# RESOURCE-DISTORTION OPTIMAL VIDEO CODING AND COMMUNICATIONS

*Haohong Wang*
Qualcomm Inc., San Diego, USA
haohongw@qualcomm.com

*Aggelos Katsaggelos*
Northwestern University, Evanston, USA
aggk@ece.northwestern.edu

## ABSTRACT

In this paper, we describe two major issues in object-based video coding and communications and provide solutions based on the MPEG-4 coding standard. We first consider the general problem of bit allocation among shape, texture and motion in video coding, and provide optimal solutions based on MINMAX (Minimum Maximum) and MINAVE (Minimum Average) distortion criteria, respectively. Then, we discuss the resource allocation problem in video communications, and demonstrate a number of unequal error protection schemes including separated packetization, joint source-channel coding and data hiding. Experimental results demonstrate significant gains by using these algorithms.

## 1. INTRODUCTION

Object-based video is one of the most important topics for interactive multimedia applications such as video telephony and video games. In such applications, a video sequence is organized as a collection of visual objects, which are separately encoded and transmitted. These objects can be thought of a sequence of two-dimensional images with arbitrary shapes. MPEG-4 [1] is the first International multimedia standard that addresses object-based video representation. In MPEG-4, an arbitrarily shaped video object is defined by its shape, texture and motion. Each frame of the video object is called a video object plane (VOP). The texture of a VOP is composed of three color components, a luminance (Y) and two chrominance components (U, V). The shape of a VOP, also called an alpha-plane, is specified by a binary array corresponding to the rectangular bounding box of the VOP specifying whether an input pixel belongs to the VOP or not, or a set of transparency values. We have extensively studied a number of problems in video objects coding and transmission [2-5]. In this paper we address two major issues.

The first issue is how to allocate bit among shape, texture and motion in video compression, which was studied in [2, 5-7]. The difficulty of this optimal bit allocation problem is due to the dependencies between shape and texture of an object, as well as, the dependencies at the macroblock level due to motion compensated predictive coding and differential encoding.

The second issue is how to efficiently allocation communication resources, such as power, bandwidth and cost for object-based video communications. So far there has been very limited reported work [3-4, 8-9] on this topic. One important reason is that arbitrarily shaped video objects make the video processing and transmission much more complicated.

In this paper, we provide optimal solutions or frameworks addressing the above problems. In section 2, we discuss the optimal bit allocation issue, and in section 3 we deal with the resource allocation problem. We draw conclusions in the last section.

## 2. RATE-DISTORTION OPTIMAL VIDEO CODING

The problem at hand is to control both the shape and texture coding parameters to reach the best video quality within a frame bit budget. We therefore formulate the following optimization problem

$$\text{Minimize } D\text{, subject to } R \le R_{budget}, \tag{1}$$

where $R$ is the total bit rate per frame, $D$ is the distortion, and $R_{budget}$ is the available bit budget per frame.

*A. Rate*

Let us denote by $\{m_1, m_2,...,m_N\}$ the $N$ raster-scan ordered macroblocks in the VOP, and by $s_i$ the shape data associated with $m_i$. The set of coding parameters for a macroblock is referred to as the shape (or texture) decision vector. Let us denote by $V=\{V_1,V_2,...,V_N\}$ the set of admissible shape decision vectors for the macroblocks, by $v_i$ a shape decision vector for the $i$th macroblock ($v_i \in V_i$), and $v=\{v_1, v_2, ..., v_N\}$ the set of shape decision vectors for the VOP. Then, $s_i$ is a function of the shape decision vectors $v_i$, that is, $s_i=g_i(v_{i-a},...,v_i)$ where $g_i$ denotes a shape function and $a$ the number of past macroblocks, macroblock $m_i$ depends on. Typically $a$>1 because the encoding of a BAB (binary alpha block) depends on its neighbors to the left, above, and above-left. Similarly, let us denote by $W=\{W_1,W_2,...,W_N\}$ the set of admissible texture decision vectors, $w_i$ a texture decision vector for $m_i$ ($w_i \in W_i$), and $w=\{w_1, w_2, ..., w_N\}$ the set of texture decision vectors for the VOP. Let us denote by $R_{S_i}(v_{i-a},...,v_i)$ the shape bit rate, and $R_{T_i}(s_i, w_{i-b},...,w_i)$ the texture bit rate for macroblock $m_i$, where $b$ is the number of previous macroblocks the texture of $m_i$ depends on. The reason that $R_{T_i}(s_i,w_{i-b},...,w_i)$ depends on $s_i$ is that the texture for transparent macroblocks inside the VOP is not coded. Clearly

