THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 6, July–August, 1979
Printed in U.S.A.

An Adaptive Intraframe DPCM Codec Based Upon Nonstationary Image Model

By N. F. MAXEMCHUK and J. A. STULLER

(Manuscript received November 21, 1978)

This paper introduces a nonstationary model for images and develops an adaptive intrafield DPCM *codec based upon the model. The codec attempts to minimize the mean-square coding error at each sample point in the picture. The quantizer in the resulting adaptive codec is found to be similar to that previously obtained from visual masking considerations. Comparative simulation results using 256 × 256 pixel rasters are given for two- and three-bit/pixel versions of the adaptive codec, the three-bit/pixel Graham codec, and three-bit/pixel previous element* DPCM.

I. INTRODUCTION

This paper introduces a nonstationary model for images and develops an adaptive intrafield codec based upon the model. The codec adaptively estimates both the mean and probable range of values of the next picture sample to be encoded and adapts the predictor and quantizer accordingly. In so doing, the coder attempts to minimize the mean-square coding error (MMSE) at each sample point in the picture. The MMSE distortion measure is generally acknowledged to be a poor indicator of image quality.[1] However, when it is applied on a *point* (rather than area) basis in conjunction with the image model presented here, the coder adaptation and resulting coding quality are found to be comparable to that previously obtained from visual masking considerations. This result follows from a property of human vision, stressed by Graham,[2] concerning the strong connection that exists between image chaos (unpredictability) and the visual system's tolerance to noise-like coding distortion. Because of this property, we obtain good image quality at two bits per pel and excellent quality at three bits per pel in a DPCM codec designed solely using the MMSE criterion—masking phenomena are in large part accounted for automatically when the

source model more adequately represents actual images and when the distortion criterion is applied on a point basis.

II. SOURCE MODEL

This section introduces a nonstationary causal source model for the intrafield video process that will be used to develop the adaptive predictive intrafield codec of Section IV. Motivation for the model is intuitive and follows from an examination of a representative video signal of the type we wish to encode, such as that shown in Fig. 1. This is a frame of two interlaced fields, each having 256 pixels per line and 128 lines, with amplitudes stored as 8-bit quantities. The essential characteristic of this (and any) image is that it is a projective transformation of a collection of physical objects. As a consequence, the image is partitioned into regions of luminance elements whose amplitudes are interrelated by the physical structure of the objects they represent. The result is an array of pixels composed of distinct regions having slowly varying "brightness" and "texture" with abrupt boundaries (the picture outline) separating one region from another. We find it natural to view this array as a field that is partitioned into regions of independent quasi-stationary subfields. Two underlying random phenomena are involved: the random amplitudes of picture elements within a

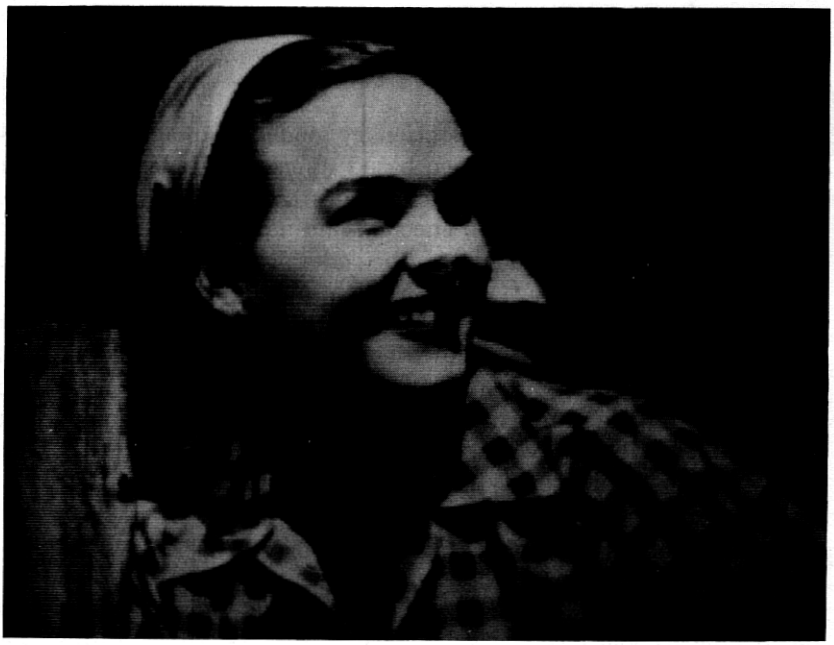

Fig. 1—Checker girl original.

given subfield and the random selection of the subfield with respect to raster coordinates. A source model that incorporates both phenomena is shown in Fig. 2.

Figure 2 models the image generation process as a composite of Q autoregressive sources, $q = 1, 2, \cdots, Q$, and one white source, $q = 0$. Switches S1 and S2 determine which source generates output luminance $s_{mn} \equiv s_t$, where m and n are, respectively, the line number in the field and the column number of the pixel and t is the time the pixel is encountered during conventional line scanning. The autoregressive sources, characterized by predictors 1 through Q and "innovations" process $w_{mn} \equiv w_t$, provide a set of Q possible processes from which the regions of slowly varying brightness and texture of a subfield in an actual image can be approximated. The random variables w_t are assumed zero mean, independent, and characterized by a single known probability density function (pdf) $g(w)$. The predictors F_q in sources $q = 1, 2, \cdots, Q$ are taken to be linear functions of pixels from the local past neighborhood of coordinate $(m, n) \equiv t$. Section IV discusses the specific predictors chosen for the codec of this paper (Table I). Source 0 models those pixels of an actual image that either have no structural relation to previous pixels or whose relation to these pixels is not adequately modeled by sources 1 through Q. Such pixels tend to occur in highly chaotic regions of the image and at certain boundaries at which new subfields are initiated. Since this source represents the extreme of chaos possible in an image, its output is taken to be a

