# OPTIMAL BIT ALLOCATION AMONG DEPENDENT QUANTIZERS FOR THE MINIMUM MAXIMUM DISTORTION CRITERION

*Guido M. Schuster*[1] *Aggelos K. Katsaggelos*[2]

[1]U.S. Robotics, Network System Div., Skokie, IL, USA, gschuste@usr.com
[2]Northwestern University, Electrical and Computer Engineering Dept., Evanston, IL, USA, aggk@ece.nwu.edu

## ABSTRACT

In this paper we introduce an optimal bit allocation scheme for dependent quantizers for the minimum maximum distortion criterion. First we show how minimizing the bit rate for a given maximum distortion can be achieved in a dependent coding framework using dynamic programming (DP). Then we employ an iterative algorithm to minimize the maximum distortion for a given bit rate, which invokes the DP scheme. We prove that it converges to the optimal solution. Finally we present a comparison between the minimum total distortion criterion and the minimum maximum distortion criterion for the encoding of an H.263 Intra frame. In this comparison we also point out the similarities between the proposed minimum maximum distortion approach and the Lagrangian multiplier based minimum total distortion approach.

## 1. INTRODUCTION

There exists an inherent tradeoff between the rate and the distortion of a lossy compression scheme. One common approach to mathematically formulate this relationship is to minimize the total distortion for a given maximum bit rate, or vice versa, to minimize the bit rate for a given maximum distortion. This bit allocation problem has been studied in [1] for independent quantizers and in [2] for dependent quantizers. The philosophy behind the minimum total distortion approach is that if the average (total) distortion is minimized then, in the long run, the best quality is obtained. As shown in [3] the Lagrangian multiplier method is well suited for these kind of constrained optimization problems. It is important to note that with this popular approach, a large variability among the distortions of the different sources is possible. When the sources are consecutive in time or space, such as different frames of a sequence or different blocks in a frame, this variability in quality can be very disturbing and the perceived quality is fairly low even though the average distortion is minimized.

A different approach to formalize the relationship between the rate and the distortion is to minimize the maximum distortion for a given bit rate, or vice versa, to minimize the bit rate for a given maximum distortion. The philosophy behind this approach is that by minimizing the maximum distortion, no single source distortion will be extremely bad and hence the overall quality will be almost constant. The minimum maximum distortion criterion is a good choice if the goal is to achieve an almost constant distortion which is as small as possible for the available bit rate.

The minimum maximum distortion problem for *independent* quantizers is studied in [4] and a solution which is based on an iterative descent procedure is presented. The procedure is quite simple in that one starts by giving zero bits to each source and then allocates enough bits to the source with the highest distortion so that its distortion is reduced by the smallest amount possible. Then the source with the largest distortion is found and again bits are allocated to that source until its distortion drops. This is repeated until all bits are used up. In [4] an extension of this scheme to dependent sources is suggested using a dual formulation but it is deemed as formidably complicated due to the number of variables involved.

In this paper, we propose a minimum maximum distortion optimal bit allocation algorithm for *dependent* quantizers which is not based on the dual formulation, but uses an iterative scheme which invokes a DP algorithm. We show that the proposed algorithm converges to the optimal solution and is computationally efficient.

The paper is organized as follows. In section 2. we introduce the notation and assumptions. In section 3. we propose a solution to the minimum rate case, which is based on DP. In section 4. we develop a solution to the minimum distortion case. The proposed algorithm is based on an iterative algorithm and the optimal solution to the minimum rate case. In section 5. we discuss an example of the proposed minimum maximum distortion approach. Finally in section 6. we compare the results of the proposed approach with the minimum total distortion approach and present our conclusions.

