we can use any suitable spatial interpolation. To interpolate at a sample at a time t that is not a multiple of T_1, the methods of the previous section can be applied.

The difficulties in spatiotemporal interpolation mainly arise when the input sampling structure Λ_1 is not separable in space and time, which is generally the case of interlace. This encompasses both interlaced-to-interlaced conversion, such as in conversion between NTSC and PAL television systems, and interlaced-to-progressive conversion (also called deinterlacing). The reason this introduces problems is that individual fields are undersampled, contrary to the

assumption in all the previously discussed methods. Furthermore, as we have seen, there may also be significant aliasing in the spatiotemporal frequency domain due to vertical motion. Thus, a great deal of the research on spatiotemporal interpolation has been addressing these problems due to interlace, and a wide variety of techniques have been proposed, many of them very empirical in nature.

Deinterlacing

Deinterlacing generally refers to a 2:1 interpolation from an interlaced grid to a progressive grid with sampling lattices

$$\begin{bmatrix} X & 0 & 0 \\ 0 & Y & 0 \\ 0 & T/2 & T \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} X & 0 & 0 \\ 0 & Y & 0 \\ 0 & 0 & T/2 \end{bmatrix}$$

respectively (see Fig. 11). Both input and output lattices consist of fields at time instants $mT/2$. However, because each input field is vertically undersampled, spatial interpolation alone is inadequate. Similarly, because of possible spatiotemporal aliasing and difficulties with motion estimation, motion-compensated interpolation alone is inadequate. Thus, the most successful methods use a nonlinear combination of spatially and temporally interpolated values, according to local measures of which is most reliable. For example, in Fig. 11, sample A might best be reconstructed using spatial interpolation, sample B with pure temporal interpolation and sample C with motion-compensated temporal interpolation. Another sample like D may be reconstructed using a combination of spatial and motion-compensated temporal interpolation. See [7] for a detailed presentation and discussion of a wide variety of deinterlacing methods. It is shown there that some adaptive motion-compensated methods can give reasonably good deinterlacing results on a wide variety of moving and fixed imagery.

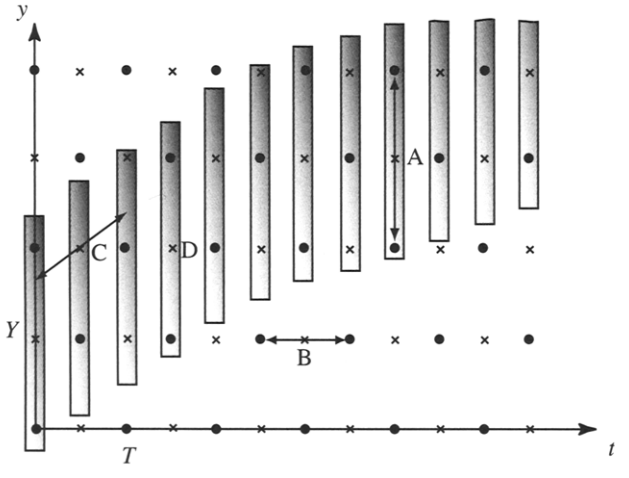

FIGURE 11 Input and output sampling structures for deinterlacing.

5 Conclusion

This chapter has provided an overview of the basic theory related to sampling and interpolation of time-varying imagery. In contrast to other types of signals, it has been shown that it is *not* desirable to limit the spectrum of the continuous signal to a *fixed* three-dimensional frequency band prior to sampling, since this leads to excessive loss of spatial resolution. It is sufficient to ensure that the replicated spectra due to sampling do not overlap. However, optimal reconstruction requires the use of motion-compensated temporal interpolation.

The interlaced scanning structure that is widely used in video systems has a fundamental problem whereby aliasing in the presence of vertical motion is inevitable. This makes operations such as motion estimation, coding and so on more difficult to accomplish. Thus, it is likely that interlaced scanning will gradually disappear as camera technology improves and the full spatial resolution desired can be obtained with frame rates of 50–60 Hz and above.

Spatiotemporal interpolation will remain an important technology to convert between the wide variety of scanning standards in both new and archival material. Research will continue into robust, low-complexity methods for motion-compensated temporal interpolation that can be incorporated into any receiver. Further work is also required to fully exploit the model of Fig. 8 or similar models in the video sampling structure conversion problem in ways similar to what has been done for still images [8].

References

[1] E. Dubois, "The sampling and reconstruction of time-varying imagery with application in video systems," *Proc. IEEE*, 73, 502–522, Apr. 1985.

[2] T. Kalker, "On multidimensional sampling," in *The Digital Signal Processing Handbook* (V. Madisetti and D. Williams, eds.), ch. 4, 4-1–4-21, CRC Press, 1998.

[3] H. Aly and E. Dubois, "Design of optimal camera apertures adapted to display devices over arbitary sampling lattices," *IEEE Signal Process. Lett.*, 11, 443–445, Apr. 2004.

[4] E. Dubois, "Motion-compensated filtering of time-varying images," *Multidimens. Syst. Signal Process.*, 3, 211–239, 1992.

[5] B. Girod and R. Thoma, "Motion-compensating field interpolation from interlaced and non-interlaced grids," in *Proc. SPIE Image Coding*, 594, 186–193, 1985.

[6] E. Dubois and J. Konrad, "Estimation of 2-D motion fields from image sequences with application to motion-compensated processing," in *Motion Analysis and Image Sequence Processing* (M. Sezan and R. Lagendijk, eds.), ch. 3, 53–87, Kluwer Academic Publishers, 1993.

[7] G. de Haan, "Deinterlacing-an overview," *Proc. IEEE*, 86, 1839–1857, Sept. 1998.

[8] H. Aly and E. Dubois, "Image up-sampling using total-variation regularization with a new observation model," *IEEE Trans. Image Process.*, 2004, in press.

[9] P. Mertz and F. Gray, "A theory of scanning and its relation to the characteristics of the transmitted signal in telephotography and television," *Bell Syst. Tech. J.*, 13, 464–515, 1934.

[10] D. Petersen and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in n-dimensional euclidean spaces," *Informat. Contr.*, 5, 279–323, 1962.

[11] M. Isnardi, "Modelling the television process," Technical Report 515, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, May 1986.

[12] E. Dubois, G. de Haan, and T. Kurita, "Special issue on motion estimation and compensation technologies for standards conversion," *Signal Processing: Image Communication*, 6, June 1994.

Further Information

The classic paper on television scanning is [9]. The use of lattices for the study of spatiotemporal sampling was introduced in [10]. A detailed study of cameral and display aperture models for television can be found in [11]. Research papers on spatiotemporal interpolation can be found regularly in the IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology and Signal Processing: Image Communication. See [12] for a special issue on motion estimation and compensation for standards conversion.