$$R = R_{syntax} + \sum_{i=1}^{N} R_{S_i}(v_{i-a},...,v_i) + \sum_{i=1}^{N} R_{T_i}(s_i, w_{i-b},...,w_i), \tag{2}$$

where $R_{syntax}$ represents the bits allocated to the data structure syntax of the VOP, and it is a fixed value if we assume each VOP frame is packed in a separate packet.

### *B. Distortion*

The MINAVE and MINMAX distortion criteria have been widely used. The MINAVE criterion is defined by

$$D = \sum_{i=1}^{N} D_i(s_i, w_{i-b}, \ldots, w_i), \tag{3}$$

where $D_i(s_i, w_{i-b}, \ldots, w_i)$ is the distortion for macroblock $m_i$. The MINMAX criterion is defined by

$$D = \max_{i \in [1,2,\ldots,N]} D_i(s_i, w_{i-b}, \ldots, w_i) \cdot \tag{4}$$

In this work, we compare both distortion criteria. We use the mean-squared error (MSE) as the distortion metric at the macroblock level, that is

$$D_i(s_i, w_{i-b}, \ldots, w_i) = \sum_{x=0}^{15}\sum_{y=0}^{15} d_{i,Y}(x,y)^2 + \sum_{x=0}^{7}\sum_{y=0}^{7}[d_{i,U}(x,y)^2 + d_{i,V}(x,y)^2], \tag{5}$$

where $d_{i,Y}(x,y)$, $d_{i,U}(x,y)$, and $d_{i,V}(x,y)$ are the differences in intensity values for the $Y$, $U$ and $V$ components at pixel $(x,y)$ of macroblock $m_i$. As an example, let us denote at pixel $(x,y)$ by $A$ and $\tilde{A}$ the original and reconstructed alpha values, by $Y$ and $\tilde{Y}$ the corresponding luminance values, and by $G$ the corresponding background pixel intensity value, respectively. Then

$$d_{i,Y}(x,y) = \tilde{A}\tilde{Y} + (1-\tilde{A})G - AY, \tag{6}$$

where we assume that the alpha value could be either $0$ (transparent) or $1$ (opaque). It is noted here that in obtaining the $\tilde{Y}$ value intensities for pixel $(x,y)$ of $m_i$, the parameters $w_{i-1}$, $w_{i-2}$ ..., and $w_{i-b}$ were also used, due to the differential encoding of DCT coefficients and motion vectors. So, this distortion metric accounts for errors due to both shape and texture encoding, and it follows the conventional distortion criteria for image/video quality evaluation.

### *C. MINMAX approach*

Using MINMAX criterion, (1) can be rewritten as

$$\text{Minimize} \max_{i \in [0,1,\ldots,N-1]} D_i(s_i, w_{i-b}, \ldots, w_i), \text{ such that}$$

$$\sum_{i=1}^{N} R_{S_i}(v_{i-a}, \ldots, v_i) + \sum_{i=1}^{N} R_{T_i}(s_i, w_{i-b}, \ldots, w_i) \le R_{\max}, \tag{7}$$

where $R_{max}=R_{budget}-R_{syntax}$.

