Principled Halftoning Based on Human Vision Models

Jeffrey B. Mulligan and Albert J. Ahumada, Jr.
NASA Ames Research Center
Moffett Field, CA, 94035-1000

ABSTRACT

When models of human vision adequately measure the relative quality of candidate halftonings of an image, the problem of halftoning the image becomes equivalent to the search problem of finding a halftone that optimizes the quality metric. Because of the vast number of possible halftones, and the complexity of image quality measures, this principled approach has usually been put aside in favor of fast algorithms that seem to perform well. We find that the principled approach can lead to a range of useful halftoning algorithms, as we trade off speed for quality by varying the complexity of the quality measure and the thoroughness of the search.

High quality halftones can be obtained reasonably quickly, for example, by using as a measure the vector length of the error image filtered by a contrast sensitivity function, and, as the search procedure, the sequential adjustment of individual pixels to improve the quality measure. If computational resources permit, simulated annealing can find nearly optimal solutions.

1. INTRODUCTION

The problem of halftoning a gray scale image is the problem of finding a binary image which appears as similar as possible to the gray scale image. This problem can be broken down into two parts: first, we must find a measure which captures the "similarity" of two images; then, after we have done this, we must be able to find the binary image which maximizes this "similarity."

In order to be useful, a measure of similarity between two images must incorporate our knowledge of the human visual system. To illustrate the importance of this, we first consider a simple metric that does *not* do this, namely the familiar root-mean-square (RMS) error. At this point, we introduce some notation: let g_i represent the value of the gray-level image $\mathbf{g}$ at the ith pixel (we use the single index i to represent the two dimensions of spatial position to simplify the following expressions). Similarly, let h_i represent the quantized value of the corresponding pixel in the output halftone image $\mathbf{h}$. We assume that $-1 \le g_i \le 1$, and $h_i = \pm 1$. We define e_i to be the *local error:*

$$e_i = h_i - g_i .$$

The total squared error, ξ, is simply:

$$\begin{aligned} \xi &= \sum_i e_i^2, \\ &= -2 \sum_i h_i g_i + \sum_i \left(1 + g_i^2 \right). \end{aligned} \tag{1}$$

To obtain the RMS error, we divide this quantity by the number of pixels and take the square root. Because this is a monotonic transformation of ξ, any halftone image $\mathbf{h}$ which minimizes the RMS error must also minimize ξ.

be completely accurate for regions where the target image is *not* uniform,[3] it is nevertheless a good starting point. The contrast sensitivity function (CSF) describes the visibility of signals as a function of spatial frequency; for each spatial frequency and orientation, the sensitivity is defined to be the reciprocal of the contrast of the weakest signal that can be seen. Obviously, we are not concerned with errors which cannot be seen; therefore it makes sense to consider, instead of the raw error e_i, the error filtered to weight frequencies according to their detectability. We now consider the space domain representation of the CSF. The CSF can be represented in the space domain by a linear shift invariant filter; the detectability of the quantization error can be estimated from the degree to which it excites this filter. The CSF itself only specifies the amplitude spectrum of the filter; to uniquely determine the filter we assume that it introduces no spatial phase shifts. We will use the symbols f_i to refer to the value of this filtered error at the ith pixel. We also introduce $w_{i,j}$ to represent the filter weight with which the error from the jth pixel, e_j, contributes to f_i:

$$f_i = \sum_j w_{i,j}\ e_j .$$

Our choice of a zero-phase filter together with shift invariance imply symmetry: $w_{i,j} = w_{j,i}$. The shift invariance assumption is that

$$w_{i,j} = w_{k,l} \quad \text{iff} \quad \vec{p}_i - \vec{p}_j = \vec{p}_k - \vec{p}_l ,$$

where $\vec{p}_i$ is a vector representing the position of the ith pixel. Shift invariance is not required for the following discussion, however.