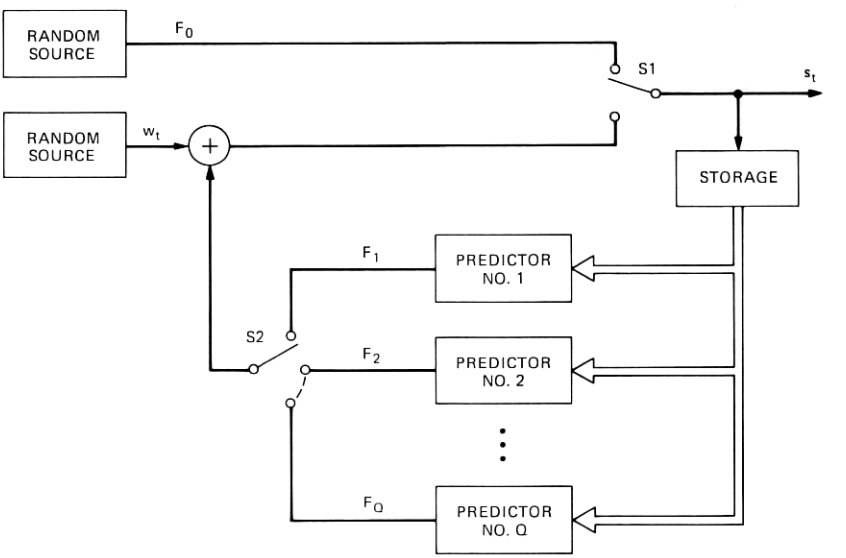

Fig. 2—Source model. Switches S1 and S2 are governed by eq. (2) in the text.

sequence of random variables $F_0 \equiv F_0(m, n)$, each uniformly distributed over [0, 255].

Switches S1 and S2 of the model determine which source is used for final output at each raster coordinate and thus determine image outline as well as more subtle structural changes. "Outline" and "subtle structural changes" are subjectively perceived qualities of an image that are difficult to quantify probabilistically. However, in an actual image "structure" tends to vary slowly: boundaries between regions are an exception, but even here the discontinuity is generally only along one dimension. This suggests that probabilistic information regarding the source in operation at time t can be inferred by appropriate processing of the pixels in the local past vicinity (in the same field) of the pixel in question. To arrive at a source model that characterizes this quasistationary in the simplest way, we model the image source as choosing sources, $q = 0, 1, 2, \cdots, Q$, *independently* according to *unknown* first-order probabilities $P[q; (m, n)]$ that are slowly varying functions of coordinates (m, n). We further assume that the Q textures generated by sources $q = 1, 2, \cdots, Q$ are *a priori* equally likely for a *random* choice of coordinate (m, n):

$$E\{P[q; (m, n)]\} = c \quad q = 1, 2, \cdots, Q \tag{1a}$$

and

$$E\{P[q; (m, n)]\} = \epsilon \ll c \quad q = 0, \tag{1b}$$

where c and ϵ are constants satisfying $\epsilon + Qc = 1$ and the expectation is taken over the raster coordinates.

An alternative approach would be to model the sequence of q_{mn} as stationary Markov. However, this approach was not taken since the assumption of nonstationary *independent* q leads to a relatively simple codec that is robust with respect to both varied picture inputs and channel errors. The assumption of equality of expectations in (1b) leads to mini-max performance with respect to variations in textural content of the picture to be encoded. Biasing this *a priori* distribution toward one predictor would make the codec more susceptible to poor performance on a picture which does not match this distribution. The problem faced by the codec is to estimate probabilities $P[q; (m, n)]$ by suitable processing of past pixel outputs and use the estimates to best advantage for bandwidth compression.

To summarize, the proposed model of the video process has the form (Fig. 2)

$$s_t = \begin{cases} F_0, \text{ with probability } P(0; t) \\ F_q + w_t, \text{ with probability } P(q; t), \quad 1 \le q \le Q, \end{cases} \tag{2}$$

where F_0 is an independent random variable uniform over [0, 255], F_q is a given linear function of pixels in the local past neighborhood of s_t

(Table I), and w_t is an independent zero mean random variable characterized by probability density function (pdf) $g(w)$. Probabilities $p(q;\ t)\ q = 0,\ 1,\ \cdots,\ Q$ vary slowly with respect to at least one coordinate of the raster and satisfy (1); otherwise, these probabilities are unknown. Note that the model embeds the elusive variety of gross image structure in the unknown probabilities $P(q;\ t)$, $0 \leq q \leq Q$. These represent the probabilistic information that the encoder hopes to learn by suitable processing of past image source outputs.

Figure 3 illustrates a representative output generated by the model. In obtaining this output, $g(w)$ was assumed Laplacian, and the $P(q;\ t)$ were estimated from the image of Fig. 1 by a procedure described in Section III. Significant increases in structural similarity to Fig. 1 are possible by modeling the sequence of w_t as nonstationary. In the interest of codec simplicity, however, this additional complexity is not included in the model.

III. ANALYTICAL DEVELOPMENTS

The design of the DPCM codec of Section IV requires specification of both the quantizer and the predictor. Complete statistical information pertinent to this design is contained in the conditional pdf of s_t given

Fig. 3—Representative output of source model. Output of the source model of Fig. 2, where w_t is Laplacian and probabilities $P(q;\ t)$ of eq. (2) are estimated from Fig. 1.

the set of past pixels $\{s_t, t < t\} \equiv S_t^-$. This section describes how this conditional pdf can be etimated at the source output.

