Least-Squares Model-Based Video Halftoning *†

D. P. Hilgenberg T. J. Flohr C. B. Atkins J. P. Allebach C. A. Bouman

School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-1285

Abstract

A technique for halftoning video sequences is presented. First, an error metric is designed, incorporating psychophysical spatio-temporal contrast sensitivity data for the human visual system. An efficient strategy for producing halftone sequences that minimize the error is then developed. The technique is compared with video halftoning algorithms that do not consider the temporal aspect of the human visual system.

1 Introduction

The computational power of current computer workstations permits the real-time display of video sequences. However, due to memory restrictions, quantization is necessary for lengthy sequences. To minimize visible artifacts, it is desirable to incorporate a visual model in the quantization algorithm; this approach has proved to be successful in 2-dimensional image quantization. Pappas and Neuhoff [1], Analoui and Allebach [2], and Mulligan and Ahumada [3] have all achieved good results in model-based halftoning; the general color quantization case has been investigated by Kolpatzik and Bouman [5] and Flohr *et al.* [6]. Gotsman [7] extended model-based halftoning to video sequences, but did not explicitly incorporate a spatio-temporal visual model. Mulligan [8] used an *ad hoc* spatio-temporal model to perform video halftoning in the context of dither screens.

In this paper, we employ a linear, shift-invariant spatio-temporal visual model based on contrast sensitivity data due to Kelly [4]. With this model, we define a visually-weighted metric for the error between two image sequences. For this work, we restrict ourselves to binary quantization (halftoning) of achromatic sequences. We extend the "direct binary search" (DBS) minimization strategy [2, 9] to the temporal dimension in order to optimize halftone sequences with respect to the error metric. The computational demand presented by the quantities of data inherent in video precludes applying this search strategy to halftone image sequences in real time. Nonetheless, we are able to achieve good results in a reasonable amount of time, so that we can assess the merits of model-based video halftoning and compare it with other techniques. Additionally, there is at least one practical video application that does not require real-time halftoning: the quantization of image sequences for storage on CD-ROM.

---

*This work was supported by Hewlett-Packard Company.

†SPIE/IS&T conference on *Human Vision, Visual Processing, and Digital Display V*, vol. 2179, pp.207-217, San Jose, CA Feb. 7–10, 1994.

The paper is organized as follows. In Section 2, we introduce the model for the spatio-temporal contrast sensitivity of the human visual system. Section 3 is devoted to defining our error metric and reducing it to a form easily implemented on a digital computer. In Section 4, we present our optimization strategy, and demonstrate its computational efficiency. In Section 5, we present and discuss our results.

2 Model of the Human Visual System

In this section, we define our model of the human visual system, which will play an important role in achieving visually pleasing halftone sequences. Let $\mathbf{x} = (x, y, t)$ denote a vector-valued point in the 3-dimensional Cartesian space $\mathbf{X} = \mathbb{R}^3$, where the x- and y-axes are spatial coordinates, and the t-axis is time. Denote by L^2 the class of finite-energy, integrable functions on $\mathbf{X}$. The human visual system is then assumed to be a functional mapping $H : L^2 \to L^2$ upon which we place the restriction that it be linear and invariant to both temporal and spatial shifts, so that it is completely specified by its impulse response $a : \mathbf{X} \to \mathbb{R}$ or its frequency response $A(\mathbf{u}) = \int_{\mathbf{X}} a(\mathbf{x}) \exp(-i\mathbf{u}^T\mathbf{x}) d\mathbf{x}$, where $\mathbf{u} = (u, v, \omega)$ are the spatial and temporal frequencies corresponding to $\mathbf{x}$.

The goal is to define A based on empirical measurements of the contrast sensitivity of the human visual system. The experiment which yields the necessary data is as follows. An achromatic sinusoidal grating with spatial frequency components (u, v) is displayed on a monitor screen; its amplitude is modulated sinusoidally with frequency ω about its nominal luminance L, yielding the stimulus function $L + \Delta L \cos(\omega t) \cos(ux + vy)$. The subject of the experiment is then directed to vary ΔL until the threshold above which the stimulus can be perceived is reached. The process is repeated for various values of (u, v, ω), hence defining a function $\Delta L(u, v, \omega)$. The contrast sensitivity function

$$A(\mathbf{u}) = A(u, v, \omega) = \frac{L}{\Delta L(u, v, \omega)}$$

is a measure of sensitivity to achromatic stimuli with frequency (u, v, ω) and average luminance L.