At this point, we must make an assumption about how errors at different parts of the image combine to determine the overall quality of the halftoned representation. For mathematical simplicity, we assume that the combined detectability of the filtered errors is described by the squared Euclidean norm of the vector whose components are the values of the filtered error: We seek the halftoned image which minimizes the total squared filtered error. At this point we will redefine the symbol ξ (which we introduced previously to represent the squared norm of the vector of raw errors) to be the squared norm of the vector of *filtered* errors:

$$\xi = \sum_i f_i^2$$

In this paper, we consider methods that can be specified by three rules: 1) a rule for selecting or generating an initial image; 2) a rule for the sequential selection of image pixels to be revised; 3) a rule for the adjustment of the halftone values at the visited locations.

At this point we introduce a bit of additional notation. Let $\mathbf{h}_n$ be the candidate halftone image after n pixels have been examined. Let k be the index of the nth pixel visited. We define $\mathbf{h}_n^+$ to be $\mathbf{h}_n$ with h_k set to +1 and $\mathbf{h}_n^-$ to be $\mathbf{h}_n$ with h_k set to -1. We use our chosen rule to obtain $\mathbf{h}_0$; we use our scanning rule to obtain a pixel index k from the iteration count n. In the following sections we consider various rules for obtaining $\mathbf{h}_{n+1}$ given $\mathbf{h}_n$.

2.1. Strict Descent

A simple method for reducing ξ is to start from some initial quantized image, and sequentially consider individual pixels in the quantized image, changing them if the change reduces the total error ξ. We refer to this method as strict descent because the total error decreases monotonically.

Consider the effect of the state of h_k, the quantized value at the kth pixel, on the total error ξ. We begin by rewriting the expression for the filtered error to make explicit the dependence on h_k:

$$f_i = \left(\sum_{j \neq k} w_{i,j}\ e_j \right) + w_{i,k}\,(h_k - g_k)\,.$$

Now,

$$\begin{aligned} \xi &= \sum_i f_i^2\,, \\ &= \sum_i \left[\left(\sum_{j \neq k} w_{i,j}\ e_j \right) + w_{i,k}\,(h_k - g_k) \right]^2 . \end{aligned}$$

By expanding the square twice, and collecting all terms which do not depend on the value of h_k into a constant C_i, we obtain

$$\xi = \sum_i \left[2\, w_{i,k}\, h_k \left(\sum_{j \neq k} w_{i,j}\ e_j \right) - 2\, w_{i,k}^2\, h_k\, g_k + C_i \right], \tag{2}$$

where

$$C_i = \left(\sum_{j \neq k} w_{i,j}\ e_j \right)^2 - 2\, w_{i,k}\, g_k \left(\sum_{j \neq k} w_{i,j}\ e_j \right) + w_{i,k}^2\,(1 + g_k^2)\,.$$

To find the value for h_k which minimizes ξ, we consider the values of ξ for the images $\mathbf{h}_n^+$ and $\mathbf{h}_n^-$, corresponding to the two possible states of h_k. Let ξ^+ be the total error associated with the halftone $\mathbf{h}_n^+$, and let ξ^- be the total error associated with the halftone $\mathbf{h}_n^-$. Our rule says that we will pick the image with the lower associated total error:

$$\mathbf{h}_{n+1} = \mathbf{h}_n^+ \quad \text{iff} \quad \xi^+ - \xi^- < 0$$

The remainder of this section is devoted to the evaluation of this difference. Expressions for ξ^+ and ξ^- are obtained from equation (2) by substituting the appropriate value of h_k. We then take the difference:

$$\begin{aligned} \xi^+ - \xi^- &= 4 \sum_i \left[w_{i,k} \left(\sum_{j \neq k} w_{i,j}\ e_j \right) - w_{i,k}^2\, g_k \right], \\ &= 4 \sum_i w_{i,k} \left(\sum_{j \neq k} w_{i,j}\ e_j \right) \;-\; 4\, g_k \sum_i w_{i,k}^2, \\ &= 4 \sum_i w_{i,k} \left(\sum_j w_{i,j}\ e_j - w_{i,k}\, e_k \right) \;-\; 4\, g_k \sum_i w_{i,k}^2. \end{aligned} \tag{3}$$