It can be easily shown that, for *given* $P(q; t)$, $q = 0, 1, \cdots, Q$, the probability density of s_t conditioned on S_t^- is (for $0 \le s_t \le 255$)

$$p(s_t \mid S_t^-) = \frac{P(0; t)}{255} + \sum_{q=1}^{Q} g(s_t - F_q(S_t^-))P(q; t), \tag{3}$$

where $F_q(S_t^-)$ denotes the qth predictor F_q of s_t as an explicit function of past pixels S_t^-. An estimate of density function (3) is obtained by replacing $P(q; t)$ in the above by its estimate, as described below.

Let the number of times the qth source had been output in a local past region R_t of N points neighboring $(m, n) = t$ (Fig. 4) be denoted by $n(q)$. Due to the nature of the source model, $n(q)$ cannot be measured at the source output. However, a reasonable and computable approximation to it is given by the expectation $E\{n(q)|S_t^-\}$, where the expectation assumes a random selection of (m, n) and is over the density (w_t). By the quasi-stationarity of $P(q; t)$, we then set

$$\hat{P}(q; t) = \frac{E\{n(q)|S_t^-\}}{N}, \tag{4}$$

which becomes (appendix)

$$\hat{P}(q; t) = \frac{1}{N} \sum_{j=1}^{N} \hat{P}(q_j = q \mid S_t^-), \tag{5}$$

where $\bar{P}(q_j = q \mid S_t^-)$ is the conditional probability that the jth pixel in R_t (Fig. 4) was output by source q based upon *a priori* probabilities $E\{P[q; (m, n)]\}$ of (1). Further manipulations (appendix) give

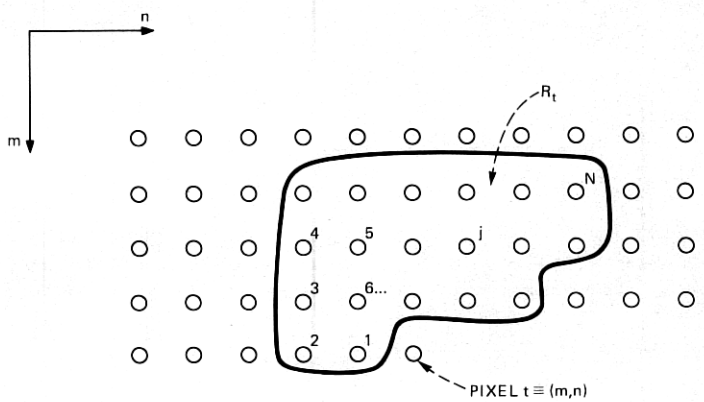

Fig. 4—Illustration of region R_t. This region consists of N pixels in a local past vicinity of pixel (m, n). Note that the numbering of coordinates $j = 1, 2, \cdots, N$ is arbitrary, as are the region boundaries.

$$\hat{P}(q;\, t) = K \frac{\epsilon}{255} \quad q = 0 \tag{6a}$$

$$\hat{P}(q;\, t) = \frac{KC}{N} \sum_{j=1}^{N} g(e_q(t;\, j)) \quad 1 \le q \le Q, \tag{6b}$$

where K satisfies

$$\sum_{q=0}^{Q} \hat{P}(q;\, t) = 1. \tag{7}$$

Equation (5) interprets $\hat{P}(q;\, t)$ as an arithmetic average of *a posteriori* probabilities of q over region P_t, and eq. (6) show how this average can be computed. The term $e_q(t;\, j)$ in (6b) is the difference between the actual value of the jth pixel in region R_t and the predicted value of this pixel given by predictor F_q and is therefore the implied value of the jth innovations variable in R_t under the hypothesis that predictor q was in operation at the source. Explicitly,

$$e_q(t;\, j) = s_t^j - F_q(S_t^{j-}), \tag{8}$$

where s_t^j is the jth pixel in R_t and S_t^{j-} is the set of pixels previous to s_t^j. Equation (6b) estimates $\hat{P}(q;\, t)$, by summing the relative probabilities of the innovations implied under the hypothesis that source q was in operation over region R_t. Note that if $g(\cdot)$ has its peak at zero, then $\hat{P}(q;\, t)$, $1 \le q \le Q$, will be large for those q corresponding to small prediction error $e_q(t,\, j)$ over the N point region. If none of the Q predictors is consistent with past local data, then all terms in the sum of (5b) will be small for $1 \le q \le Q$, and the normalization in (6) will make $\hat{P}(0;\, t)$ large. Further description of (4) to (7) is included in the derivation in the appendix.

The codec described in Section IV predicts s_t by the estimated mean of predictable source outputs:

$$\hat{s}_t = \frac{\sum_{q=1}^{Q} F_q(S_t^-)\hat{P}(q;\, t)}{\sum_{q=1}^{Q} \hat{P}(q;\, t)}. \tag{9}$$

An important characteristic of this prediction rule is its insensitivity to small variations in data S_t^- regardless of the relative values of N and Q. This is in contrast to the covariance method in linear prediction described in a review paper by Makhoul[3] in which small sample size can lead to an ill-conditioned system of equations whose inversion is the adapted predictor. Since (9) is a weighted average of stable (and generally good) estimates F_q, stability persists even for $N < Q$, and some thought indicates that the resulting prediction of $\hat{s}_t$ works in an intuitively reasonable way even if N is only unity.

IV. THE CODEC

In this section, a codec resulting from the source model is described. A block diagram of the encoder is shown in Fig. 5. The codec has been used to code pictures using two and three bits per pel.

The encoder operates by forming Q estimates $F_q(X_t^-)$, $1 \leq q \leq Q$, of source output S_t based upon the previously reconstructed field elements X_t^-. Estimates of source probabilities $P(q; t)$, $0 \leq q \leq 1$, are made with eq. (6) to (7) using previously reconstructed pixels X_t^j in place of S_t^j. Estimates $F_q(X_t^-)$ and probabilities $P(q; t)$ are used to predict the next encoder input pixel s_t, according to (9) and the most likely distribution of values s_t according to (3), with X_t^- replacing S_t^-.

