Operational Rate-Distortion Optimal Bit Allocation between Shape and Texture for MPEG-4 Video Coding

Haohong Wang, Guido M. Schuster, and Aggelos K. Katsaggelos*

Image and Video Processing Lab (IVPL)
Department of Electrical and Computer Engineering
Northwestern University, Evanston, IL 60208, USA
Email: {haohong, aggk}@ece.northwestern.edu

* Abteilung Elektrotechnik
Hochschule für Technik, Rapperswil
CH-8040 Rapperswil, Switzerland
Email: guido.schuster@hsr.ch

ABSTRACT

MPEG-4 is the first multimedia standard that supports the decoupling of a video object into object shape and object texture information, which consequently brings up the optimal encoding problem for object-based video. In this paper, we present an operational rate-distortion optimal bit allocation scheme between shape and texture for MPEG-4 encoding. Our approach is based on the Lagrange multiplier method, while the adoption of dynamic programming techniques enables its higher efficiency over the exhaustive search algorithm. Our work will not only benefit the further study of joint shape and texture encoding, but also make possible the deeper study of optimal joint source-channel coding of object-based video.

1. INTRODUCTION

MPEG-4 [1-2] is the first international video communication standard that enables object-based video coding. By transmitting object information along with texture, MPEG-4 enables content-based interactivity, which will greatly benefit a number of multimedia applications, as well as, computer games and related applications with increased interactivity with the audio-visual content.

In order to build an efficient MPEG-4 video encoder, the fundamental problem of optimal bit allocation between shape and texture has to be solved first. The difficulty of this problem lies in that the coding of the object shape is not independent of the coding of its texture, in other words, a jointly optimal coding scheme of shape and texture needs to be developed. In [3], an optimal vertex-based shape encoder is proposed, taking into consideration the texture information of the video frames. By utilizing a variable-width tolerance band, which is proportional to the degree of trust in the accuracy of the shape information at that location, the encoder spends more bits on parts of the object boundary where high accuracy is required, while fewer bits on other less important parts. In [4], a joint shape and texture rate control algorithm is proposed for MPEG-4 encoders, and in [5], an improved statistical rate-distortion model is presented for the encoding of MPEG-4 object shape. However, none of these approaches optimally solved the bit allocation problem.

In this paper, we are proposing an operational rate-distortion optimal bit allocation scheme between shape and texture for encoding of MPEG-4 object-based video. The algorithm is based on the Lagrange multiplier method. Dynamic programming is applied to significantly reduce the computational complexity of our approach.

The rest of the paper is organized as follows. The next section provides a brief overview of the MPEG-4 video coding algorithms. Since we only consider intra mode, only this part is covered. In section 3, the problem formulation is presented. Section 4 demonstrates the optimal solution. Section 5 provides our experimental results, and we draw conclusions in the last section.

2. OVERVIEW OF MPEG-4 VIDEO CODEC

In this section, the relevant parts of MPEG-4 video coding algorithms are reviewed. Further details can be found in [1-2].

2.1. Shape coding

The binary shape information is coded utilizing the macroblock-based structure, where binary alpha data are grouped within 16x16 binary alpha blocks (BAB). The BABs can be classified into transparent, opaque and border macroblocks by summing up the pixels within the BAB and then accessing a threshold (*Alpha_TH*). To reduce the bit-rate, lossy representation of a BAB might be adopted. It successively down-samples the original BAB by a conversion ratio factor (CR) of two or four, and then up-samples the approximated sub-image back to the full-resolution. The selection of CR also depends on the threshold *Alpha_TH*. It is important to point out that both values of BAB type and CR are transmitted to the decoder, while the value of *Alpha_TH* is not transmitted. This means that *Alpha_TH* only indirectly affects the rate-distortion characteristic of the video codec, and therefore it is considered in our optimization.