Kelly [4] has performed the above experiment, and conveniently supplies a mathematical expression for a curve which fits his data over a wide range of frequencies. Several points with respect to Kelly's work are worth addressing. First, it is assumed that the contrast sensitivity function depends only on the magnitude of (u, v), not its angle; hence A is restricted to dependence on (α, ω), where $\alpha = \sqrt{u^2 + v^2}$. Second, Kelly's data correspond to "stabilized" viewing conditions, wherein small, jerky, random eye movements are tracked and compensated for. These movements, called saccades, constitute the mechanism predominantly responsible for our ability to perceive stationary scenes; indeed, the data reflect this fact: the contrast sensitivity function drops off as ω decreases to zero. The applicability of these data for situations wherein viewing conditions are not stabilized may be brought into question. We shall address this issue further in Section 5, in which we present our results.

The functional form of the contrast sensitivity function is

$$A(\alpha, \omega) = \left[6.1 + 7.3 \left|\log\left(\frac{\omega}{3\alpha}\right)\right|^3\right] \omega\alpha \exp\left(-\frac{2(\omega + 2\alpha)}{45.9}\right).$$

We choose this frequency response, which, along with the assumptions of linearity and shift invariance, defines the human visual system H. The energy spectrum $|A(\alpha, \omega)|^2$ is plotted in Figure 1.

In the next section, we shall define an error metric, the minimization of which would ideally produce the most visually appealing halftone sequence for a given continuous-tone sequence.

3 Error Metric

In this section, we describe our model for the overall system which operates on an image sequence during the viewing process. This model encompasses the human visual system described in the previous section, and also the effects of the display upon which the sequence is viewed. Based on this model, we define the global error measure which we wish to minimize. Finally, we demonstrate that a few reasonable assumptions suffice to simplify the error metric from one which depends on continuous-space functional arguments to one which comprises only a finite set of parameters, facilitating the design of optimization algorithms for digital computer implementation.

Let $\mathbf{M} = \mathbf{Z}^3$, where $\mathbf{Z}$ is the set of integers. We denote by $\mathbf{m} = (m, n, k) \in \mathbf{M}$ the discrete vector-valued argument whose components correspond to spatial and temporal indices. Let $g : \mathbf{M} \to [0, 1]$ be the continuous-tone image sequence we wish to quantize. The goal is to generate a halftone sequence $h : \mathbf{M} \to \{0, 1\}$ which is the "best" approximation to g.

To this end, we assume that g is viewed on a display, yielding the output signal $D\{g\}$, which is then perceived by the human visual system to be $H\{D\{g\}\}$. Similarly, h is passed through the cascade of the systems D and H to produce the perceived halftone sequence $H\{D\{h\}\}$.

Let us first examine the display system response D. The display transforms a discrete-parameter sequence into a function of continuous spatial and temporal arguments, denoted by $\mathbf{x} = (x, y, t) \in \mathbf{X}$. We assume that gamma correction or the like is employed, if necessary, so that the display intensity is linearly related to the input data. This linearization is assumed to be incorporated in D. Further, we model D as having a spatio-temporal impulse response $d : \mathbf{X} \to \mathbb{R}$ that is invariant to spatial or temporal shifts, so that D is completely characterized by d. Then we may write

$$D\{h\}(\mathbf{x}) = \sum_{\mathbf{m}\in\mathbf{M}} h(\mathbf{m}) d(\mathbf{x} - \mathbf{T}\mathbf{m}), \tag{1}$$

where the 3×3 matrix $\mathbf{T}$ transforms the abstract, discrete index $\mathbf{m}$ into a physical spatio-temporal offset $\mathbf{x}$.

$$\mathbf{T} = \begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0}^T & T_0 \end{bmatrix}.$$

The 2×2 matrix $\mathbf{P}$ relates the position index (m, n) to a physical screen displacement (x, y). Under the assumption of a rectangular lattice, $\mathbf{P}$ is diagonal. The indices (m, n) are assumed to be ordered so that they need only be multiplied by a spatial sampling period to yield the resulting screen displacement (x, y). Therefore,

$$\mathbf{P} = \begin{bmatrix} X_0 & 0 \\ 0 & Y_0 \end{bmatrix},$$

where X_0 and Y_0 are the vertical and horizontal spacing between display pixels. T_0 is the time interval between samples, and $\mathbf{0} = (0, 0)^T$. Note that the diagonal elements of $\mathbf{T}^{-1}$ are the spatial and temporal sampling rates. An analogous expression to (1) holds for $D\{g\}$.

We now consider the human visual system H. As discussed in the previous section, H is assumed to be a linear, shift-invariant system with spatio-temporal impulse response $a = a(\mathbf{x})$. Letting $h_d = D\{h\}$ and $g_d = D\{g\}$, we have

$$H\{D\{h\}\} = a * h_d,$$

and, similarly,

$$H\{D\{g\}\} = a * g_d.$$

The asterisk denotes convolution in both time and space, while the subscript "d" denotes "display." We define the error function $e_d = h_d - g_d$, and the perceived error function $\epsilon_d = a * e_d$.