The encoder has been implemented using an $N = 4$ point learning region R_t (Fig. 6) and $Q = 6$ predictors. The predictors used are given in Table I.

Note that with these six predictors the form of predictor (7) can be any one of the most common fixed predictors used in intrafield coders. This varies over the picture so that the best predictor (or best weighted sum) considering the recent past will be used at each sample point.

The pictures which were encoded consisted of 256 lines in two interleaved fields and 256 samples per line. The previous line elements were taken from the previous line in the same field. In this environment, no advantage was obtained by including elements more than one line away in the estimates. Similarly, no visible improvement was obtained using elements that were more than two elements away on the same line. The slope estimator, F_5, and the planer estimator, F_6, were found to be particularly useful in the system which uses two bits

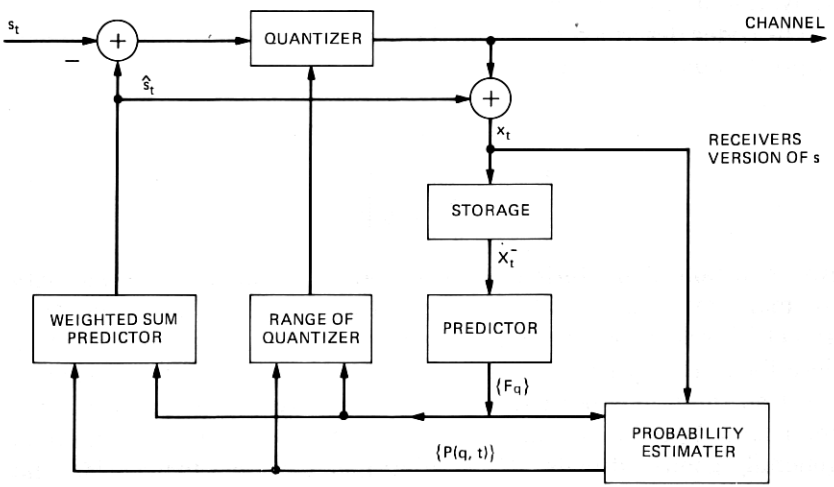

Fig. 5—The encoder.

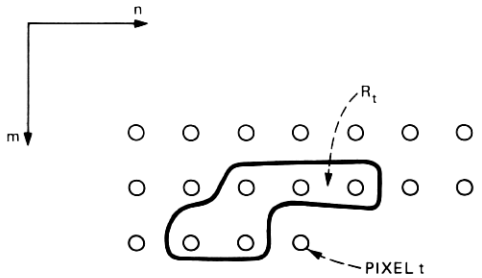

Fig. 6—Four-point region R_t used by codec.

Table I

$F_1(n, m) = \hat{x}(n-1, m)$.
$F_2(n, m) = \hat{x}(n-1, m-1)$.
$F_3(n, m) = \hat{x}(n, m-1)$.
$F_4(n, m) = \hat{x}(n+1, m-1)$.
$F_5(n, m) = 2\hat{x}(n-1, m) - \hat{x}(n-2, m)$.
$F_6(n, m) = \hat{x}(n-1, m) + \hat{x}(n, m-1) - \hat{x}(n-1, m-1)$.

per pel. These estimators allowed the coder to respond more quickly to edges within the picture, and reduced slope overload.

Ideally, the quantizer should be adapted at each point to the estimated probability distribution of s_t. In view of the complex form of (3), this type of redesign is not feasible, and the following ad-hoc curve-fitting technique was used to simplify the adaptation algorithm. The density function $g(w)$ was taken as Laplacian, $g(w) = \alpha/2 \exp(-\alpha \mid w \mid)$. The Max quantizer[4] for this was determined. Each side of the distribution (3) about the mean $\hat{s}_t$ was then approximated by an exponential distribution, and the axis was simply scaled appropriately in codec operation to place the quantization levels. The parameter of each exponential distribution was selected so that it had the same first moment about $\hat{s}_t$ as the corresponding portion of the actual distribution as described in Fig. 7.

When the estimated probability of occurrence of the random estimator is near zero and the estimators $q = 1$ through 6 are identical corresponding to a perfectly flat region in the picture, the parameter of the exponential defining the quantizer assumes its smallest value. In this situation, the parameter of the exponential defining the quantizer is approximately α, the parameter of the Laplacian distribution defining the innovation term in the model. Therefore, α determines the minimum values of the levels of the quantizer, and these, in turn, determine the amount of granularity due to quantization noise in flat regions of the picture and the ability of the coder to respond to unexpected edges. The smaller the value of α, the lower the granular quantization noise; the larger the value of α, the quicker the coder can

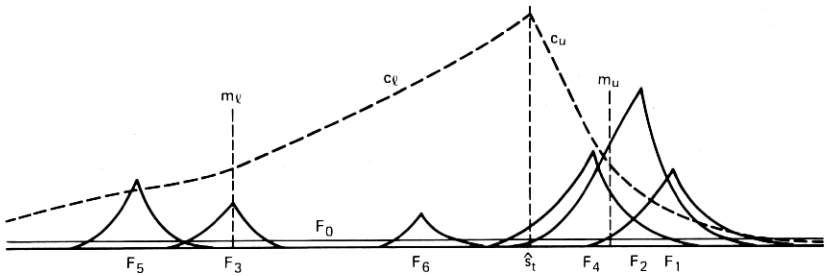

Fig. 7—Illustration of encoder's derivation of predictor and quantizer. F_0 refers to the distribution of the white source output; $\hat{s}_t$ is the weighted sum predictor; m_u and m_e are the upper and lower first moments of the actual distribution about $\hat{s}_t$; and c_u and c_e are the exponential distributions used to determine the quantizer.

respond to edges. Because of this interaction, the values of α were selected experimentally based upon visual examination of a sequence of coded pictures for the two- and three-bit/pel quantizers. For the two-bit/pel quantizer, α was selected so that the minimum value of the inner quantizer level is equal to two picture levels, when the picture is initially quantized into 256 levels. For the three-bit/pel quantizer, the inner quantization level was selected so that the inner quantization level is equal to one picture level.