The temperature parameter T_n is positive, so the state having the smaller value of ξ has the higher probability. As the temperature becomes large, the two states are chosen with nearly equal frequency. If we define $\Delta\xi$ to be the difference between the the two energies,

$$\Delta\xi = \xi^+ - \xi^-,$$

the rule can be stated in terms of the probability that the succeeding state will be $\mathbf{h}_n^+$:

$$\Pr(\mathbf{h}_{n+1} = \mathbf{h}_n^+) = \frac{\exp(-\Delta\xi/T_n)}{1+\exp(-\Delta\xi/T_n)}$$

As the temperature approaches zero, so does the probability that the poor state is chosen, and the rule approaches the deterministic rule of section 2.1:

$$\mathbf{h}_{n+1} = \mathbf{h}_n^+ \quad \text{iff} \quad \Delta\xi < 0.$$

This deterministic rule always improves the energy at every step but almost always stops at a local rather than a global optimum. Annealing schedules which slowly reduce T_n as a function of n can have a good chance of finding the optimal halftone, but they can be very slow.

Good annealing schedules can be estimated from an examination of what happens to the average error when the algorithm is run at a fixed temperature, as shown in figure 3. The figure shows the average error as a function of number iterations for 4 different temperatures. For a given temperature, the average error decreases until it reaches an asymptotic value which corresponds to the "thermal" noise in the system. The higher temperature processes reach their asymptotic values more quickly, but these values are higher than those ultimately reached by the lower temperature processes. Note that the curve in figure 3 corresponding to a temperature of 0.025 has not reached equilibrium after 100000 iterations over the entire image. The data in figure 3 were generated by a process halftoning a uniform input of zeroes, for which the optimal halftone is known to be a perfectly regular checkerboard. The log of the error for the checkerboard has a value of approximately -4, which is quite a bit lower than the best curve in figure 3.

3. RESULTS

We have defined a metric $\xi(\mathbf{g}, \mathbf{h})$, which describes how close a halftone $\mathbf{h}$ comes to perceptually simulating a continuous tone image $\mathbf{g}$. The new algorithms we have presented seek to minimize this error metric. In this framework, one can evaluate the performance of various algorithms by comparing the generated error values.

In order to apply our algorithms, the filter to be applied to the error must first be determined. In our simulations, we have approximated the contrast sensitivity function with a Gaussian. The optimal standard deviation will depend on the final viewing distance, but, as we have seen, the textures which are produced in uniform regions do not depend strongly on the filter size for standard deviations above one pixel.

Once the error filter has been determined, two other free parameters must be determined. First, an initial state of the image must be chosen. We have obtained good results by starting with a random array of light and dark pixels; other choices we have investigated include locally quantizing the image with a fixed threshold (as in figure 1), or using an image obtained from another halftoning algorithm, or using a constant image. Secondly, the order in which the points are visited must be specified. We have tried three different scanning patterns,

one in which the points are sampled randomly, and two deterministic scanning patterns. One of these deterministic scans was a traditional left-to-right top-to-bottom raster; the other was a "scattered" scan in which the two low order bits of the iteration counter determined the quadrant of the point, the next two bits selected the quadrant-within-the-quadrant, and so on recursively.

Our investigations for the case of strict descent have produced the following observations: first, the use of a raster scan produces oscillation in the output with the consequence of more iterations being required to reach equilibrium. The scattered scan pattern produces rapid convergence, but if the initial image is not random the scattered scan method produces regular textures similar to those produced by ordered dither. These are eliminated if the initial halftone is random. Random scanning seems to produce the best results of all, but more iterations are require to insure that all pixels have been visited. Images produced using all of these methods are shown for comparison in figure 4.

4. Existing methods

Among existing halftoning methods, there can be observed a tradeoff between speed and simplicity and the quality of the final image. *Ordered dither*[1] is one of the simplest methods. In this method, each pixel in the input image is compared against a position-dependent threshold, with the result of the comparison determining the state of the corresponding output pixel. The quality of the final image is greatly affected by the choice of the individual thresholds; Bayer[5] has demonstrated a construction for obtaining well-balanced micro patterns. The biggest advantage of this method is its speed and simplicity. Since the operations carried out at a given pixel are independent of the values of all other pixels, the algorithm is eminently suited to implementation on a parallel machine.