The problem (7) can be solved by first solving another problem, that is

$$\text{Minimize} \sum_{i=1}^{N} R_{S_i}(v_{i-a}, \ldots, v_i) + \sum_{i=1}^{N} R_{T_i}(s_i, w_{i-b}, \ldots, w_i),$$

$$\text{such that} \max_{i \in [1,2,\ldots,N]} D_i(s_i, w_{i-b}, \ldots, w_i) \le D_{\max} \cdot \tag{8}$$

This can be done because the distortion is a non-increasing function of the bit rate for an optimal codec, that is, by increasing the number of available bits the codec's performance will either remain the same or improve. Hence, when $D_{max}$ sweeps from zero to infinity, $R^*(D_{\max})$, the solution to (8) traces out the staircase-like curve. Therefore, a bisection presented in [10] can be used to find the optimal $D^*_{\max}$ that satisfies $R^*(D^*_{\max}) \le R_{\max}$, and therefore solves problem (7). Let us now define the set of admissible decision vectors $U=V\times W$. For each macroblock $m_i$, there is a decision vector $u_i=(v_i, w_i)$ and (8) can be simplified as

$$\text{Minimize}_{u} \sum_{i=1}^{N}[R_i(u_{i-a}, \ldots, u_i)],$$

$$\text{such that} \max_{i \in [1,2,\ldots,N]} D_i(u_{i-a}, \ldots, u_i) \le D_{\max}, \tag{9}$$

where $R_i(u_{i-a}, \ldots, v_i) = R_{S_i}(v_{i-a}, \ldots, v_i) + R_{T_i}(s_i, w_{i-b}, \ldots, w_i)$, $D_i(u_{i-a}, \ldots, u_i) = D_i(s_i, w_{i-b}, \ldots, w_i)$ ,and $u = [u_1, u_2, \ldots, u_N]^{\mathrm{T}}$ (without loss of generality, we assume that $a \ge b$ ).

To implement the algorithm for solving the optimization problem (9), we create a cost function $C_k(u_{k-a}, \ldots, u_k)$, which represents the minimum total rate and distortion up to and including macroblock $m_k$ with the distortion constraint in (9), given that $u_{k-a}, \ldots, u_k$ are decision vectors for macroblocks $m_{k-a}, \ldots, m_k$. Clearly, the solution for $\min_{u_{N-a}, \ldots, u_N} C_N(u_{N-a}, \ldots, u_N)$ is also the optimal solution for (9). The key observation for deriving an efficient algorithm is the fact that given $a+1$ decision vectors $u_{k-a-1}, \ldots, u_{k-1}$ for macroblocks $m_{k-a-1}, \ldots, m_{k-1}$, and the cost function $C_{k-1}(u_{k-a-1}, \ldots, u_{k-1})$, the selection of the next decision vector $u_k$ is independent of the selection of the previous decision vectors $u_1$, $u_2$, ..., $u_{k-a-2}$. This is true since the cost function can be expressed recursively as

$$C_k(u_{k-a}, \ldots, u_k) = \min_{u_{k-a-1}, \ldots, u_{k-1}} [C_{k-1}(u_{k-a-1}, \ldots, u_{k-1}) + p_k(u_{k-a}, \ldots, u_k)], \tag{10}$$

where

$$p_k(u_{k-a}, \ldots, u_k) = \begin{cases} \infty & \text{if } D_k(u_{k-a}, \ldots, u_k) > D_{\max} \\ R_k(u_{k-a}, \ldots, u_k) & \text{if } D_k(u_{k-a}, \ldots, u_k) \le D_{\max} \end{cases} \tag{11}$$

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 dynamic programming. With the cost function defined, this problem can be converted into a graph theory problem of finding the shortest path in a directed acyclic graph (DAG) [10]. The computational complexity of the algorithm is O($N\times|U|^{max(a,b)+1}$) ($|U|$ is the cardinality of $U$), which depends directly on the value of $a$ and $b$, but still much more efficient than the exponential computational complexity of an exhaustive search algorithm.