The random variable F_0 in the source model of Fig. 2 is uniformly distributed over the range of possible values the sample can assume. In implementing the codec, it was found to be desirable to assume that the range of F_0 is somewhat reduced. Limiting the span of F_0 is particularly necessary in the system which transmits two bits per pel. This can be seen as follows. Assume that probability $P(0, t)$ is estimated by the encoder to be close to unity. In this situation, if F_0 has range [0, 255] the four quantization levels will be spread over the entire range of possible sample values. It is then likely that none of the estimators will be close to the reconstructed value x_t even though an estimator can have closely approximated the actual value s_t. Thus, the random estimator may be used for the next sample. This creates an instability in the coder which can propagate into flat regions of the picture. To eliminate this type of instability, the maximum range of the quantizer was limited. To be consistent with limiting the maximum range of the quantizer, the span of F_0 was limited to a symmetrical region about $\hat{s}_t$ of (9). The maximum span of the quantizer was also set experimentally. In the two-bit/pel system, the maximum span of the quantizer was set so that the inner level of the quantizer is eight picture levels. And in the three-bit/pel system, the maximum span of the quantizer was set so that the inner level in the quantizer is four picture levels. In the two-bit/pel system, there are only two quantization levels on each side of the predicted value. In this system, the maximum span of the

quantizer determined the ability of the encoder to track sudden changes in the picture. Therefore, it is necessary to make the maximum quantizer span as large as possible, without making the encoder unstable. In the three-bit/pel system, four quantization levels are on each side of the predictor. In this system, restricting the maximum quantizer span was necessary to prevent the quantizer span from frequently exceeding the range of possible picture levels and wasting quantization levels. This is why a smaller maximum value of the inner quantization level was selected for the three-bit/pel system than for the two-bit/pel system.

In Figs. 8 and 9, the quantization span for various parts of the picture in the two- and three-bit/pel systems is shown. In these pictures, the average of the upper and lower quantization spans is displayed. The white areas correspond to the smallest span of the quantizer and the black levels the largest span. It is interesting to note that the resulting quantizer adaptation is similar to that which would be expected if a masking function were used.[5] However, this quantizer adaptation was arrived at strictly by mathematical techniques, minimizing the expected point mean-squared error with a varying probability distribution of next sample values, rather than by the psychovisual considerations used to derive masking functions. This result is

Fig. 8—Quantizer range adaptation of the two-bit/pel codec.

Fig. 9—Quantizer range adaptation of the three-bit/pel codec.

consistent with Graham's early observations concerning the strong connection between image chaos and the visual system's tolerance to noise-like distortion.[2]

V. RESULTS

This adaptive predictor with a two- and three-bit/pel quantizer has been implemented and compared with an adaptive predictor using Graham's rule[2] and the three-bit/pel fixed quantizer suggested in the Graham paper, and a previous element DPCM encoder with a fixed three-bit/pel quantizer. Our two-bit/pel adaptive predictor has considerably less slope overload than the previous element predictor having three bit/pel quantizer, but is not quite as good as the Graham predictor having a three bit/pel quantizer. Our adaptive predictor with a three bit/pel quantizer has less slope overload than the Graham predictor with a three bit/pel quantizer. In addition, the estimates at edges within the picture are accurate enough to virtually eliminate the edge business in moving sequences which is characteristic of many adaptive predictors. To demonstrate these characteristics, the difference between the original picture of the checker girl, Fig. 1, and the result of processing by these four techniques is shown in Figs. 10 to 13.