At the other end of the spectrum is *error diffusion*,[6] introduced by Floyd and Steinberg. In error diffusion, as each pixel is quantized the quantization error is "diffused" or spread to neighboring pixels which have yet to be quantized. When these pixels are eventually quantized, their final output values are chosen so as to compensate for previous quantization errors, as well as the desired value at the point itself. This method produces high-quality halftones and is particularly good at reproducing sharp edges. Because of its inherently serial nature, however, a parallel implementation is impossible.

Knuth[7] has introduced a hybrid method he calls *dot diffusion.* In this method, small groups of pixels are quantized independently (as in ordered dither), but within each group quantization errors are shared between neighbors. This method obtains results nearly as good as those produced by error diffusion using an algorithm for which a parallel implementation exists.

Of the methods described above, it is generally agreed that error diffusion produces the best results. We might ask, however, is this in fact the best we can do? The serial nature of the algorithm means that the errors get propagated in the direction the image is scanned, which can result in subtle artifacts. From an aesthetic standpoint, we would prefer an algorithm in which the errors are diffused isotropically in all directions. It seems intuitively clear that the pattern of errors will depend somehow on the weights with which the errors are spread to nearby pixels, but there are no recipes telling how to choose the weights to obtain a particular error spectrum.

These problems have been addressed by recent work[4,8,9,10] applying the theory of neural networks to the problem of halftoning. The development presented above is similar to that followed by these earlier efforts. Our approach differs, however, in that we have eliminated the intermediate image in the Hopfield network which is subjected to a nonlinearity in order to produce the final halftone image. In our method the pixels are processed sequentially, whereas when the intermediate image is used it is updated in parallel, or "lock-step" fashion. We believe that this difference produces faster convergence for our method in the strict descent case.

In figure 5, we present a comparison of the present method with the most popular competitor, error diffusion. The particular implementation of the error diffusion used the weights given in Newman and Sproull's text,[11] and used a serpentine raster which processed alternate scan lines in alternate directions.

5. DISCUSSION

Unlike error diffusion, the present method produces halftones which possess uniform textural properties at all gray levels. This is not the case for the error diffusion algorithm, as can be seen in figure 5, where the error diffusion image shows the formation of regular patterns at gray levels producing dot densities of 25%, 50% and 75%. Regular patterns such as these can interact with nonlinearities in the output device (such as ink spread in a printer dot) to produce artifactual distortions in the tone scale. While output nonlinearities will also distort images processed using the algorithm described in this paper, the uniformity of the halftone texture with changes in gray level will prevent the introduction of false contours by the artifacts, making the method robust with respect to failings of the output device.

In our development, we used the contrast sensitivity function to filter the error in order to assess the quality of a candidate halftone. A more refined version might use a more accurate representation of the CSF, but when the image is to be viewed from a variety of distances this may produce little improvement. One aspect of visual sensitivity which probably can be exploited is the oblique effect, which describes the fact that humans have greater sensitivity for signals oriented either vertically or horizontally than for oblique signals.[12]

All approaches based on the CSF presuppose that the quantization errors will be near threshold. For the case where the halftoning noise will be visible (as is commonly the case when presenting halftone images on cathode-ray tube displays), minimum detectability of the error may not be the most appropriate criterion. Since the ultimate goal is communication of the information in the original image, what is really desired is a halftoning noise which can be perceptually disassociated from the underlying image; we would like to produce a halftone in which the noise is seen as a transparent veil overlying an uncorrupted percept of the original image. While these ideas have not been fully developed, we believe that the present method, by making explicit the relationship between the design of the error filter and the resulting error spectrum, will make it easy to design adaptive schemes which tailor the quantization noise to the picture content.