### *D. MINAVE approach*

Using MINAVE criterion, (1) can be rewritten as

$$\text{Minimize } \sum_{i=1}^{N} D_i(s_i, w_{i-b},\ldots,w_i), \text{ such that}$$

$$\sum_{i=1}^{N} R_{S_i}(v_{i-a},\ldots,v_i) + \sum_{i=1}^{N} R_{T_i}(s_i, w_{i-b},\ldots,w_i) \le R_{\max}, \qquad (12)$$

and further be simplified based on the similar steps in section *2.C* into

$$\text{Minimize}_{u} \sum_{i=1}^{N} D_i(u_{i-a},\ldots,u_i),$$

$$\text{such that } \sum_{i=1}^{N} [R_i(u_{i-a},\ldots,u_i)] \le R_{\max}. \qquad (13)$$

Based on Lagrangian relaxation problem (13) can be converted into a unconstrained problem that

$$\text{Minimize}_{u} \sum_{i=1}^{N} [R_i(u_{i-a},\ldots,u_i) + \lambda D_i(u_{i-a},\ldots,u_i)], \qquad (14)$$

where $\lambda$ is the Lagrange multiplier. By creating a cost function similarly as described in section *2.C*, we can convert (14) into a graph theory problem and solve it using dynamic programming.

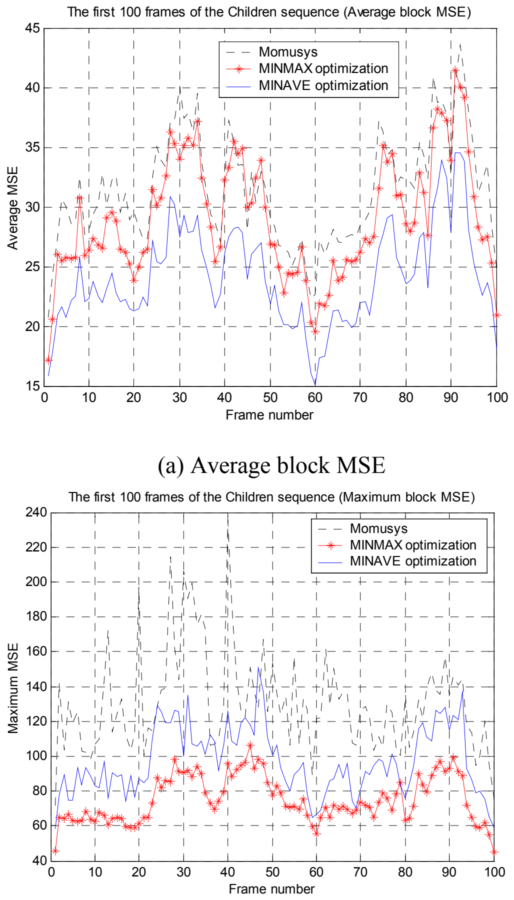
### *E. Experimental results*

We implemented our codec using the proposed encoding schemes and the basic coding algorithms specified in MPEG-4 video verification model, and compared the MINMAX and MINAVE codecs with an MPEG-4 software implementation, MoMuSys [1], in the case where their VOP bit rates are matched. We conducted experiments on the 100 frames of the “Children” sequence in Inter mode, and we run MoMuSys with initial Alpha_TH=0 and initial QP=10, and target bit rate for VOL (Video Object Layer) equal to *40,000*. In this experiment $R_{budget}$, the maximum VOP bit rate from (1), is set to be equal to the VOP bit rate obtained by MoMuSys. Clearly, $R_{budget}$ changes from frame to frame following the profile generated by the MoMuSys rate controller, and the proposed coder will minimize the resulting frame distortion for the given VOP bit rate. As expected and as shown in Fig. 1, the MINAVE approach outperformed the other methods in terms of the average block MSE, and the MINMAX obtained the best maximum block MSE. As expected, the MoMuSys codec was outperformed by both the MINAVE and MINMAX approaches, which points to the benefits of using optimal bit allocation schemes.