The border macroblocks need to be further processed by context-based arithmetic encoding (CAE). That is, a template of 10 pixels is used to define the context for predicting the alpha value of the current pixel as shown in Fig. 1. Also, a probability table is predefined for the context of each pixel. Then, the sequence of pixels within the BAB drives an arithmetic encoder with its pair of alpha value and probability. The concept of CAE makes the encoding of a BAB depend on its neighbors to the left, above, and above-left.

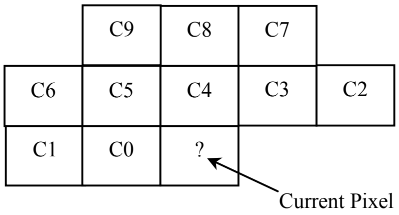


Figure 1. Context for Intra mode

2.2. Texture coding

In MPEG-4, the texture content of the macroblock's bitstream depends to a great extent on the reconstructed shape information. First, low-pass extrapolation (LPE) padding is applied on each 8x8 boundary (non-transparent and non-opaque) blocks. This involves taking the average of all the luminance/chrominance values over all the opaque pixels of the block, and all transparent pixels are given this average value. Then, a 2D 8x8 DCT transform is applied on each block, and the resulting DCT coefficients are quantized. The DC coefficient is quantized (QDC) using a step size of 8. For AC coefficients, MPEG-4 utilizes either the quantization method of H.263 or that of MPEG-2. In this paper, the former method is adopted, that is, a quantization parameter determines the quantization process, and the same value applies to all coefficients in a macroblock but may change from one macroblock to the next. After quantization, the adaptive DC prediction is applied by selecting either the QDC value of immediately previous block or that of the block immediately above it. With the same prediction direction, either coefficients from the first row or the first column of a previously coded block are used to predict the co-sited coefficients of the current blocks. Finally, the differential coefficients after prediction are encoded by VLC encoding. Clearly, the selection criterion involves the current block, the previous block, the block above and to the left, and the block immediately above, which also makes a macroblock depend on its neighbors to the left, above, and above-left.

2.3. Video data structure

It is important to note that MPEG-4, like all video standards, only standardizes the video decoding, while leaves its encoding part open for possible new coming technologies. Our work is therefore based on the standard decoder structure. Figure 2 shows the video data structures for I-VOP.

Bab_type	CR	ST	BAC	MCBP	DC data
DC Marker	AC_pred_flag	CBPY	AC data		

Figure 2. Data partitioned shape texture syntax for I-VOP

(Note: CR=Conversion Ratio, ST=Scan Type, BAC=Block Alpha Code, MCBPC=Macroblock type & Coded Block Pattern for Chrominance,CBPY=Coded Block Pattern for Luminance)

Our optimization work addresses the making of decisions on the coding parameters so that the video is coded to meet certain constraints in video quality, bit rate, etc. From the data structure shown above, the adjustable parameters could only include Bab_type, CR, ST, MCBP, CBPY, and quantization step size, which decides DC and AC data.

3. PROBLEM FORMULATION

Our goal is to control both the shape and texture coding parameters in order to minimize the total (shape plus texture) bit rate required to transmit a video sequences at some acceptable level of quality. We can write this optimization formally as

$$\text{Minimize } R\text{, subject to } D \le D_{max}, \qquad (1)$$

where R is the total bit rate per frame, D is the distortion, and D_{max} is the maximum tolerable distortion. The same techniques can be applied to solve the dual problem, that is,

$$\text{Minimize } D\text{, subject to } R \le R_{budget,}, \qquad (2)$$

where R_{budget} is the bit budget.

Our study of the optimal bit allocation between shape and texture is restricted to the frame level. In other words, we do not attempt to optimally allocate the bits among the different frames of a video sequence. The reader interested in that problem is referred to [6].

