Joint Optimal Object Shape Estimation and Encoding

Lisimachos P. Kondi, Gerry Melnikov and Aggelos K. Katsaggelos

Abstract

A major problem in object oriented video coding and MPEG-4 is the encoding of object boundaries. Traditionally this problem is treated separately from the texture encoding problem. In this paper, we present a vertex-based shape coding method which is optimal in the operational rate-distortion (ORD) sense and takes into account the texture information of the video frames. This is accomplished by utilizing a variable-width tolerance band whose width is a function of the texture profile. As an example, this width is inversely proportional to the magnitude of the image gradient. Thus, in areas where the confidence in the estimation of the boundary is low and/or coding errors in the boundary will not affect the application (object oriented coding, MPEG-4, etc.) significantly, a larger boundary approximation error is allowed. We present experimental results which demonstrate the effectiveness of the proposed algorithm.

Index Terms

Shape coding, boundary coding, MPEG-4, operational rate-distortion theory, shape-adaptive DCT.

I. Introduction

Shape representation and encoding is a problem with relatively long history (for a recent review, see [3] and references therein).

In our previous work [4], [5], [6], [7], [8], we introduced operational rate-distortion optimal and efficient shape coding schemes, in the intra- and inter-mode, utilizing any order curves for the approximation, such as, straight lines and B-splines. These schemes utilize Graph Theory

A preliminary version of this paper was presented in [1], [2].

The work of Aggelos K. Katsaggelos was supported in part by the Motorola Center for Communications.

and Dynamic Programming in order to reduce their computational complexity and typically outperform shape coding techniques proposed by MPEG-4 [3].

A variety of distortion measures can be utilized with these operational rate-distortion optimal shape coding techniques. If errors at all parts of the boundary are given the same importance, small but distinct features of the original boundary may disappear (be "cut off") in its approximation, for low bit rates. Therefore, important features of the shape should be preserved. Coupled with this observation is the fact that the number of bits utilized for the encoding of the texture inside the approximated shape, also depends on the way the shape is approximated. Traditionally the shape and texture information are encoded independently from each other, as reported in the results in the literature but also as represented by MPEG-4.

In this paper we address both of these issues, namely, the adaptive rate-distortion optimal encoding of shape and the encoding of the associated texture.

In introducing the adaptivity in shape or boundary encoding [1], a quantitative measure of the relative importance of each part of the boundary needs to be defined first. The objective of this paper is not to focus on ways for determining such relative importance of parts or features of the boundary, but to propose a way for incorporating such information into the ORD optimal boundary encoding process. It is emphasized that the relative importance of the boundary segments is application dependent, and therefore it should be proper to be addressed separately by each application.

The value of the curvature at each pixel on the curve can be used, for example, in determining the relative significance of each boundary segment. In this paper, we utilize an adaptive distortion measure which is based on a measure of the sharpness of the underlying intensity edge, as expressed, for example, by the magnitude of the gradient of the intensity. Thus, in areas where the magnitude of the gradient is high, a closer approximation of the boundary is forced, whereas in areas with low gradient magnitude, a higher approximation error is allowed. This idea was first presented in [1] and was later re-discovered in [9]. The justification for this is that in areas of low gradient magnitude, a higher approximation error would be less perceivable. Furthermore, if a gradient-based boundary estimation method was employed in the first place, our confidence

[1]The terms shape and boundary are used interchangeably in this paper assuming one can uniquely determine the one from the other.

in the accuracy of the boundary estimation would not be very high in areas of low gradient magnitude. This is also true in the case of object oriented video coding. The motion and object estimation cannot be very accurate in areas with low gradient magnitude and furthermore, larger boundary approximation errors in these areas would not impact motion compensation much. A second reason for using the magnitude of the intensity gradient to determine the amount of error in approximating the corresponding boundary, is that this will have an effect on the quality of the encoded texture, if an MPEG-4 type of approach is utilized in encoding the texture, as outlined next.