To better understand and compare the reconstructed video quality by the MINMAX and MINAVE schemes, we employ the peak signal-to-noise ratio (PSNR), as a common distortion metric to evaluate the reconstructed video quality. The PSNR is obtained by

$$D_{PSNR} = 10\log_{10} \frac{1.5 \times 16^2 \times N \times 255^2}{D}, \qquad (15)$$

where $D$ is defined in (3) and the factor 1.5 comes from the down sampling of the chrominance components by a factor of 2.

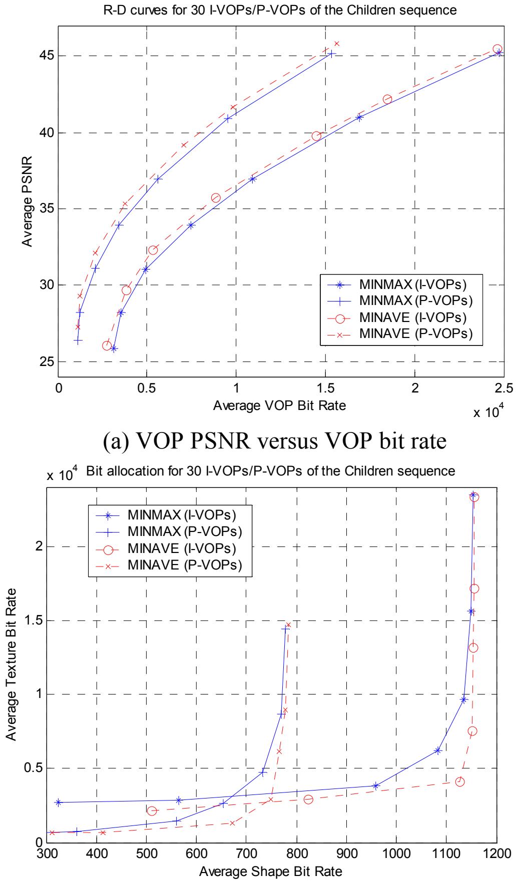


(a) Average block MSE

(b) Maximum block MSE

Figure 1. Comparison of the various codecs

We study the rate-distortion characteristics of the encoding schemes and explore the optimal bit allocation schemes between shape and texture by encoding the first 30 frames of the “children” sequence in both intra and inter mode. In Fig. 2 (a), the rate-distortion curves obtained by the proposed optimal MINMAX and MINAVE approaches are shown, and the corresponding average shape and texture bit rates are shown in Fig. 2 (b) (notice that there is a one-to-one correspondence between operating points). It is clear from Fig. 2 (a) that Inter-mode encoding saves a considerable number of bits compared to Intra-mode encoding for both schemes. Furthermore, in both intra and inter modes, the MINMAX approach performed almost as good as the MINAVE approach. We can find from Fig. 2 (b) that shape information only represents a small portion (in most cases, less than *20%*) of the overall bit budget. Moreover, in both modes, the MINAVE approach is more willing to spend bits on shape than texture compared to the MINMAX approach (because the MINMAX curves are always above the MINAVE curves). This means that the shape has a stronger impact on the total frame distortion than texture and therefore the lossless encoding of shape is the first concern of the encoder as the bit rate increases.



(a) VOP PSNR versus VOP bit rate

(b) Shape bit rate versus texture bit rate

Figure 2. Experimental results for "Children" sequence

## 3. RESOURCE-DISTORTION OPTIMAL VIDEO COMMUNICATIONS

The general problem for resource allocation is to choose coding parameters for the shape and texture of a VOP (or frame) and transmission parameters, so as to minimize the total expected distortion, given a cost constraint and a transmission delay constraint, that is,

$$\text{Minimize } E[D_{tot}], \text{ s.t. } C_{tot} \le C_{\max} \text{ and } T_{tot} \le T_{\max}, \qquad (16)$$

where $E[D_{tot}]$ is the expected total distortion for the frame, $C_{tot}$ is the total cost, $T_{tot}$ is the total transmission delay, $C_{max}$ is the maximum allowable cost, and $T_{max}$ is the maximum amount of time that can be used to transmit the entire frame. In a wireless network, the cost we consider is the transmission powers for the data packets, while in a DiffServ network, the cost represents the price for each QoS channel over which the data packets are transmitted. The expected distortion can be estimated by certain statistical approaches [11] based on the decoder error concealment strategy and the specific channel models.