The global error metric is chosen to be the energy in the perceived error function:

$$\begin{aligned}
E = ||\epsilon_d||_2^2 &= \int_{\mathbf{X}} \epsilon_d^2(\mathbf{x})d\mathbf{x} \\
&= \int_{\mathbf{X}} \{[a * (h_d - g_d)](\mathbf{x})\}^2 \, d\mathbf{x} \\
&= \int_{\mathbf{X}} \{[a * (D\{h\} - D\{g\})](\mathbf{x})\}^2 \, d\mathbf{x} \\
&= \int_{\mathbf{X}} \{[a * (D\{e\})](\mathbf{x})\}^2 \, d\mathbf{x} \\
&= \int_{\mathbf{X}} \left\{ \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m})[a * d](\mathbf{x} - \mathbf{T}\mathbf{m}) \right\}^2 d\mathbf{x} \\
&= \int_{\mathbf{X}} \left\{ \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m})p(\mathbf{x} - \mathbf{T}\mathbf{m}) \right\}^2 d\mathbf{x},
\end{aligned}$$

where we have introduced the notation $e = h - g$ for the quantization error, and $p = a * d$ for the perceived spatio-temporal display impulse response. Carrying out the squaring operation, we have

$$\begin{aligned}
E &= \sum_{\mathbf{n}\in\mathbf{M}} e(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m}) \int_{\mathbf{X}} p(\mathbf{x} - \mathbf{T}\mathbf{m})p(\mathbf{x} - \mathbf{T}\mathbf{n})d\mathbf{x} \\
&= \sum_{\mathbf{n}\in\mathbf{M}} e(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m})r(\mathbf{n} - \mathbf{m}),
\end{aligned}$$

where r is a sequence of samples of the autocorrelation function of the perceived spatio-temporal impulse response; the samples are taken on the same discrete-valued grid as that which constitutes the support of the continuous-tone and halftone sequences. The definition of r is

$$r(\mathbf{m}) = \int_{\mathbf{X}} p(\mathbf{x})p(\mathbf{x} + \mathbf{T}\mathbf{m})d\mathbf{x}.$$

Observe that r is the integral of the product of two shifted copies of p. The absolute offsets of these copies are irrelevant; only their relative positions affect the value of the integral. This is the reason that r depends only upon the difference between $\mathbf{m}$ and $\mathbf{n}$. For the same reason, r is symmetric: $r(\mathbf{m}) = r(-\mathbf{m})$. If we neglect the interpixel and interframe display effects, it can be shown (the proof is based on Appendix A in [5]) that

$$r(\mathbf{m}) = K \int R(\mathbf{u}) \exp(i\mathbf{u}^T\mathbf{m})d\mathbf{u},$$

where

$$R(\mathbf{u}) = \sum_{\mathbf{n}\in\mathbf{M}} \left| A(\mathbf{T}^{-1}(\mathbf{u} - 2\pi\mathbf{n})) \right|^2 \tag{2}$$

and K is a positive constant. The integral is taken over any cube whose sides have length 2π.

We now have an expression for E which contains no integrals, only sums. We approximate r by a finite vector by appropriately sampling and windowing $R(\mathbf{u})$, and taking its inverse DFT. It is now reasonable to introduce a shorthand matrix notation for E. Define the vector $\mathbf{e}$ whose components are some ordering of the elements in the sequence e, and the matrix $\mathbf{R}$ whose rows are appropriate permutations of the sequence r, such that the matrix multiplication $\mathbf{Re}$ constitutes a vector representation of the convolution $r * e$. $\mathbf{R}$ is an autocorrelation matrix regardless of the ordering chosen; hence, it is symmetric: $\mathbf{R} = \mathbf{R}^T$. We may now write $E = \mathbf{e}^T\mathbf{Re}$, which emphasizes the quadratic nature of the cost function. Since E is a nonnegative quantity, we see immediately that $\mathbf{R}$ is nonnegative definite. Writing $\mathbf{e}$ in terms of the halftone and continuous-tone image sequences, $\mathbf{e} = \mathbf{h} - \mathbf{g}$, yields

$$E = \mathbf{h}^T\mathbf{Rh} - 2\mathbf{h}^T\mathbf{Rg} + \mathbf{g}^T\mathbf{Rg}. \tag{3}$$

Observe that $\mathbf{g}$ is fixed; the goal is to select $\mathbf{h}$ to minimize E. Hence, the last term is irrelevant with respect to the optimization problem. The middle term is a cross-correlation quantity which couples $\mathbf{h}$ to $\mathbf{g}$; it is this term which enforces the fidelity of the halftone sequence to the original. The first term, which is the perceptually weighted energy in the halftone sequence, is a measure of the noise induced by the quantization process; it carries no information about the continuous-tone sequence: indeed, if the second term were absent, the error would be minimized by taking $\mathbf{h}$ identically equal to zero.