3.1 Notation

We assume that the current frame has N macroblocks $m_1, m_2, \ldots, m_N$ in W columns, and they are numbered following the horizontal scan order. Every macroblock has a bab_type $b_i \in B_i$, a CR $c_i \in C_i$, an ST $s_i \in S_i$, an MCBP $p_i \in P_i$, and a CBPY $y_i \in Y_i$ associated with it, where B_i is the set of all admissible BAB type for m_i, C_i is the set of all admissible CR for m_i, S_i is the set of all admissible ST for m_i, P_i is the set of all admissible MCBP for m_i, and Y_i is the set of all admissible CBPY for m_i. Let us define a decision vector $v_i=[b_i, c_i, s_i, p_i, y_i] \in V_i$ for every macroblock m_i. $V_i = B_i \times C_i \times S_i \times P_i \times Y_i$ is the admissible decision vector set for m_i. All macroblocks within the frame share a unique quantizer step size $q \in Q$, where Q is the set of all admissible quantizer step size for the frame.

3.2 Rate

Let us denote by $R_i(v_{i-W-1}, \ldots, v_i)$ the rate for macroblock m_i. As mentioned in section 2 and shown in Fig. 3, the encoding of current macroblock will only depend on macroblock 1, 2 and 3. In other words, after the decision vectors for macroblocks 1, 2, 3 and current macroblock are settled, the rate for the current macroblock is fixed. Clearly,

$$R = R_{syntax} + \sum_{i=1}^{N} R_i(v_{i-W-1}, \ldots, v_i), \tag{3}$$

where R_{syntax} represents the bits allocated on the data structure syntax of the VOP.

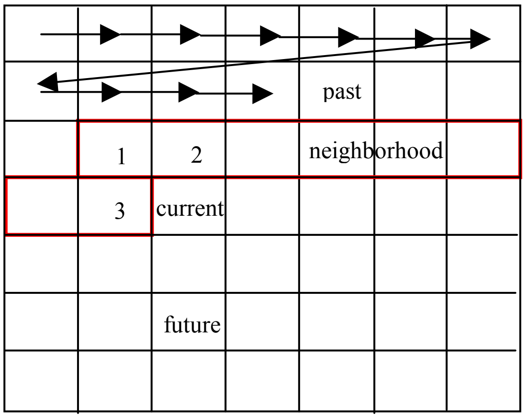


Figure 3 Layout of macroblocks in the VOP

3.3 Distortion

We use the mean-squared error (MSE) as the distortion metric. Let us denote by $D_i(v_{i-W-1}, \ldots, v_i)$ the distortion for macroblock m_i. Clearly,

$$D = \sum_{i=1}^{N} D_i(v_{i-W-1}, \ldots, v_i) \tag{4}$$

and

$$D_i = \sum_{x=0}^{15}\sum_{y=0}^{15} d_{i,Y}(x,y)^2 + \sum_{x=0}^{7}\sum_{y=0}^{7}[d_{i,U}(x,y)^2 + d_{i,V}(x,y)^2]$$

where $d_{i,Y}(x,y)$, $d_{i,U}(x,y)$, and $d_{i,V}(x,y)$ are differential intensity values for the Y, U and V components at pixel (x,y).

The peak signal-to-noise ratio (PSNR) can be obtained by

$$D_{PSNR} = 10\log_{10}\frac{1.5 \times N \times 255^2}{D}, \tag{5}$$

where the factor 1.5 comes from the down sampling of the chrominance components by a factor of 2.

4. OPTIMAL SOLUTION

We derive a solution to problem (1) using the Lagrange multiplier method to relax the constraint, so that the relaxed problem can be solved using a shortest path algorithm. The same steps can be followed in solving the dual problem (2). We first define a Lagrangian cost function

$$J_\lambda(V) = R + \lambda D = R_{syntax} + \sum_{i=1}^{N}[R_i(v_{i-W-1}, \ldots, v_i) + \lambda D_i(v_{i-W-1}, \ldots, v_i)],$$

where λ is the Lagrange multiplier. It has been shown in [7] and [8] that if there is a λ^* such that $V^* = \arg[\min_V J_{\lambda^*}(V)]$, and which leads to $D=D_{max}$, then V^* is also an optimal solution to (1). It is well known that when λ sweeps from zero to infinity, the solution to (1) traces the convex hull of the operational rate distortion function, which is a non-increasing function. Hence, bisection or the fast convex search we present in [9] can be used to find λ^*. Therefore, if we can find the optimal solution to the unconstrained problem

$$\min \sum_{i=1}^{N}[R_i(v_{i-W-1}, \ldots, v_i) + \lambda D_i(v_{i-W-1}, \ldots, v_i)], \tag{6}$$

we can find the optimal λ^*, and the convex hull approximation to the constrained problem (1).