The observation from previous section that the shape may have a stronger impact on the reconstructed video quality than texture directly motivates the unequal protection of the shape and texture components of the video objects in video encoding and transmission. Unequal protection implies, (1) the important data get higher priority in allocating bits during source coding, (2) the important data is encoded with more redundancy during channel coding, (3) the important data get heavier error protection during transmission. Therefore, a cost-distortion optimal source-channel scheme that jointly considers source coding, channel coding, data hiding and error concealment within the optimization framework will solve the problem (16). The detail framework and solutions will be presented in [3,4]. Here we provide a brief summary of the basic ideas.

Other than allocating more bits for shape information in source and channel coding, data hiding is an effective way to favor error-resilient shape data transmission over unreliable network. In MPEG-4 a data partitioned packetization scheme is applied to increase error resilience. In this scheme, the shape and texture data are packed in a same packet but separated by a motion marker. This way, when the partition containing texture data is corrupted but shape and motion data are received correctly, the motion vector obtained can be used to conceal the corrupted texture. However, since the decoding of the texture partition relies on information stored in the shape partition, such as the texture motion vector and coding modes, the whole packet will be discarded when the shape data is corrupted, even if the texture data is uncorrupted. We propose to embed shape and motion information into the texture data to make it self-decodable. Thus the texture data can be used even if the shape partition is corrupted. In addition, the embedded shape and motion data could help to partially recover the lost shape and motion paritition.

Simulation results of encoding "Children" sequence and transmitting them over the simulated wireless channels with SNR=6dB are shown in Fig. 3, which demonstrated the advantage of joint source-channel coding with data hiding. We compare three approaches, (1) EEP method, where the shape and texture partition within a packet are equally protected (by channel coding) but different packet is allowed to be unequally protected; (2) UEP method, where the shape and texture partitions are unequally protected by channel coding; (3) Hybrid method, where the shape and texture partition are unequally protected by channel coding, and the shape and motion data are allowed to be embedded into the texture data. The data hiding is not used in EEP and UEP methods, but source coding parameters are optimized for all these methods. UEP method outperformed EEP at lower bit rate (when rate is lower than *6000*) because it can use more channel bits to protect shape data, which has a stronger impact on the decoded video quality. The data hiding works well when bit rate is high enough that there is enough DCT coefficients available for embedding shape and motion information. This way, the hybrid method inherits the virtue from both UEP and data hiding, and thus makes it work well for all ranges of the bit rate.

