APPENDIX

The sampling theorem

B

In vision, everything depends on image acquisition. In the modern age, image acquisition invariably involves sampling the underlying analog signals to convert them to digital form. Ultimately, sampling is subject to the mathematical rules that are embodied in the sampling theorem. These are vitally important if we are to obtain digital signals that accurately reflect the original analog ones. It is the purpose of this appendix to remind the reader about this fundamental process.

B.1 THE SAMPLING THEOREM

The Nyquist sampling theorem underlies all situations where continuous signals are sampled and is especially important where patterns are to be digitized and analyzed by computers. This makes it highly relevant both with visual patterns and with acoustic waveforms: hence it is described briefly in this section.

Consider the sampling theorem first in respect of a 1-D time-varying waveform. The theorem states that a sequence of samples (Fig. B.1) of such a waveform contains all the original information and can be used to regenerate the original waveform exactly, but only if (1) the bandwidth W of the original waveform is restricted and (2) the rate of sampling f is at least twice the bandwidth of the original waveform—i.e., $f \geq 2W$. Assuming that samples are taken every T seconds, this means that $1/T \geq 2W$.

At first, it may be somewhat surprising that the original waveform can be reconstructed exactly from a set of discrete samples. However, the two conditions for achieving this are very stringent. What they are demanding in effect is that the signal must not be permitted to change unpredictably (i.e., at too fast a rate) or else accurate interpolation between the samples will not prove possible (the errors that arise from this source are called “aliasing” errors).

Unfortunately, the first condition is virtually unrealizable, since it is close to impossible to devise a low-pass filter with a perfect cutoff. Recall from Chapter 3, Image Filtering and Morphology that a low-pass filter with a perfect cutoff will have infinite extent in the time domain, so any attempt at achieving

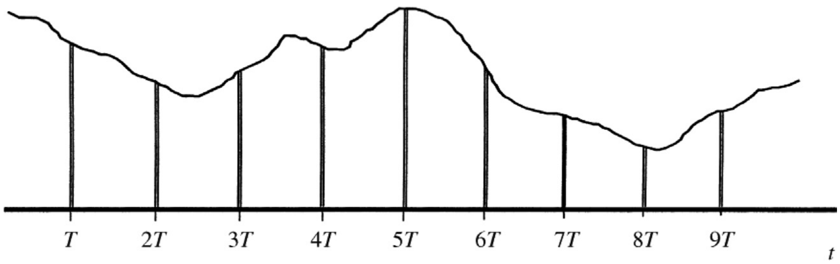

FIGURE B.1

The process of sampling a time-varying signal: a continuous time-varying 1-D signal is sampled by narrow sampling pulses at a regular rate $f_r = 1/T$ which must be at least twice the bandwidth of the signal.

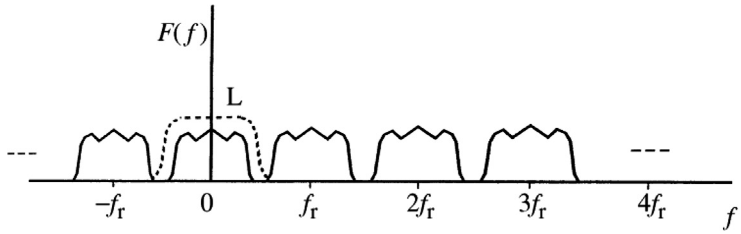

FIGURE B.2

Effect of low-pass filtering to eliminate repeated spectra in the frequency domain (f_r, sampling rate; L, low-pass filter characteristic). This diagram shows the repeated spectra of the frequency transform $F(f)$ of the original sampled waveform. It also demonstrates how a low-pass filter can be expected to eliminate the repeated spectra to recover the original waveform.

the same effect by time domain operations must be doomed to failure. However, acceptable approximations can be achieved by allowing a "guard-band" between the desired and actual cutoff frequencies. This means that the sampling rate must be higher than the Nyquist rate (in telecommunications, satisfactory operation can generally be achieved at sampling rates around 20% above the Nyquist rate—see Brown and Glazier, 1974).