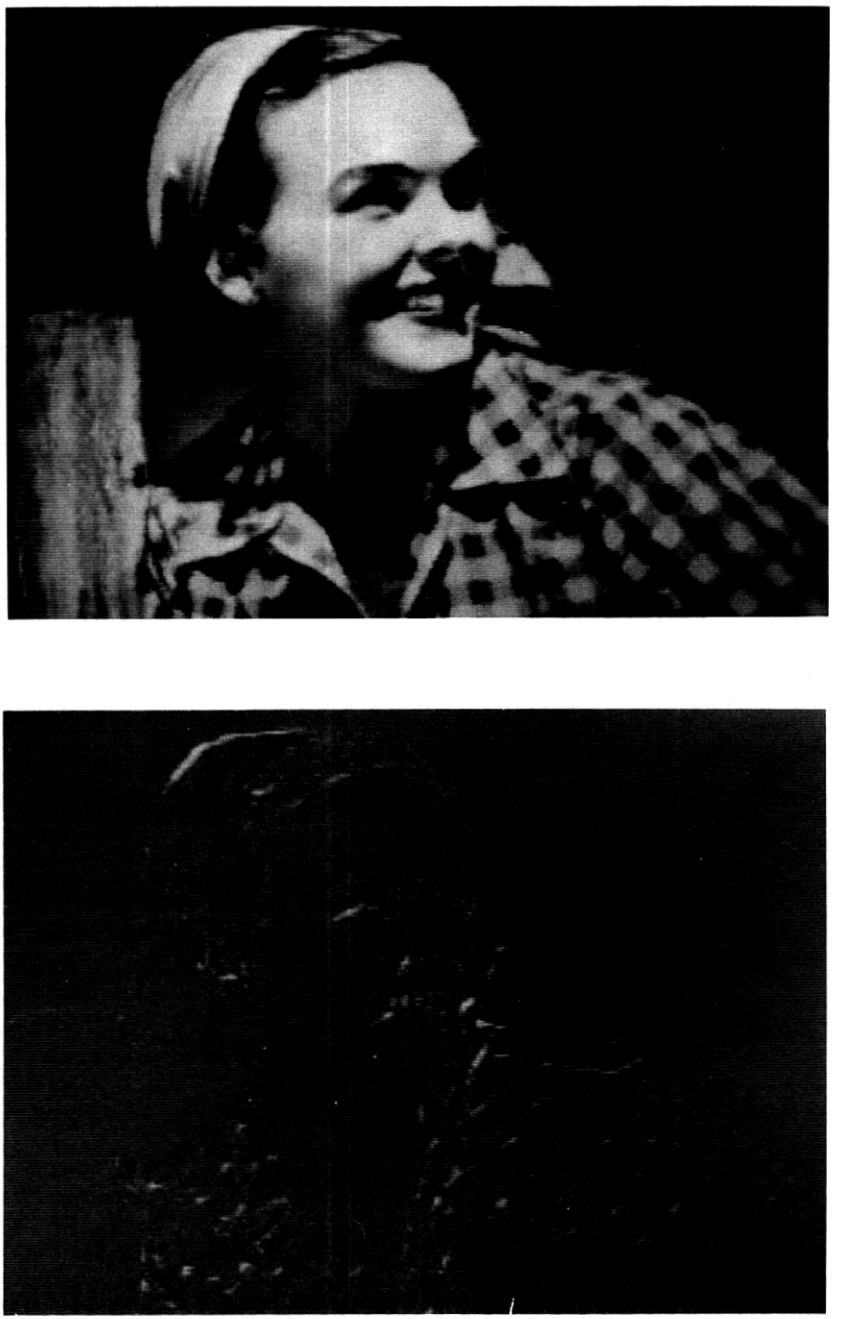

Fig. 10—Two-bit/pel codec performance. Top: Decoder output. Bottom: Difference between decoder output and original.

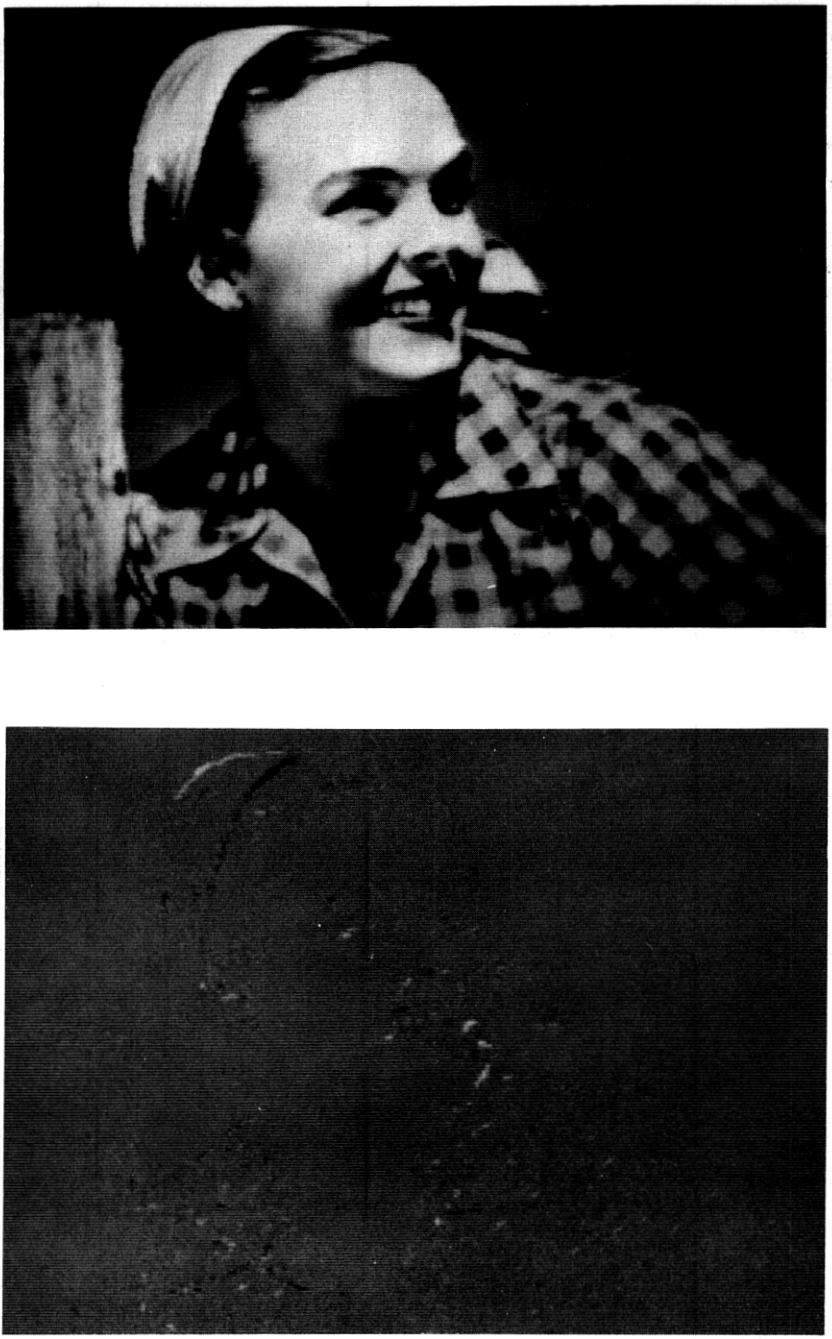

Fig. 11—Three-bit/pel codec performance. Top: Decoder output. Bottom: Difference between decoder output and original.