To implement the algorithm to solve the optimization problem (6), we create a cost function $C_k(v_{k-W-1}, \ldots, v_k)$, which represents the minimum total rate and distortion up to and including macroblock m_k given that $v_{k-W-1}, \ldots, v_k$ are decision vectors for macroblocks $m_{k-W-1}, \ldots, m_k$. Clearly,

$$J_\lambda(V) = \min_{v_{N-W-1}, \ldots, v_N} C_N(v_{N-W-1}, \ldots, v_N) \tag{7}$$

The key observation for deriving an efficient algorithm is the fact that given $W+1$ decision vectors $v_{k-W-2}, \ldots, v_{k-1}$ for macroblocks $m_{k-W-2}, \ldots, m_{k-1}$, and the cost function $C_{k-1}(v_{k-W-2}, \ldots, v_{k-1})$, the selection of the next decision vector v_k is independent of the selection of the previous decision vectors $v_1, v_2, \ldots, v_{k-W-3}$. This is true since the cost function can be expressed recursively as

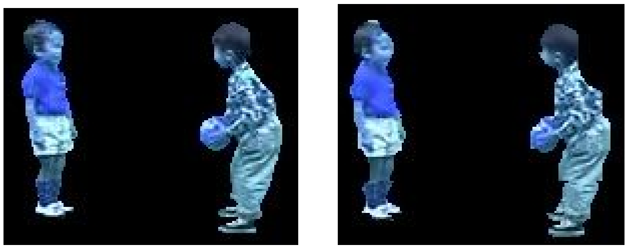
$$C_k(v_{k-W-1}, \ldots, v_k) = \min_{v_{k-W-2}}[C_{k-1}(v_{k-W-2}, \ldots, v_{k-1})] + R_k(v_{k-W-1}, \ldots, v_k) + \lambda D_k(v_{k-W-1}, \ldots, v_k). \tag{8}$$

The recursive representation of the cost function above makes the future step of the optimization process independent from its past step, which is the foundation of the dynamic programming technique. In [10], a similar DP algorithm is applied to solve the bit allocation problem between displacement vector field and displaced frame difference in motion-compensated video coding. The computational complexity of our algorithm is $O(N \times |V_i|^{W+2})$ ($|V_i|$ is the cardinality of V_i), which depends directly on the width of the VOP, but still much more efficient than the exponential computational complexity of an exhaustive search algorithm.

5. EXPERIMENTAL RESULTS

A number of experiments have been conducted, some of which are reported here. Figure 4 shows a comparison of the results obtained using the optimal approach (see Fig.

4(a)) and using Momusys without optimization (see Fig. 4(b)) by given a bit budget of *8,000* bits for the first frame of children sequence.



(a) PSNR=*25.69* (b) PSNR=*21.74*

Figure 4 Comparison of reconstructed frame quality

In the second experiment, the R-D curve obtained by our optimal approach is compared with the results from Momusys by exhaustively trying all combinations of the parameters (*Alpha_TH* and *QP*) as shown in Fig. 5. Clearly, our result in addition to providing solutions on the convex hull of all operating points, also demonstrates some gains in RD quality, due to the selection of adjustable parameters different than *Alpha_TH*.