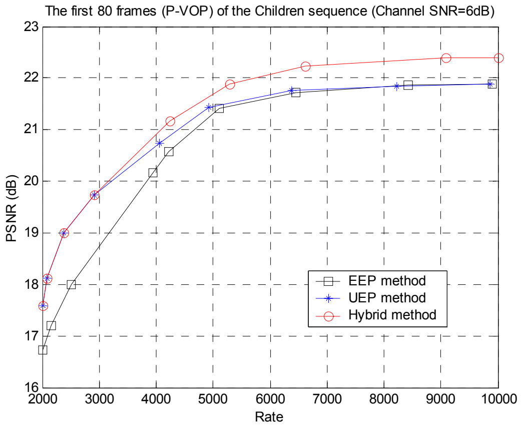


Figure 3. Comparison of three approaches

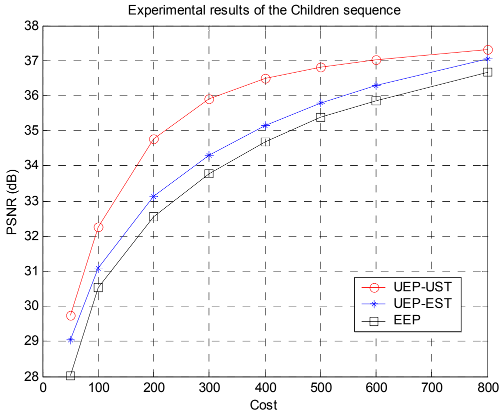


Figure 4. Cost-distortion curves of various systems

Another approach to enable unequal error protection is to use a separated packetization scheme, where the shape and texture are packed into separate packets. Within the framework of cost-distortion optimization, we jointly consider source coding, packet classification and error concealment. The optimal solution will dynamically allocate the resources based on the importance of the packets. In Fig. 4 we compared three error protection schemes: (1) UEP-UST, an unequal error protection scheme using the separated packetization scheme, where the shape and texture data are placed in separate packets and therefore can be transmitted over different service channels; (2) UEP-EST, an unequal error protection scheme using combined packetization where the packets containing both shape and texture data can be transmitted over different service channels; (3) EEP, an equal error protection scheme using combined packetization and all the packets are transmitted over a same channel. By jointly adapting the source coding parameters along with the transmission parameters, the UEP approaches outperformed the EEP approach. In addition, UEP-UST outperformed UEP-EST because the former approach has increased flexibility in adjusting the parameters.

## 4. CONCLUSIONS

In this paper, we demonstrated a number of advanced algorithms for two major issues of object-based video coding and communications: bit allocation problem and resource allocation problem. We studied and compared the optimal MINMAX and MINAVE approaches in video coding. In addition, we overviewed two unequal error protection schemes for video communications. The experimental results demonstrate that the proposed algorithms have significantly outperformed other approaches.

## REFERENCES

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

[2] H. Wang, G. M. Schuster, and A. K. Katsaggelos, “Rate-distortion optimal bit allocation for object-based video coding”, to appear, *IEEE Trans. Circuits and System for Video Technology*, 2005.

[3] H. Wang, F. Zhai, Y. Eisenberg, and A. K. Katsaggelos, “Cost-distortion optimized unequal error protection for object-based video communications”, to appear, *IEEE Trans. Circuits and System for Video Technology*, 2005.

[4] H. Wang, S. A. Tsaftaris, and A. K. Katsaggelos, “Joint source-channel coding for wireless object-based video communications utilizing data hiding”, *IEEE Trans. Image Processing*, submitted.

[5] H. Wang and A. K. Katsaggelos, “Optimal object-based video coding: a MINMAX approach”, *IEEE Trans. Consumer Electronics*, submitted.

[6] A. Vetro, H. Sun, and Y. Wang, “MPEG-4 rate control for multiple video objects”, *IEEE Trans. Circuits and System for Video Technology*, Vol. 9, NO. 1, pp. 186-199, Feb. 1999.

[7] L. P. Kondi, F. W. Meier, G. M. Schuster, and A. K. Katsaggelos, "Joint optimal object shape estimation and encoding”, in *Proc. SPIE Conf. on Visual Communications and Image Processing*, pp. 14-25, SPIE vol. 3309, San Jose, CA, Jan. 28-30, 1998.

[8] S. Worrall, S. Fabri, A. H. Sadka, and A. M. Kondoz, “Prioritization of data partitioned MPEG-4 video over mobile networks”, *European Transactions on Telecommunications, Special Issue on Packet Video*, Vol. 12, Issue No. 3, pp. 169-174, May/June 2001.

[9] W. R. Heinzelman, M. Budagavi, and R. Talluri, “Unequal error protection of MPEG-4 compressed video”, in *Proc Proceedings of the International Conference on Image Processing*, pp. 530-534, Oct. 1999.

[10] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion based video compression: optimal video frame compression and object boundary encoding*, Kluwer Academic Publishers, 1997.

[11] R. Zhang, S. L. Regunathan, and K. Rose, “Video coding with optimal inter/intra-mode switching for packet loss resilience”, *IEEE J. SAC*, Vol 18, No. 6, pp. 966-976, June 2000.