## 2. NOTATION AND ASSUMPTIONS

We assume that the rate $r_i(\cdot)$ and the distortion $d_i(\cdot)$ of a source $i$ depend on the quantizer selections of neighboring sources in a neighborhood defined by two non-negative integers $a$ and $b$. Therefore the total rate $R(\cdot)$, which is the sum of the source rates, is defined as follows,

$$R(x_0,\ldots,x_{N-1}) = \sum_{i=0}^{N-1} r_i(x_{i-a},\ldots,x_{i+b}), \qquad (1)$$

and the total distortion $D(\cdot)$, which is the maximum of the source distortions, is defined as follows,

$$D(x_0,\ldots,x_{N-1}) = \max_{i\in[0,\ldots,N-1]}\{d_i(x_{i-a},\ldots,x_{i+b})\}, \qquad (2)$$

where $x_i$ is the quantizer of source $i$ selected from the set of admissible quantizers $X_i$ for source $i$ and $N$ is the total number of sources. The quantizers $x_{-a},\ldots,x_{-1}$ and $x_N,\ldots,x_{N-1+b}$ are the boundary parameters and can be set to any desired value.

## 3. THE MINIMUM RATE CASE

First we consider the case of finding the minimum rate for a given maximum distortion $D_{max}$. This can be formulated as follows,

$$\min_{x_0,\ldots,x_{N-1}} R(x_0,\ldots,x_{N-1}) \quad \text{s.t.:} \quad D(x_0,\ldots,x_{N-1}) \le D_{max}. \qquad (3)$$

The key observation for the proposed optimal bit allocation procedure is that the maximum distortion $D_{max}$ is a constraint which applies to each source distortion $d_i(x_{i-a},\ldots,x_{i+b})$, as is clear from Eq. (2). Therefore any quantizer selection which results in a single source distortion being larger than $D_{max}$ cannot belong to the optimal quantizer sequence. Of the remaining quantizer sequences, the one which results in the smallest rate is the optimal solution to the above problem. These two concepts can be combined using the following definitions,

$$g_i(x_{i-a},\ldots,x_{i+b}) = \quad (4)$$
$$\begin{cases} \infty & : \quad d_i(x_{i-a},\ldots,x_{i+b}) > D_{max} \\ r_i(x_{i-a},\ldots,x_{i+b}) & : \quad d_i(x_{i-a},\ldots,x_{i+b}) \le D_{max} \end{cases},$$

and

$$G(x_0,\ldots,x_{N-1}) = \sum_{i=0}^{N-1} g_i(x_{i-a},\ldots,x_{i+b}). \quad (5)$$

Finding the minimum of the objective function in Eq. (5) is equivalent to the original problem of finding

$$R^*(D_{max}) = \min_{x_0,\ldots,x_{N-1}} R(x_0,\ldots,x_{N-1}) \quad (6)$$
$$\text{s.t.:} \quad D(x_0,\ldots,x_{N-1}) \le D_{max},$$

where $R^*(D_{max})$ is the smallest rate for a given maximum distortion $D_{max}$.

We find the minimum of the objective function in Eq. (5) using DP. We denote by $g_l^*(x_{l-a+1},\ldots,x_{l+b})$ the minimum of the partial objective function up to and including neighborhood $l$, that is

$$g_l^*(x_{l-a+1},\ldots,x_{l+b}) = \min_{x_0,\ldots,x_{l-a}} \sum_{i=0}^{l} g_i(x_{i-a},\ldots,x_{i+b}). \quad (7)$$

From Eq. (7) it follows that,

$$g_{l+1}^*(x_{l+1-a+1},\ldots,x_{l+1+b}) \quad (8)$$
$$= \min_{x_0,\ldots,x_{l+1-a}} \sum_{i=0}^{l+1} g_i(x_{i-a},\ldots,x_{i+b})$$
$$= \min_{x_{l+1-a}} \left[ \min_{x_0,\ldots,x_{l-a}} \left( \sum_{i=0}^{l} g_i(x_{i-a},\ldots,x_{i+b}) + g_{l+1}(x_{l+1-a},\ldots,x_{l+1+b}) \right) \right].$$

Since $g_{l+1}(x_{l+1-a},\ldots,x_{l+1+b})$ does not depend on $x_0,\ldots,x_{l-a}$, it can be moved outside the inner minimization. Then the resulting inner minimization is equal to $g_l^*(x_{l-a+1},\ldots,x_{l+b})$ in Eq. (7) and the following DP recursion formula results,

$$g_{l+1}^*(x_{l+1-a+1},\ldots,x_{l+1+b}) = \min_{x_{l+1-a}} \quad (9)$$
$$[g_l^*(x_{l+1-a},\ldots,x_{l+b}) + g_{l+1}(x_{l+1-a},\ldots,x_{l+1+b})]$$