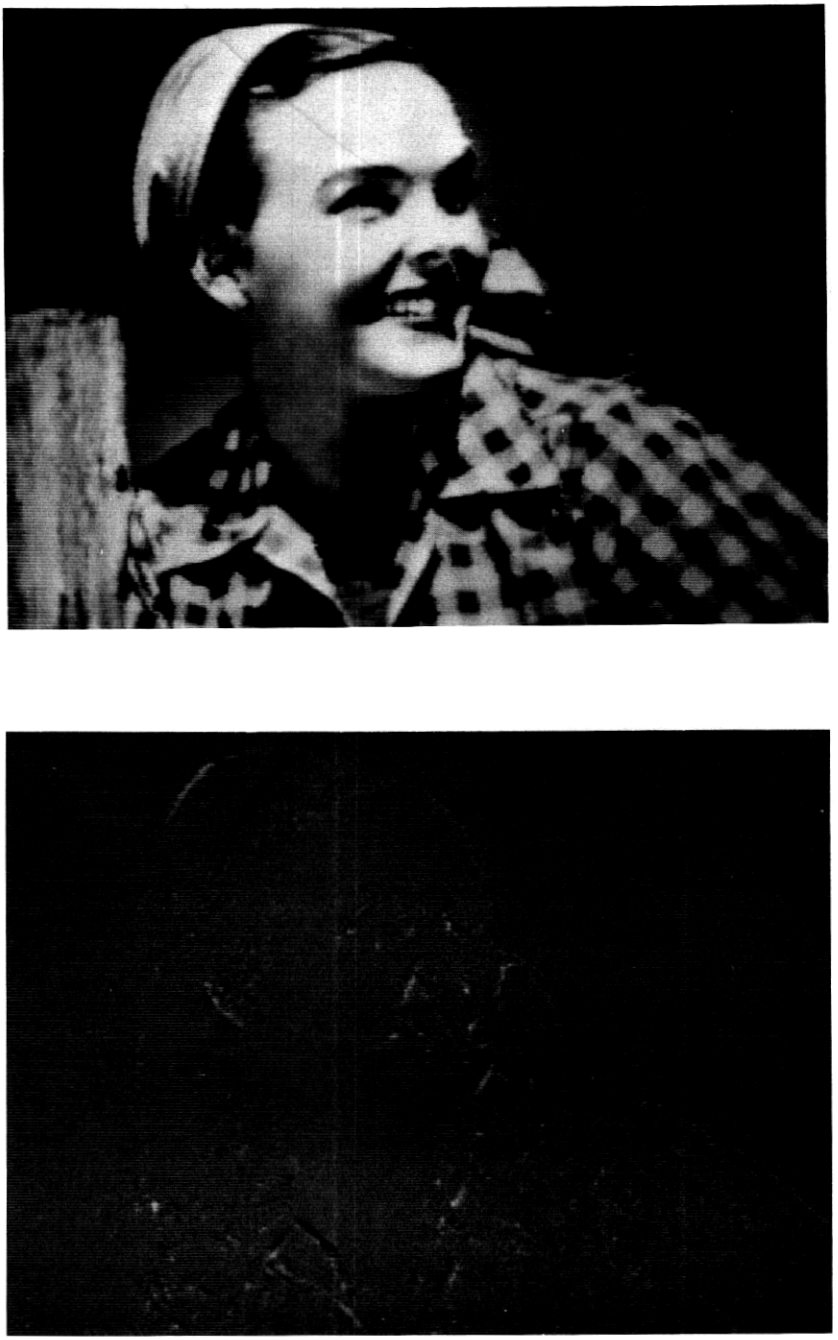

Fig. 12—Performance of three-bit Graham codec. Top: Decoder output. Bottom: Difference between decoder output and original.