The MPEG-4 standard allows for the encoding of the texture of video objects using *Shape-Adaptive Discrete Cosine Transform (SA-DCT)* [10]. With a block-based hybrid motion compensated video coding approach, video frames are encoded by taking the Discrete Cosine Transform of 8×8 blocks of the intensity or the displaced frame difference. When shape information is used, however, there are 8×8 blocks which are partially occupied by an object. Using an 8×8 DCT for such blocks, we would need to transmit 64 coefficients, although the actual number of pels belonging to the object would be smaller. SA-DCT provides for a way of encoding such blocks using a number of coefficients that is equal to the number of the object pels in the block. This is accomplished by shifting the object pels towards the origin of the block and then taking one dimensional DCTs row-wise and then column-wise. The length of these one-dimensional DCTs can be of any size less than or equal to eight.

Compression is accomplished by quantization of the DCT coefficient followed by entropy coding. It is expected that if a block contains an edge or an area with high gradient magnitude, more bits would be required for its encoding. Thus, in order to achieve a more efficient encoding of the texture, it would be beneficial to have an accurate boundary approximation in areas with high gradient magnitude. If an object and its background are encoded using SA-DCT, in areas close to the object boundary, the same block can belong to both the object and its background. If the boundary is accurately approximated, SA-DCT can be used to encode the block twice, once for the object and once for the background and the edge will not impact the encoding efficiency too much. Furthermore, it is expected that SA-DCT will be more efficient if the boundary edges are horizontal or vertical, as opposed to diagonal. Therefore, our boundary encoding algorithm is allowed to be biased towards choosing horizontal or vertical edges, through the use of the appropriately designed variable length codes (VLCs).

The rest of the paper is organized as follows. In section II, an introduction of optimal shape coding is presented along with a way to incorporate our confidence in the boundary estimation into the shape coding problem formulation. In section III, the shape coding algorithm is explained. In section IV, experimental results are presented and in section V, conclusions are drawn.

II. Proposed Algorithm

We consider the problem of the optimal approximation of a discrete connected boundary in the rate distortion sense. That is, the encoding of the boundary is sought with the smallest possible bit rate at an acceptable distortion or with the smallest possible distortion satisfying a bit budget constraint. We have developed a number of approaches for solving these two problems for both the intra- and inter-frame coding modes and under various order curves and distortion criteria [4], [5], [7], [6]. In all cases the problem reduces to finding the shortest path in a Directed Acyclic Graph (DAG).

Let $B = \{b_0, \ldots, b_{N_B-1}\}$ denote the connected boundary which is an ordered set, where b_j is the j-th point of B and N_B is the total number of points in B. This boundary needs to be approximated by a curve of order n (for $n = 1$ a polygon results). That is, the number and the location of the control points (vertices for a polygon) of the curve need to be determined. Let $A = \{a_1, a_2, a_3, \ldots\}$ be the set of admissible vertices or control points. A needs to be defined before the coding of the boundary begins. Usually, $A \supseteq B$, i.e., A is a superset of B and all original boundary points are eligible to become control points. In our previous work [4], [5], [6], [7], we defined a "distortion band" of width $2 \cdot D_{max}$ along the boundary B. The boundary approximation must lie within the distortion band.

In this paper, we allow the distortion band to have variable width along the boundary. We call this new band a *tolerance* band. The definition of the tolerance band requires a D_{max} for every boundary point. We denote this as $D_{max}[i], i = 0, \ldots, N_B - 1$. In order to construct the tolerance band, we draw circles from each boundary point b_i with radius $D_{max}[i]$. The tolerance band consists of the set of all point that lie inside the circles.

As mentioned earlier, a number of approaches can be followed in defining $D_{max}[i]$, depending on the objectives of the application. In this work, $D_{max}[i]$ is defined in a way that is inversely proportional to the image gradient. The algorithm proceeds as follows: The gradient is first

calculated for the whole image; that is, for an image $f(x, y)$ is defined as:

$$\nabla f(x, y) = [\partial f/\partial x \ \partial f/\partial y]^T = [f_x \ f_y]^T. \tag{1}$$