Having established the DP recursion formula, the forward DP algorithm can be used to find the optimal solution to problem (5). Note that the forward DP algorithm is also called the Viterbi algorithm. First, the recursion needs to be initialized,

$$g_{a-1}^*(x_0,\ldots,x_{a+b-1}) = \sum_{i=0}^{a-1} g_i(x_{i-a},\ldots,x_{i+b})$$
$$\forall\, [x_0,\ldots,x_{a+b-1}] \in X_0 \times \ldots \times X_{a+b-1}. \quad (10)$$

We introduce a back pointer which will be used to remember the optimal selection,

$$i_{a-1}(x_0,\ldots,x_{a+b-1}) = [x_{-a},\ldots,x_{b-1}]. \quad (11)$$

Next, the recursion is started, hence the DP recursion formula (9) is applied for $l = a-1$ up to and including $l = N-2$, that is,

$$g_{l+1}^*(x_{l+1-a+1},\ldots,x_{l+1+b}) = \min_{x_{l+1-a}} \quad (12)$$
$$[g_l^*(x_{l+1-a},\ldots,x_{l+b}) + g_{l+1}(x_{l+1-a},\ldots,x_{l+1+b})],$$
$$\forall\, [x_{l+1-a+1},\ldots,x_{l+1+b}] \in X_{l+1-a+1} \times \ldots \times X_{l+1+b}.$$

Again the back pointer is assigned, using the argument $x_{l+1-a}^*(x_{l+1-a+1},\ldots,x_{l+1+b})$ which minimizes Eq. (12),

$$i_{l+1}(x_{l+1-a+1},\ldots,x_{l+1+b}) \quad (13)$$
$$= [x_{l+1-a}^*(x_{l+1-a+1},\ldots,x_{l+1+b}), x_{l+1-a+1},\ldots,x_{l+b}],$$
$$\forall\, [x_{l+1-a+1},\ldots,x_{l+1+b}] \in X_{l+1-a+1} \times \ldots \times X_{l+1+b}.$$

Then the final solution is found by observing that,

$$\min_{x_0,\ldots,x_{N-1}} G(x_0,\ldots,x_{N-1})$$
$$= \min_{x_{N-a},\ldots,x_{N-1}} g_{N-1}^*(x_{N-a},\ldots,x_{N-1+b}). \quad (14)$$

The arguments $[x_{N-a}^*,\ldots,x_{N-1}^*]$ which minimize Eq. (14) are required for the final back pointer,

$$i_N = [x_{N-a}^*,\ldots,x_{N-1}^*, x_N,\ldots,x_{N-1+b}]. \quad (15)$$

Note that $x_N,\ldots,x_{N-1+b}$ are fixed parameters and hence they are identical to their optimal values $x_N^*,\ldots,x_{N-1+b}^*$. Using the described approach, we can find the minimum value of the objective function, but we also need to know the optimal quantizers $x_0^*,\ldots,x_{N-1}^*$. These quantizers can be found by following the back pointers during the backtracking stage. Note that $i_N$ identifies the optimal quantizers $x_{N-a}^*,\ldots,x_{N-1+b}^*$. The remaining quantizers can be found as follows. For $l = N-1,\ldots,a$

$$x_{l-a}^* = [i_l(x_{l-a+1}^*,\ldots,x_{l+b}^*)]_1, \quad (16)$$

where $[\cdot]_1$ refers to the first element in the vector.

If the optimal solution $[x_0^*,\ldots,x_{N-1}^*]$ results in an infinite rate $R^*(D_{max})$, then there exits no solution which can satisfy the maximum distortion requirement for each source. When the rate for the optimal solution is finite, then the solution optimally selects the admissible quantizers such that the rate $R^*(D_{max})$ is minimized for a given maximum distortion $D_{max}$.