In the next section, we shall use the above expression to devise an efficient algorithm for minimizing E, thereby optimizing the halftone image sequence.

4 Minimization Strategy

We now consider a technique for optimizing the halftone sequence with respect to the error metric defined in the previous section. The goal is to choose the vector $\mathbf{h}$ which minimizes

$$E = \mathbf{h}^T\mathbf{Rh} - 2\mathbf{h}^T\mathbf{Rg} + \mathbf{g}^T\mathbf{Rg}.$$

To achieve this minimization, we employ an extension of a method which has proved to be quite effective for conventional 2-D (spatial) halftoning: the so-called "direct binary search" (DBS) algorithm [2, 9]. The idea is straightforward: supply an initial condition for $\mathbf{h}$; thence visit each index $\mathbf{m} \in \mathbf{M}$ in some prescribed order, investigating at each $\mathbf{m}$ the change in E due to toggling $h(\mathbf{m})$ (*i.e.* from 0 to 1 or 1 to 0) or exchanging it with any of its (differing) adjacent neighbors. If any of these changes, either toggling or exchanging, causes the error to decrease, then the change which most reduces the error is made. For example, in the 3-D case, each pixel has 26 adjacent neighbors, counting those adjoining diagonally. Hence, including the option of toggling, there are 27 possible changes to consider at each pixel $\mathbf{m}$. If any of these changes reduces the error, then the one which reduces it the most is performed. After accepting or rejecting a change at a particular pixel, the same thing is done at the next site. This process continues until an entire pass has been made through the index set $\mathbf{M}$ without discovering a single change which results in a decrease in error.

As we only accept changes that reduce the error, and since the error is nonnegative, this stopping condition is guaranteed to occur.

We do not claim that the global minimum will necessarily be achieved; in fact it is quite unlikely given the enormity of the search space. However, it has been our experience in applying this strategy to 2-D halftoning that very good minima are reached; only with considerably more computation, such as stochastic approaches, have we managed to surpass the performance of the direct binary search. A typical halftone produced by DBS is depicted in Figure 2.

We now develop the basis for efficient implementation. Recall that the error metric can be written as

$$E = \sum_{\mathbf{n}\in\mathbf{M}} e(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m}) r(\mathbf{n}-\mathbf{m}).$$

Consider changing $h(\mathbf{n})$ to $h(\mathbf{n})+\Delta(\mathbf{n})$. Since $e = h - g$, this is the same as changing $e(\mathbf{n})$ to $e(\mathbf{n})+\Delta(\mathbf{n})$. Hence, the resulting change in the global error is

$$\begin{aligned}
\Delta E &= \sum_{\mathbf{n}\in\mathbf{M}} [e(\mathbf{n})+\Delta(\mathbf{n})] \sum_{\mathbf{m}\in\mathbf{M}} [e(\mathbf{m})+\Delta(\mathbf{m})] r(\mathbf{n}-\mathbf{m}) - \sum_{\mathbf{n}\in\mathbf{M}} e(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m}) r(\mathbf{n}-\mathbf{m}) \\
&= \sum_{\mathbf{n}\in\mathbf{M}} \Delta(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} \Delta(\mathbf{m}) r(\mathbf{n}-\mathbf{m}) + 2\sum_{\mathbf{n}\in\mathbf{M}} \Delta(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m}) r(\mathbf{n}-\mathbf{m}) \\
&= \sum_{\mathbf{n}\in\mathbf{M}} \Delta(\mathbf{n}) \sum_{\mathbf{m}\in\mathbf{M}} \Delta(\mathbf{m}) r(\mathbf{n}-\mathbf{m}) + 2\sum_{\mathbf{n}\in\mathbf{M}} \Delta(\mathbf{n}) c(\mathbf{n}),
\end{aligned}$$

where $c = r * e$ is the sequence which results from passing the error sequence e through a digital filter whose impulse response is r, the autocorrelation of the perceived display spatio-temporal impulse response. Performing this convolution is computationally intensive, given its 3-dimensional nature and the sheer size of a typical sequence of images. However, the full-blown convolution need only be calculated once (and can be achieved in reasonable time by the fast Fourier transform); it is then stored in a look-up table, and updated as e changes due to toggling or exchanging elements of the halftone sequence. We shall see shortly that these updates require only modest computation.