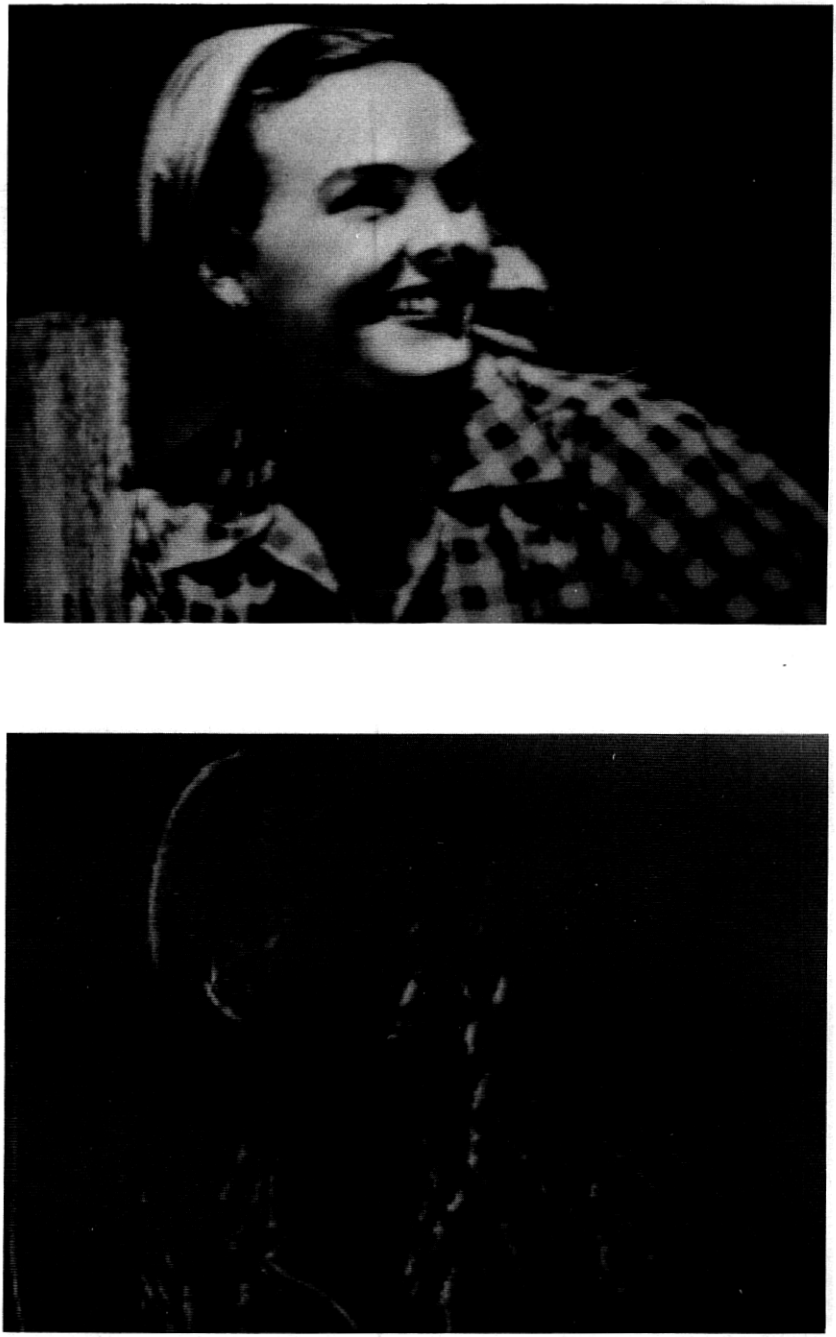

Fig. 13—Performance of three-bit previous element DPCM. Top: Decoder output. Bottom: Difference between decoder output and original.

APPENDIX

This appendix traces the development from eq. (4) to (7) of Section III.

There are $M = (Q + 1)^N$ possible vectors $\mathbf{V} = (q_1, q_2, \cdots q_j, \cdots q_N)$ of source options in the N point region R_t of Fig. 4. Number these vectors $i = 1, 2, \cdots, M$ and let $\mathbf{V}_i$ denote the ith vector. Define $n(q \mid \mathbf{V}_i)$ as the number of components in $\mathbf{V}_i$ that equal the specific value q. Then (4) becomes

$$\hat{P}(q; t) = E\left|\frac{n(q)}{N}\right| \tag{10a}$$

$$= \frac{1}{N}\sum_{i=1}^{M} n(q \mid \mathbf{V}_i)\bar{P}(\mathbf{V}_i \mid S_t^-). \tag{10b}$$

In taking the expectation in (10a) we have treated $t \equiv (m, n)$ as a randomly chosen raster point for which $\bar{P}(q) \equiv E\{P(q; t)\}$ of eq. (1) applies. Because the q are selected independently, $\hat{P}(\mathbf{V}_i \mid S_t^-)$ of (10b) is related to $\hat{P}(q)$ by

$$\hat{P}(\mathbf{V}_i \mid S_t^-) = \frac{P(S_t^- \mid \mathbf{V}_i)}{P(S_t^-)}\bar{P}(\mathbf{V}_i) \tag{11a}$$

$$= \frac{P(S_t^- \mid \mathbf{V}_i)}{P(S_t^-)}\prod_{j=1}^{N}\bar{P}(q_{ij}), \tag{11b}$$

where $p(S_t^- \mid (\cdot))$ denotes the probability density function of the vector of values in S_t^-, and q_{ij} is the jth component of $\mathbf{V}_i$. [The random selection of t cannot affect the independence of the components of $\mathbf{V}_i$ in (11a); hence, (11b)].

Substituting

$$n(q \mid \mathbf{V}_i) = \sum_{j=1}^{N}\delta_{q-q_{ij}} \tag{12}$$

into (10b) and summing over i, (10b) becomes

$$\hat{P}(q; t) = \frac{1}{N}\sum_{j=1}^{N}\bar{P}(q_j = q \mid S_t^-), \tag{13}$$

where $\bar{P}(q_j = q \mid S_t^-)$ is the conditional probability that the jth pixel in R_t was generated by source q. Note that (13) gives $\hat{P}(q; t)$ as an average of *a posteriori* probabilities of q over the region R_t, where

$$\bar{P}(q_j = q \mid S_t^-) = \frac{P(S_t^- \mid q_j = q)}{P(S_t^-)}\bar{P}(q). \tag{14}$$

We now partition the set of pixels S_t^- into (*i*) a subset of pixels future to s_j but past to S_t (call it S_j^+); (*ii*) the pixel s_j; and (*iii*) a subset of

pixels S_t^j past to pixel s_j. The elements in S_j^+ when conditioned on s_j and S_t^j do not depend upon q_j, and it follows after straightforward manipulations that

$$\bar{P}(q_j = q \mid S_t^-) = K\, g(s_t - F_q(S_t^j))\bar{P}(q); \quad 1 \leq q = \leq Q,$$

and

$$\bar{P}(q_j = 0 \mid S_t^-) = \frac{K}{255}\bar{P}(0), \tag{15}$$

where K is a normalizing constant. Substitution of eqs. (15) and (1) into (10) yields (5).

REFERENCES

1. Z. L. Budrikis, "Visual Fidelity Criterion and Modeling," Proc. IEEE, *60* (July 1972), pp. 771 to 779.
2. R. E. Graham, "Predictive Quantizing of Television Signals," IRE Wescon Convention Record, Part 4, 1958, pp. 142 to 157.
3. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, *63* (April 1975), pp. 561 to 580.
4. J. Max, "Quantizing for Minimum Distortion," IEEE, Trans. Inform. Theory, *IT-21* (July, 1975), pp. 373 to 378.
5. A. N. Netravali and B. Prasada, "Adaptive Quantization of Picture Signals Using Spatial Masking," Proc. IEEE, *65* (April 1977), pp. 536 to 548.