## 4. THE MINIMUM DISTORTION CASE

In the minimum distortion case, the maximum number of bits ($R_{max}$) is given and the goal is to select the quantizers in such a fashion that the resulting maximum distortion is as small as possible. Mathematically this can be expressed as follows,

$$\min_{x_0,\ldots,x_{N-1}} D(x_0,\ldots,x_{N-1}) \quad \text{s.t.:} \quad R(x_0,\ldots,x_{N-1}) \le R_{max}. \quad (17)$$

The main difference to the minimum rate problem is the fact that the maximum rate $R_{max}$ is not an upper limit for each source, but for the sum of the sources. Also this is a minimum maximum problem, since the total distortion is defined as the maximum over all source distortions.

The proposed optimal bit allocation algorithm for the minimum distortion case is based on the fact that we can

optimally solve the minimum rate case. In other words, for every given $D_{max}$ we can find the quantizer sequence which results in $R^*(D_{max})$, the minimum rate for encoding the combined sources, where each source distortion has to be below the maximum distortion $D_{max}$. We use the following lemma to formulate an iterative procedure to find the optimal solution for the minimum distortion problem.

**Lemma 1** *$R^*(D_{max})$ is a non-increasing function of $D_{max}$.*

Proof: Let $D^2_{max} \geq D^1_{max}$, $[{}^1x^*_0, \ldots, {}^1x^*_{N-1}]$ be the optimal solution of Eq. (6) for $D_{max}$=$D^1_{max}$, and $[{}^2x^*_0, \ldots, {}^2x^*_{N-1}]$ the optimal solution of Eq. (6) for $D_{max}$=$D^2_{max}$. Since $D^1_{max} \leq D^2_{max}$, $[{}^1x^*_0, \ldots, {}^1x^*_{N-1}]$ is a possible solution of Eq. (6) for $D_{max}$=$D^2_{max}$, using $R^*(D^1_{max})$ bits. Since $[{}^2x^*_0, \ldots, {}^2x^*_{N-1}]$ is the optimal solution of Eq. (6) for $D_{max}$=$D^2_{max}$, it follows that $R^*(D^2_{max}) \leq R^*(D^1_{max})$. ■

The above lemma is intuitively clear since it simply states that if a greater maximum error is permissible, then we should be able to encode the sources with a smaller number of bits. Note that even though this seems obvious, this only holds true because we can solve the minimum rate case optimally.

Having shown that $R^*(D_{max})$ is a non-increasing function, we can use the bisection method to find the optimal $D^*_{max}$ such that $R^*(D^*_{max}) = R_{max}$, which solves the minimum distortion problem of Eq. (17). The bisection method starts with two points $(D^l_{max}, R^*(D^l_{max}))$ and $(D^u_{max}, R^*(D^u_{max}))$ which bracket the optimal solution (see Fig. 1). Then a middle point $(D^m_{max}, R^*(D^m_{max}))$ is found by invoking the minimum rate algorithm for $D_{max} = D^m_{max} = (D^l_{max}+D^u_{max})/2$. The new bracketing points of the optimal solution are then the middle point and the one of the original points which is closer to the optimal solution. This procedure is then iterated until the optimal solution is found or the bracket is small enough for the purpose at hand.

Since this is a discrete optimization problem, the function $R^*(D_{max})$ is not continuous and exhibits a staircase characteristic (see Fig. 1). This implies that there might not exist a $D^*_{max}$ such that $R^*(D^*_{max}) = R_{max}$. In this case, the proposed algorithm will still find the optimal solution, which is of the form $R^*(D^*_{max}) < R_{max}$, but only after an infinite number of iterations. This is true, since after every iteration, the length of the interval which contains the optimal solution is cut in half. Hence after any finite number of iterations, the interval still contains infinitely many possible solutions, but after an infinite number of iterations, the length of the interval is zero and the optimal solution is found. In practice, if we have not found a $D_{max}$ such that $R^*(D_{max}) = R_{max}$ after a given maximum number of iterations, we terminate the algorithm.

## 5. EXAMPLE

In this section we present an example to compare the minimum total (or average) distortion and the minimum maximum distortion approaches. The dependent image coding scheme we use for this example is the intra frame scheme employed in TMN4 [5], which is the test model four of the H.263 standard.