Other investigators have recently considered similar approaches, but instead of flipping the states of individual pixels they considered instead the effects of exchanging the values of adjacent pixels having different states, i.e. "moving" the white pixels around. We note that every such exchange can in principle be generated by the sequential setting and clearing of the individual pixels, although this scenario often involves an improbable intermediate state having a high error value. Nonetheless, as long as a method has a nonzero probability of

reaching all possible states, it will approach the global optimum given enough time to do so. Conversely, if only a fraction of the states can be reached, as in the case when only exchanges are allowed, then it is extremely unlikely that the global optimum will be in the space of images which will be considered. One group following this approach[13] began by setting the total number of white pixels in accordance with the D.C. value of the image; it can be seen from the tone scale compression in figure 2, however, that the optimal image will not have this property if it has unequal areas of values near white and black. Fixing the number of white pixels in the image is equivalent to giving an infinite weight to the value of the CSF at zero frequency. We note that by using a Gaussian error filter in our work, we have already given a disproportionate weight to the low frequency terms, since the CSF is actually bandpass in nature.

Lastly, we must mention one final point which is something of an embarrassment. In the early stages of this work, we were primarily concerned with improving on the results produced by error diffusion, and paid little attention to ordered dither. When we later included ordered dither in our comparisons, we were surprised to find that it produced lower numerical values of the filtered error than either error diffusion or the method presented here, in spite of the fact that the images produced by ordered dither were the least visually pleasing. We interpret this fact to reflect the inadequacy of a simple circularly symmetric filter as a model of the human visual system; in future work, we plan to develop a refined version of the algorithm which uses an array of oriented filters, which we think will be sensitive to the structured artifacts in the ordered dither images, and will give the images rankings which are more in line with perceptual judgements.

6. CONCLUSIONS

In this work we have tried to demonstrate how one could design a halftoning algorithm based on a computational model of the human visual system. The particular model which we have considered here is only a first stab at the problem, but it serves to illustrate some of the principles which we hope to use to extend to more complex models. In particular, we hope to extend the method to exploit the visual system's relatively poor resolution (both in space and time) to chromatic variation.

7. ACKNOWLEDGEMENTS

This work was supported by NASA RTOP 506-7151.

References

1. Ulichney, Robert, *Digital Halftoning*, MIT Press, Cambridge, Mass., 1987.
2. Cambell, F. W. and Robson, J. G., ''Application of Fourier analysis to the visibility of gratings,'' *J. Physiol. (Lond.)*, vol. 197, pp. 551-566, 1968.
3. Ahumada, A. J. Jr., ''Putting the noise of the visual system back in the picture,'' *J. Opt. Soc. Am. A*, vol. 4, pp. 2372-2378, 1987.
4. Carnevali, P., Coletti, L., and Patarnello, S., ''Image Processing by Simulated Annealing,'' *IBM J. Res. Develop.*, vol. 29, pp. 569-579, 1985.
5. Bayer, B. E., ''An optimum method for two-level rendition of continuous-tone pictures,'' *Conference Record of the Intl. Conf. on Communications*, pp. 26-11 - 26-15, 1973.

6. Floyd, R. and Steinberg, L., ‘‘An adaptive algorithm for spatial gray scale,’’ *SID 1975 Symp. Dig. Tech. Papers*, p. 36, 1975.
7. Knuth, Donald E., ‘‘Digital halftones by dot diffusion,’’ *ACM Trans. Graphics*, vol. 6, pp. 245-273, 1987.
8. Anastassiou, D., ‘‘Neural net based digital halftoning of images,’’ *Proc. IEEE Symp. Circuits & Systems*, pp. 507-510, 1988.
9. Ling, Daniel T. and Just, Dieter, ‘‘Neural Networks for Halftoning of Color Images,’’ *IBM Research Report*, vol. RC 16588, 1991.
10. Sullivan, J., Ray, L., and Miller, R., ‘‘Design of minimum visual modulation halftone patterns,’’ *IEEE Trans. Systems Man Cybernetics*, vol. 21, pp. 33-38, 1991.
11. Newman, William M. and Sproull, Robert F., *Principles of interactive computer graphics*, McGraw-Hill, New York, NY, 1979.
12. Olzak, L. A. and Thomas, J. P., ‘‘Seeing spatial patterns,’’ in *Handbook of perception and human performance*, ed. K. R. Boff, L. Kaufman, J. P. Thomas, pp. 7.1-7.56, 1986.
13. Analoui, M. and Allebach, J. P., ‘‘Model-based halftoning by direct binary search,’’ *Proc. SPIE Conference on Human Vision, Visual Processing and Digital Display III*, pp. paper 1666-09, 1992.