Out of a number of possible implementations, the Sobel edge detector masks are used to calculate the gradient [11], that is, estimates $\hat{f}_x$ and $\hat{f}_y$ for f_x and f_y are obtained by using gradient operators of the form:

$$\hat{f}_x = \mathbf{w}_1^T \mathbf{x}, \tag{2}$$

$$\hat{f}_y = \mathbf{w}_2^T \mathbf{x}, \tag{3}$$

where $\mathbf{x}$ is the vector containing image pels in a local image neighborhood and $\mathbf{w}_1$ and $\mathbf{w}_2$ are the Sobel edge detector masks, where,

$$\mathbf{w}_1 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \tag{4}$$

and

$$\mathbf{w}_2 = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \tag{5}$$

The magnitude of the gradient is then computed, that is,

$$|\nabla f(x, y)| = \sqrt{f_x^2(x, y) + f_y^2(x, y)}. \tag{6}$$

Let us now denote by $gradmin$ and $gradmax$, respectively, the minimum and maximum of the magnitude of the image gradient for the whole image. Let us also denote the desired minimum and maximum values of $D_{max}[i]$ as T_{min} and T_{max}, respectively. Then, a linear mapping is performed between the gradient value of each boundary point and the width of the tolerance band. If the magnitude of the gradient at the boundary point b_i is grad[i], then the width of the tolerance band at this point is given by:

$$D_{max}[i] = T_{min} + \lambda(grad[i] - gradmax), \tag{7}$$

where

$$\lambda = \frac{T_{max} - T_{min}}{gradmin - gradmax}. \tag{8}$$

In practice, we need to define a threshold for the gradient magnitude. The boundary points whose gradient magnitude exceeds the threshold should have the minimum possible $D_{max}[i]$. Clearly, $gradmax$ is equal to the threshold in that case. An example of the application of this approach is shown in Fig. 1. The variable width distortion band is shown in it, with $T_{max} = 3.0$ and $T_{min} = 0.8$.

Having established notation and introduced the tolerance band we now consider one specific formulation of the shape encoding problem. We assume B-splines are used for the boundary approximation (the k-th spline segment is denoted by $Q_k(p_{k-1}, p_k, p_{k+1})$ and it is defined in terms of three control points p_{k-1}, p_k, p_{k+1}), and the following optimization problem is considered

$$\min_{p_0, p_1, \ldots, p_{N_p}} R(p_0, \ldots, p_{N_p+1}) \tag{9}$$

subject to

$$D(p_0, \ldots, p_{N_p+1}) \leq D_{max},$$

where $R(.)$ is the bit rate required to encode the control points, $D(.)$ is the distortion measure and D_{max} is the maximum distortion permitted. It is noted here that the optimization is with respect to both the number and the location of the control points and also that the first and last control points are the same $(p_0 = p_{N_p+1})$ and they are considered known.

For the case of B-splines, the total distortion can be expressed as follows.

$$D(p_0, \ldots, p_{N_P+1}) = \max_{k \in [1, \ldots, N_P]} d(p_{k-1}, p_k, p_{k+1}), \tag{10}$$

where $d(p_{k-1}, p_k, p_{k+1})$ is the segment distortion and can be expressed as

$$d(p_{k-1}, p_k, p_{k+1}) = \begin{cases} 0: & \text{all points of } Q_k(p_{k-1}, p_k, p_{k+1}) \\ & \text{are inside the tolerance band} \\ \infty: & \text{any point of } Q_k(p_{k-1}, p_k, p_{k+1}) \\ & \text{is outside the tolerance band} \end{cases} \tag{11}$$

This distortion measure takes a curve segment Q_k, given by the three control points p_{k-1}, p_k and p_{k+1}, as input and checks if the curve segment is inside the tolerance band.

The goal of the proposed algorithm is to find the B-spline curve whose control points can be encoded with the smallest number of bits under two conditions: 1) the approximated curve

lies within the tolerance band, and 2) the control points must be selected from the admissible control point set A.

The distortion of curve segment Q_k depends on three control points p_{k-1}, p_k and p_{k+1}. The segment distortion can be combined with the segment rate by defining a weight function w as follows,

$$w(p_{k-1}, p_k, p_{k+1}) = r(p_{k-1}, p_k, p_{k+1}) + d(p_{k-1}, p_k, p_{k+1}). \tag{12}$$

Various ways to differentially encode the location of the control points and therefore define the rate segment are considered in [4], [7], [3].