We encode the first frame of the QCIF color sequence "Mother and Daughter" using this scheme. We use the TMN4 mechanism for transmitting the quantizer step sizes which is based on a modified delta modulation scheme. In TMN4, the quantizer step size of the current macro block must be within ±2 of the quantizer step size employed for the previous macro block. Then the difference between the quantizer step sizes is entropy coded. This DPCM scheme results in a first order dependency between two consecutive blocks, since the operational rate distortion curve of the current block depends on the quantizer selected for the previous block.

First we fix the quantizer step size for all macro blocks to 10. The resulting rate ($R_{Q=10}$ = 18297 bits) and distortion are listed in Table 1. Note that the mean squared error (MSE) of the luminance (Y) channel is used as the distortion measure. In the case where we minimize the total distortion subject to a maximum bit rate (in this example, $R_{max} = R_{Q=10}$), the problem can be formulated using the Lagrangian multiplier method [6]. The solution of the relaxed (unconstrained) problem can then be found using DP. Again, the resulting rate and distortion are in Table 1 and so are the results for the proposed minimum maximum distortion scheme, where again the maximum rate $R_{max}$ was set equal to $R_{Q=10}$.

In Fig. 2 the MSE per macro block for the three implementations is shown and in Fig. 3 the corresponding quantizer selections are displayed. It is interesting to notice in Fig. 3 that there are quite a few blocks where the quantizers are the same for both optimal schemes. These blocks tend to coincide with the blocks where the MSE (see Fig. 2) is very small, i.e., blocks with no high frequency components. It is clear from Fig. 2 that the minimum maximum distortion scheme results in a more even quality for the entire frame than the minimum total distortion approach.

## 6. CONCLUSIONS

A minimum maximum distortion approach for the optimal bit allocation among dependent quantizers has been presented. This approach has similarities with the minimum total approach. For the later, an iterative scheme (for example, the very fast convex search presented in [7], or bisection) is needed to find the optimal tradeoff parameter $\lambda$, where for each iteration the relaxed problem is solved by DP. For the former, bisection is employed, where for each iteration, the minimum rate problem is solved optimally using DP. One of the main differences is that the Lagrangian multiplier approach can only find solutions which belong to the convex hull, whereas the proposed minimum maximum distortion scheme will always find the optimal solution. Furthermore, in the proposed approach, the minimum rate case can be solved without an iteration, whereas the minimum total distortion approach always requires an iterative search.

The minimum maximum distortion and minimum total distortion approaches offer different paradigms for formulating the inherent tradeoff between rate and distortion. This is also evident in Table 1. Clearly the minimum maximum distortion approach results in a higher average distortion, on the other hand, its maximum distortion is much lower than the maximum distortion of the minimum total distortion approach. This is also reflected in the smaller standard deviation. In other words, the minimum maximum distortion approach results, for example, in an encoded image, with more uniform quality than the one resulting from the minimum total distortion approach.

## REFERENCES


[1] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[2] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans-*

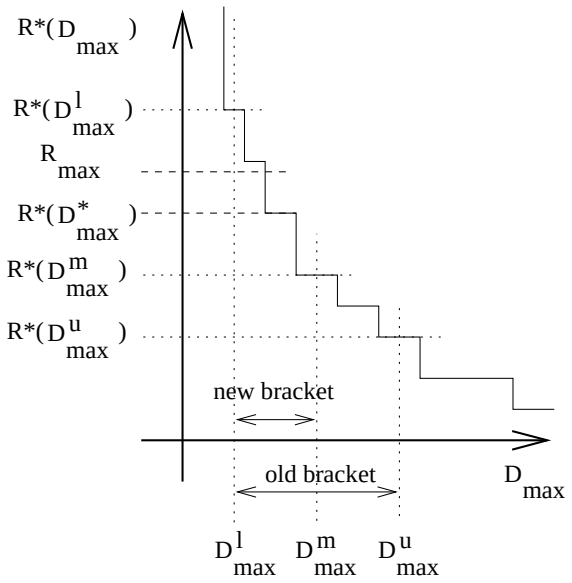


**Figure 1. The $R^*(D_{max})$ function, which is a non-increasing function, exhibits a staircase characteristic. The selected $R_{max}$ falls onto a discontinuity hence the optimal solution is of the form $R^*(D^*_{max}) < R_{max}$, instead of $R^*(D^*_{max}) = R_{max}$.**

| | Rate | Distortion (MSE) | | | |
|---|---|---|---|---|---|
| | | mean | min | max | std |
| Q=10 | 18297 | 27.8 | 0.6 | 66.8 | 17.5 |
| min total | 18431 | 27.1 | 0.6 | 65.0 | 17.3 |
| min max | 18293 | 29.9 | 0.6 | 46.2 | 15.9 |