Let us examine the specific cases corresponding to a toggle or an exchange of halftone pixels. First, we investigate ΔE due to toggling one halftone element, say $h(\mathbf{n}_0)$. In this case, $\Delta(\mathbf{n}) = u\delta(\mathbf{n}-\mathbf{n}_0)$, where δ is the Kronecker delta function, and $u = \pm 1$, the sign depending on the pixel's value prior to toggling. Hence,

$$\begin{aligned}
\Delta E &= \sum_{\mathbf{n}\in\mathbf{M}} u\delta(\mathbf{n}-\mathbf{n}_0) \sum_{\mathbf{m}\in\mathbf{M}} u\delta(\mathbf{m}-\mathbf{n}_0) r(\mathbf{n}-\mathbf{m}) + 2\sum_{\mathbf{n}\in\mathbf{M}} u\delta(\mathbf{n}-\mathbf{n}_0) c(\mathbf{n}) \\
&= \sum_{\mathbf{n}\in\mathbf{M}} \delta(\mathbf{n}-\mathbf{n}_0) r(\mathbf{n}-\mathbf{n}_0) + 2uc(\mathbf{n}_0) \\
&= r(0) + 2uc(\mathbf{n}_0),
\end{aligned}$$

where the fact that $u^2 = 1$ has been used. We now compute ΔE due to exchanging two halftone pixels, say $h(\mathbf{n}_0)$ and $h(\mathbf{m}_0)$, where it is assumed that the two have opposite values (there is nothing to be gained by exchanging identical pixels). Hence, $\Delta(\mathbf{n}) = u\delta(\mathbf{n}-\mathbf{n}_0) - u\delta(\mathbf{n}-\mathbf{m}_0)$, where again $u = \pm 1$. We now have

$$\Delta E = \sum_{\mathbf{n}\in\mathbf{M}} u[\delta(\mathbf{n}-\mathbf{n}_0) - \delta(\mathbf{n}-\mathbf{m}_0)] \sum_{\mathbf{m}\in\mathbf{M}} u[\delta(\mathbf{m}-\mathbf{n}_0) - \delta(\mathbf{m}-\mathbf{m}_0)] r(\mathbf{n}-\mathbf{m})$$

$$
\begin{aligned}
&+2\sum_{\mathbf{n}\in\mathbf{M}} u[\delta(\mathbf{n}-\mathbf{n}_0)-\delta(\mathbf{n}-\mathbf{m}_0)]c(\mathbf{n}) \\
=\;& 2r(0)-2r(\mathbf{n}_0-\mathbf{m}_0)+2u[c(\mathbf{n}-\mathbf{n}_0)-c(\mathbf{n}-\mathbf{m}_0)],
\end{aligned}
$$

where use has been made of the relation $r(-\mathbf{n}) = r(\mathbf{n})$, and again $u^2 = 1$.

Observe that ΔE can be computed with one multiplication and one addition in the case of toggling; and two multiplications and three additions in the case of exchanging. Computation of the minimum ΔE in the worst case requires investigating 26 exchanges and one toggle. However, only when neighboring pixels have opposite values are exchanges considered, so ΔE can usually be calculated with far fewer than the worst-case number of operations.

When the minimum of the ΔE quantities for a given pixel is negative, the operation (exchange or toggle) which most reduces E, *i.e.* yields the most negative ΔE, is performed. This has the effect of adding $\Delta(\mathbf{n})$ to $e(\mathbf{n})$, as described above. This necessitates updating the look-up table in which the values of the c sequence are stored. We now derive the required updates. In the case of a toggle, $\Delta(\mathbf{n}) = u\delta(\mathbf{n}-\mathbf{n}_0)$. Then the change in c is

$$
\begin{aligned}
\Delta c(\mathbf{n}) &= \sum_{\mathbf{m}\in\mathbf{M}} [e(\mathbf{m}) + u\delta(\mathbf{m}-\mathbf{n}_0)]r(\mathbf{n}-\mathbf{m}) - \sum_{\mathbf{m}\in\mathbf{M}} e(\mathbf{m})r(\mathbf{n}-\mathbf{m}) \\
&= ur(\mathbf{n}-\mathbf{n}_0)
\end{aligned}
$$

For an exchange, $\Delta(\mathbf{n}) = u\delta(\mathbf{n}-\mathbf{n}_0) - u\delta(\mathbf{n}-\mathbf{m}_0)$. The corresponding change in c is then

$$
\begin{aligned}
\Delta c(\mathbf{n}) &= u\sum_{\mathbf{m}\in\mathbf{M}} [\delta(\mathbf{m}-\mathbf{n}_0) - \delta(\mathbf{m}-\mathbf{m}_0)]r(\mathbf{n}-\mathbf{m}) \\
&= u[r(\mathbf{n}-\mathbf{n}_0) - r(\mathbf{n}-\mathbf{m}_0)]
\end{aligned}
$$