One way of recovering the original waveform is by applying a low-pass filter. This approach is intuitively correct, since it acts in such a way as to broaden the narrow discrete samples until they coalesce and sum to give a continuous waveform. Indeed, this method acts in such a way as to eliminate the "repeated" spectra in the transform of the original sampled waveform (Fig. B.2): this in itself shows why the original waveform has to be narrow-banded before sampling—so that the repeated and basic spectra of the waveform do not cross over each other and become impossible to separate with a low-pass filter. The idea may be taken further because the Fourier transform of a square cutoff filter is the sinc (sin u/u)

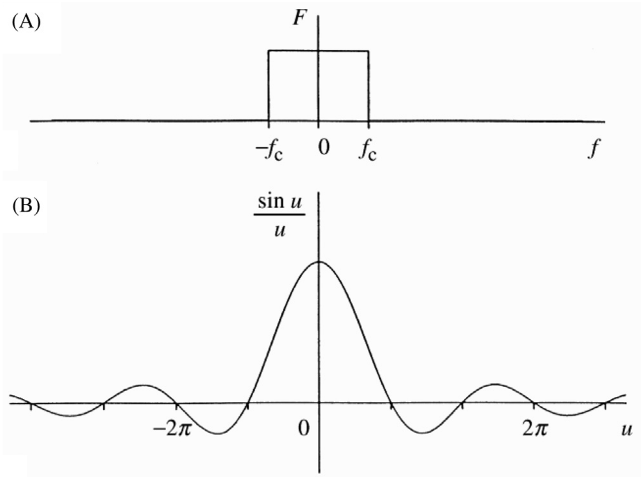

FIGURE B.3

The sinc (sin u/u) function shown in (B) is the Fourier transform of a square pulse (A) corresponding to an ideal low-pass filter. In this case, $u = 2\pi f_c t$, f_c being the cutoff frequency.

function (Fig. B.3). Hence, the original waveform may be recovered by convolving the samples with the sinc function (which in this case means replacing them by sinc functions of corresponding amplitudes). This has the effect of broadening out the samples as required, until the original waveform is recovered.

So far we have considered the situation only for 1-D time-varying signals. However, recalling that there is an exact mathematical correspondence between time and frequency domain signals on the one hand and spatial and spatial frequency signals on the other, the above ideas may all be applied immediately to each dimension of an image (although the condition for accurate sampling now becomes $1/X \geq 2W_X$, where X is the spatial sampling period and W_X is the spatial bandwidth). Here we accept this correspondence without further discussion and proceed to apply the sampling theorem to image acquisition.

Consider next how the signal from a camera may be sampled rigorously according to the sampling theorem. First, note that this has to be achieved both horizontally and vertically. Perhaps the most obvious solution to this problem is to perform the process optically, perhaps by defocussing the lens; however, the optical transform function for this case is frequently (i.e., for extreme cases of defocussing) very odd, going negative for some spatial frequencies and causing contrast reversals; hence this solution is far from ideal (Pratt, 2001). Alternatively, we could use a diffraction-limited optical system or perhaps pass the focused beam through some sort of patterned or frosted glass to reduce the spatial bandwidth artificially. None of these techniques will be particularly easy to apply nor (apart possibly from the second) will it give accurate solutions. However, this problem is not as serious

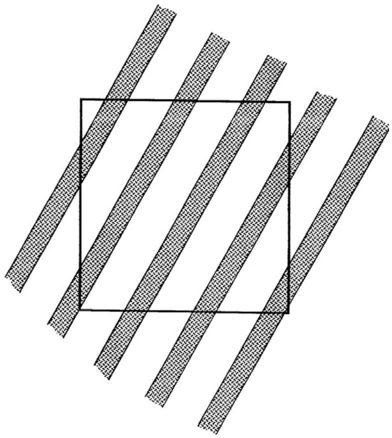

FIGURE B.4

Low-pass filtering carried out by averaging over the pixel region: an image with local high-frequency banding is to be averaged over the whole pixel region by the action of the sensing device.

as might be imagined. If the sensing region of the camera (per pixel) is reasonably large, and close to the size of a pixel, then the averaging inherent in obtaining the pixel intensities will in fact perform the necessary narrow-banding (Fig. B.4). To analyze the situation in more detail, note that a pixel is essentially square with a sharp cutoff at its borders: thus its spatial frequency pattern is a 2-D sinc function, which (taking the central positive peak) approximates to a low-pass spatial frequency filter: This approximation improves somewhat as the border between pixels becomes more fuzzy.

The point here is that the worst case from the point of view of the sampling theorem is that of extremely narrow discrete samples, but clearly this worst case is most unlikely to occur with most cameras. However, this does not mean that sampling is automatically ideal—and indeed it is not, since the spatial frequency pattern for a sharply defined pixel shape has (in principle) infinite extent in the spatial frequency domain. The review by Pratt (2001) clarifies the situation and shows that there is a tradeoff between aliasing and resolution error. Overall, quality of sampling will be one of the limiting factors if the greatest precision in image measurement is aimed for: if the bandwidth of the presampling filter is too low, resolution will be lost; if it is too high, aliasing distortions will creep in; and if its spatial frequency response curve is not suitably smooth, a guard band will have to be included and performance will again suffer.

The sampling theorem is well-covered in very many books on signal processing (see for example, Rosie, 1966), though details of how band-limiting should be carried out prior to sampling are not so readily available. See Pratt (2001) for further information and references about sampling in the imaging context.