**Table 1. Each of the 99 macro blocks ($16 \times 16$) results in a particular MSE, and the mean MSE is the mean of these 99 MSEs. The same holds for the minimum, maximum and standard deviation column.**

*actions on Image Processing*, vol. 3, pp. 533–545, Sept. 1994.

[3] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, pp. 399–417, 1963.

[4] D. W. Lin, M.-H. Wang, and J.-J. Chen, "Optimal delayed-coding of video sequences subject to a buffer size constraint," in *Proceedings of the Conference on Visual Communications and Image Processing*, vol. 2094, pp. 223–234, SPIE, 1993.

[5] Expert's Group on Very Low Bitrate Visual Telephony, *Video Codec Test Model, TMN4 Rev1.* ITU Telecommunication Standardization Sector, Oct. 1994.

[6] G. M. Schuster and A. K. Katsaggelos, "A video compression scheme with optimal bit allocation between displacement vector field and displaced frame difference," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 1996.

[7] G. M. Schuster and A. K. Katsaggelos, "Fast and efficient mode and quantizer selection in the rate distortion sense for H.263," in *Proceedings of the Conference on Visual Communications and Image Processing*, pp. 784–795, SPIE, Mar. 1996.

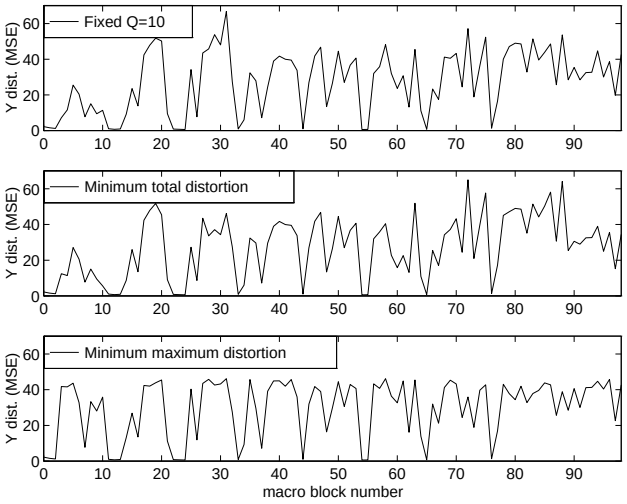


**Figure 2. MSE of each macro block of the luminance channel; First row: MSE for a fixed quantizer step size of Q=10. Second row: MSE for the minimum total distortion approach. Third row: MSE for the minimum maximum distortion approach.**

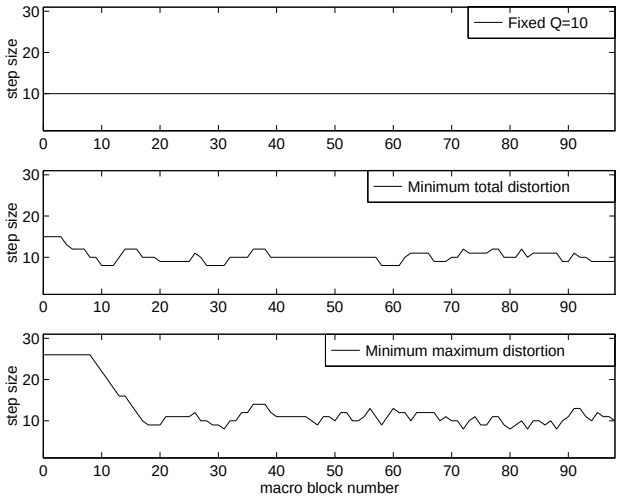


**Figure 3. Macro block quantizer step sizes; First row: fixed quantizer step size of Q=10. Second row: step sizes for the minimum total distortion approach. Third row: step sizes for the minimum maximum distortion approach.**