The support of Δc is, in the case of a toggle, simply the support of the autocorrelation sequence $r(\mathbf{n}-\mathbf{n}_0)$. Therefore, the number of entries in the c look-up table which must be updated is the number of non-zero coefficients in the filter r; call this size $|r|$. Then the number of additions required to update c in the case of a toggle is $|r|$. No multiplications are needed. On the other hand, if an exchange is performed, the support of c is the union of the supports of $r(\mathbf{n}-\mathbf{n}_0)$ and $r(\mathbf{n}-\mathbf{m}_0)$. As an exchange is computationally equivalent to two toggles, the number of additions required in this case is $2|r|$. Observe that, if a toggle or an exchange were performed at every pixel, the computation required would be equivalent to repeating the original convolution. However, only a fraction of the pixels are exchanged or toggled; furthermore, this fraction decreases rapidly toward zero as multiple passes through the image sequence are made. This feature is what makes the algorithm efficient.

5 Results

In this section, we present a comparison of the performances of three approaches to video halftoning, including our spatio-temporal model-based error minimization strategy. We compute the perceived error energy, as defined in (3), for each technique and note the degree to which the error metric correlates with the subjective quality of halftone sequences. We also describe the appearance of the sequences, and offer our interpretations of the results.

For purposes of comparison, we consider applying the 2-dimensional DBS algorithm, with a spatial-only visual model, to each frame of a continuous-tone sequence. The initial states of the frames in the halftone sequence are independent and random. This technique has the advantage of being considerably faster than a full 3-dimensional search, but does not take into account the interframe correlation of the continuous-tone sequence. Reasoning that the frames in the halftone sequence should also be correlated, Gotsman [7] modified this strategy as follows. The frames in the sequence are halftoned sequentially. The initial condition for a given halftone frame is chosen to be the result of halftoning the previous frame. The idea is that this initial condition should be "close" to a local minimum for the next frame, so that the search algorithm will settle in that local minimum, thereby maintaining strong correlation between the frames of the halftone sequence. For brevity, we will refer to the first method as "2DFI" (for "2-D frame-independent"), and the second, modified strategy as "2DFD" ("2-D frame-dependent"). The full 3-dimensional search, with the spatio-temporal visual model, will be labeled simply "3D".

To investigate the performance of these three approaches, we used the luminance component of the popular "Salesman" sequence as our continuous-tone input. This sequence has a frame rate of 30 Hz. We also created a 60 Hz version of the same sequence by simply replicating each frame. We then applied each of the three halftoning algorithms to both the 30 Hz and the 60 Hz sequences. Note that the 2-D methods do not require any modification when the frame rate changes, since the 2-D visual model has no temporal component. The spatio-temporal model, however, does change, since the T_0 element of the matrix $\mathbf{T}$ in (1) is the inverse of the frame rate.

Observe that, if DBS did indeed achieve a global minimum for each frame, then the 2DFI and 2DFD approaches would be identical, and the 2D results would be the same for 30 Hz and 60 Hz (since the 60 Hz sequence is simply a frame-replicated version of the 30 Hz sequence). Since DBS does not reach a global minimum, the results do depend upon the initial condition for each frame.

When the frame rate of a sequence is 30 Hz, the highest temporal frequency which can be achieved is 15 Hz. Referring to Figure 1, we see that the human visual system is still quite sensitive to such frequencies. We therefore deduce that the human visual system will not perform satisfactory temporal averaging at this frame rate; that is, the high-frequency error components will be visible. The 60 Hz frame rate, on the other hand, corresponds to a maximum temporal frequency of 30 Hz. Error components at this frequency should be less noticeable, so that better averaging will occur. Our subjective assessment of the results agrees with this reasoning.

Referring again to Figure 1, we observe another interesting characteristic of the contrast sensitivity data: the sensitivity is substantially reduced at very low temporal frequencies. This corresponds to the stabilized conditions under which the data were taken. Because saccadic eye movements are the chief means by which we perceive stationary scenes, it follows that compensating for and removing their effects will greatly reduce our sensitivity at very low frequencies. As a result of this dip in sensitivity, we expect that a halftone sequence which globally minimizes the error metric would have substantial low-temporal-frequency error content. However, we discovered that our 3D halftoning algorithm does not produce such results. This is largely due to the search strategy employed: for most pixel orderings, it is very unlikely that such low-temporal-frequency structure will be enforced using only toggles and nearest-neighbor exchanges.