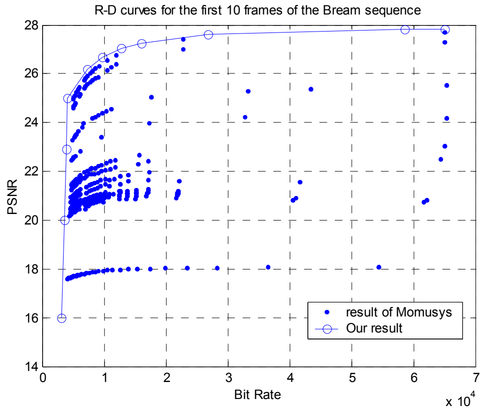


Figure 5 Comparison of R-D curves

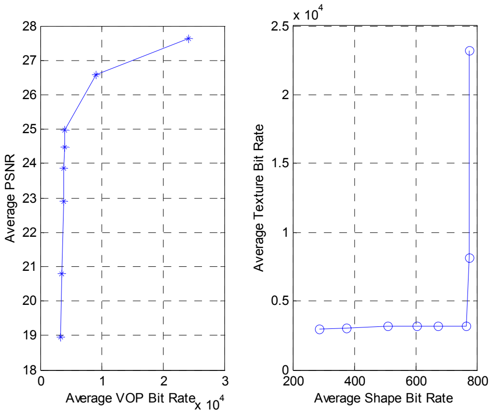


Figure 6 R-D curve and corresponding bit allocation between shape and texture

Figure 6 shows the R-D curve and corresponding bit allocation between shape and texture obtained using our optimal approach for encoding the first *30* frames of the bream sequence. In the figure, the shape bits increase rapidly from *280* to around *800* when the PSNR is between *19* and *25*, while the texture bits stay at around *3,500*. However, when the PSNR keeps increasing to *27.7dB*, the texture bits go up steeply from *3,500* to *24,000* and shape bits only show a slight change. As it can also be inferred from the figure, although the shape information only occupies 0.3-20% of the total VOP bits, it has a strong impact on the quality of the video, which also indirectly proves that the study of optimal bit allocation between shape and texture is very valuable.

6. CONCLUSION

In this paper, we presented an operational rate-distortion optimal bit allocation scheme between shape and texture for the encoding of MPEG-4 object-based video. By applying the Lagrange multiplier method and dynamic programming techniques, our optimal approach has a much more efficient performance than an exhaustive search algorithm. By studying the experimental result on bit allocation between shape and texture information, it is evident that the shape information could have a strong impact on the quality of the reconstructed video sequences.

REFERENCES

[1] MPEG-4 video VM 18.0, ISO/IEC JTC1/SC29/WG11 N3908, Pisa, Jan. 2001.

[2] N. Brady, "MPEG-4 standardized methods for the compression of arbitrarily shaped video objects", *IEEE Trans. Circuits and System for Video Technology*, Vol. 9, NO. 8, pp. 1170-1189, Dec. 1999.

[3] L. P. Kondi, G. Melnikov, and A. K. Katsaggelos, "Joint optimal coding of texture and shape", *Proc. IEEE International Conference on Image Processing*, Volume III, pp. 94-97, Thessaloniki, Greece, October 2001.

[4] A. Vetro, H. Sun, and Y. Wang, "Joint shape and texture rate control for MPEG-4 encoders", *Proc. IEEE International Conference on Circuits and Systems*, pp. 285-288, Montery, USA, June 1998.

[5] A. Vetro, Y. Wang, and H. Sun, "A Probabilistic approach for rate-distortion modeling of multiscale binary shape", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3353-3356, Salt Lake City, May 2002.

[6] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders", *IEEE Trans. Image Processing*, Vol. 3, pp. 533-545, Sept. 1994.

[7] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources", Oper. Res., vol. 11, pp. 399-417, 1963.

[8] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1445-1453, Sept. 1998.

[9] G. M. Schuster and A. K. Katsaggelos, "Fast and efficient mode and quantizer selection in the rate distortion sense for H.263", in *SPIE Proc. Conf. Visual Communications and Image Processing*, pp. 784-795, March 1996.

[10] G. M. Schuster and A. K. Katsaggelos, "A theory for the Optimal Bit Allocation between Displacement Vector Field and Displaced Frame Difference", *IEEE Journal on Selected Areas in Communications*, Vol. 15, NO. 9, pp. 1739-1751, Dec. 1997.