On the other hand, the 2DFD approach was constructed explicitly to maintain a high degree of correlation between adjacent frames. The initial condition at each frame is sufficiently close to a local minimum that few pixels need be changed to reach that local minimum. Therefore, the halftone textures from frame to frame do not change substantially. This situation corresponds to substantial low-temporal-frequency error, so that we might expect the 2DFD technique to perform

Technique	*RMS Error Metric*	*Subjective Quality*
2DFD	2.4%	Poor
2DFI	2.9%	Moderate
3D	2.6%	Better

Table 1: Comparison of three approaches to video halftoning, for 60 Hz frame rate

well with respect to a metric defined from stabilized contrast sensitivity data. This is exactly what happens.

Referring to Table 5, we see that the 2DFD technique yielded halftone sequences with the smallest spatio-temporal error, even though it does not explicitly attempt to minimize the metric by which it is being judged. However, its subjective quality is by far the worst of the three techniques; its spatial resolution is effectively lower than that of the 3D approach since the entire burden of tone reconstruction is placed upon spatial averaging. The spatial textures are extremely visible, and moving objects look as though they were being viewed through a "dirty window." Due to the frame replication used in constructing the 60 Hz continuous-tone sequence, every input frame is identical to one adjacent to it. It is easy to see that, by its design, the 2DFD technique will yield halftone sequences which also possess this property, effectively reducing the output frame rate to 30 Hz. This aggravates the problems associated with the 2DFD strategy.

On the other hand, the 2DFI approach does not force any correlation between adjacent halftone frames; the initial condition for each frame is random and independent of the others, so that two identical continuous-tone frames may yield completely different halftones. The 2DFI approach does not suffer from either the low-temporal-frequency artifacts or the reduced effective frame rate that plague 2DFD; as a result, it produces halftone sequences whose subjective quality is much better than those generated by the 2DFD technique, even though the error metric does not reflect this (see Table 5).

Finally, we compare our 3D halftoning algorithm with the 2D approaches. Subjectively, when the frame rate is 30 Hz, the 3D strategy produces halftone sequences which are qualitatively similar to the 2DFI results. At the 60 Hz frame rate, the 3D approach is subjectively better than the 2DFI: the detail rendition is improved, and the temporal noise is not as noticeable. The difference between the 3D and 2DFI techniques is subtle compared with the improvement from 2DFD to 2DFI, but it is nonetheless present. This improvement is reflected in the error metric.

6 Conclusions

As described in the previous section, the error metric does not correlate well with subjective evaluation for sequences with low-temporal-frequency error. We believe that this shortcoming in the error metric is due to our use of contrast sensitivity data taken under stabilized viewing conditions. If stabilization had not occurred, the sensitivity of the visual system to low-temporal-frequency stimuli would be much greater. We used stabilized data because Kelly [4] measured contrast sensitivity for a much denser set of frequencies than we could find for unstabilized data; he also supplied a simple mathematical expression that fits his data. In the future, it would be more appropriate to use unstabilized data.

Despite a metric which is not correct for low temporal frequencies, we observed that the

3D spatio-temporal halftoning algorithm yields good results, putting the quantization error in high spatial and temporal frequencies, so that it is less visible. We infer that our metric is not unreasonable for high-frequency quantization error. We also observed that our 3D algorithm does not place error components at very low temporal frequencies, even though doing so would reduce the error metric. This is due to the localized search strategy we use, which only allows each pixel to be toggled or exchanged with an adjacent pixel. A more global search would be needed to produce low-frequency components.

In the future, the display should be accounted for. Recall that we neglected the spatial (interpixel) and temporal (interframe) effects of the display. It would be a simple matter, given the spatio-temporal impulse response of the display, to include it in the overall algorithm. Indeed, only (2) would change in that case.

Finally, an eventual generalization of this work would allow multi-level quantization of color image sequences. This will involve additional components in the visual model, and a reformulation of the minimization strategy.

In conclusion, we have presented a metric for the difference between two achromatic image sequences, based on psychophysical data corresponding to the human visual system. We extended a successful 2-dimensional minimization strategy to the spatio-temporal domain, and used it to produce halftone sequences that are superior to those which can be obtained using only spatial psychophysical data. The technique is very useful for image sequence quantization applications which do not require real-time processing (*e.g.* CD-ROM). Future research will concentrate on developing a superior metric. Once this has been done, the design of algorithms to yield near-optimal results in real time will be pursued.

References

[1] T. N. Pappas and D. L. Neuhoff, "Least-Squares Model-Based Halftoning," *Proc. 1992 SPIE/IS&T Symposium on Electronic Imaging Science and Technology*, Vol. 1666, San Jose, CA, Feb 9–14, 1992, pp. 165–176.

[2] M. Analoui and J. P. Allebach, "Model-Based Halftoning Using Direct Binary Search," *Proc. 1992 SPIE/IS&T Symposium on Electronic Imaging Science and Technology*, Vol. 1666, San Jose, CA, Feb. 9–14, 1992, pp. 96–108.

[3] J. B. Mulligan and A. J. Ahumada, Jr., "Principled Halftoning Based on Human Vision Models," *Proc. 1992 SPIE/IS&T Symposium on Electronic Imaging Science and Technology*, Vol. 1666, San Jose, CA, Feb. 9–14, 1992, pp. 109–121.

[4] D. H. Kelly, "Motion and Vision. II. Stabilized Spatio-Temporal Threshold Surface," *J. Opt. Soc. Am.*, Vol. 69, No. 10, pp. 1340–1349, Oct. 1979.

[5] B. Kolpatzik and C. A. Bouman, "Optimized Error Diffusion for Image Display," *J. Electronic Imaging*, Vol. 1, No. 3, 1992, pp. 277–292.

[6] T. J. Flohr, B. W. Kolpatzik, R. Balasubramanian, D. A. Carrara, C. A. Bouman, and J. P. Allebach, "Model-Based Color Image Quantization," *Proc. 1993 SPIE/IS&T Symposium on Electronic Imaging and Science and Technology*, San Jose, CA, Jan. 31–Feb. 4, 1993.

[7] C. Gotsman, "Halftoning of Image Sequences," *The Visual Computer*, Vol. 9, 1993, pp. 255-266.

[8] J. B. Mulligan, "Methods for Spatio-Temporal Dithering," *SID '93 Conference Digest*, Seattle, WA, May 17–21, 1993, pp. 155–158.

[9] T. J. Flohr, C. B. Atkins, and J. P. Allebach, "Can DBS Ever Be a Practical Halftoning Technique?" *SID '93 Conference Digest*, Seattle, WA, May 17–21, 1993, pp. 223–226.

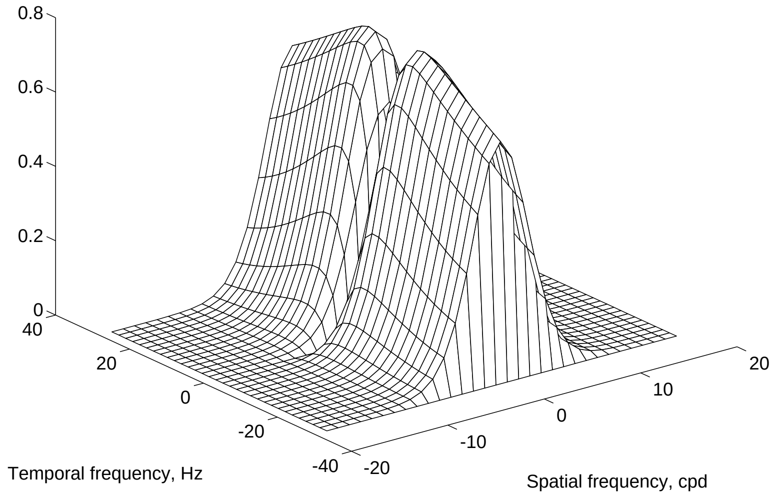

Figure 1: Energy spectrum of human visual contrast sensitivity

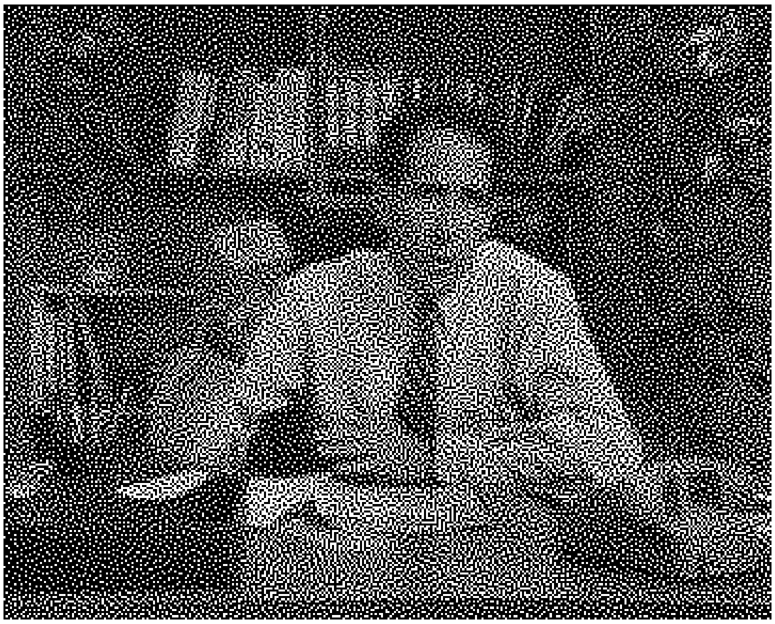

Figure 2: Typical DBS halftone