The problem can be formulated as a shortest path problem in a weighted directed graph. A vector $\vec{E}$ starts at control point $p_u = a_{i,i_b}$ and ends at control point $p_{u+1} = a_{k,k_b}$ with the condition that both admissible control points cannot be assigned to the same boundary point $(\vec{E} = a_{i,i_b} - a_{k,k_b} \in A^2; \forall i \neq k\}$. A path of order K from control point p_0 to control point p_K is an ordered set $\{p_0, \ldots, p_K\}$. The length of the path is defined as follows,

$$\sum_{k=1}^{K-1} w(p_{k-1}, p_k, p_{k+1}). \tag{13}$$

Again, note that the above definition of the weight function leads to a length of infinity for every path which includes a curve segment which has a part that lies outside the tolerance band. Therefore a shortest path algorithm will not select these paths. A specific example of a Directed Acyclic Graph for the B-spline approximation case is shown in Fig. 2. The DAG-shortest-path algorithm [12] can be used to efficiently find the shortest path of the graph.

If a polygon is used instead of B-splines to approximate the boundary, the formulation is similar with the following exceptions. A polygon edge is defined by two points, its vertices. Thus, the control point (vertex) rates and segment distortions depend on only two points ($r(p_{k-1}, p_k)$ and $d(p_{k-1}, p_k)$. Therefore, the weights $w(p_{k-1}, p_k)$ also depend on two control points (vertices), p_{k-1} and p_k.

As mentioned previously, we expect the Shape Adaptive DCT (SA-DCT) to be more efficient if the edges of the object are horizontal or vertical. Thus, we allow for a $bias < 1$ multiplicative factor for the weights of points p_{k-1} and p_k which correspond to horizontal or vertical edges. Thus,

$$w(p_{k-1}, p_k) = bias \cdot [r(p_{k-1}, p_k) + d(p_{k-1}, p_k)] \tag{14}$$

if p_{k-1} and p_k define a horizontal or vertical edge. Thus, the boundary encoding algorithm will favor horizontal and vertical edges.

III. Experimental Results

We coded both the intensity and shape of frame 0 of the "Kids" and "Bream" sequences. The intensity and shape of frame 0 of the "Kids" sequence are shown in Fig. 3 and Fig. 4, respectively. A number of experiments were conducted, some of which are reported below.

In one experiment, B-splines were used for the boundary approximation. The intensity of the objects was coded using SA-DCT with a Quantization Parameter (QP) equal to 20. A fixed-width tolerance band of one and three pels, was used, as well as, a variable-width tolerance band which is inversely related to the gradient magnitude, as discussed in this paper ($T_{max} = 3$, $T_{min} = 0.8$), were used in encoding the shape information. In Table I the number of bits required for shape and intensity encoding are shown, along with the corresponding PSNR in dB. The PSNR was calculated with respect to the intersection of the original and reconstructed shape boundaries. Figure 5 shows the boundary approximation using a variable tolerance band as discussed in this paper, for which 345 bits were used. Figure 6 shows a result of a fixed width tolerance band width of 1 pel, for which 467 bits (35.36% more) were used. By comparing the two encodings, it is clear that the important features of the two objects have been preserved in Fig. 5, while resulting in considerable savings of bits.

The above experiment was repeated using straight lines instead of B-splines to encode the boundaries. Variable distortion band results are presented with or without the use of the bias that favors horizontal and vertical directions, as described in section II. A bias coefficient of 0.5 is used. The results of this experiment are shown in Table II for a QP of 20 and in Table III for a QP of 10. It can be seen that, in both cases, the total number of bits is significantly reduced when using the variable width distortion band with a bias.

We repeated all of the above experiments for frame 0 of the "Bream" sequence. The results are shown in Tables IV, V and VI. Similar observations can be made.

We also compared the performance of the proposed shape and texture coding algorithm with results obtained using MPEG-4 (Momusys implementation). Context-based Arithmetic Encoding (CAE) was used for shape coding and 8×8 DCT was used for texture coding. The MPEG-4 results are optimal in the operational rate-distortion sense, as described in [13]. Thus, they offer

the best possible rate-distortion performance achievable by MPEG-4. The results can be seen in Fig. 7 for the "Bream" sequence. The figure shows the PSNR as a function of the total bit rate (shape plus texture). The QPs used were equal to 10, 15, 20, 25 and 30. It can be observed that the proposed algorithm typically outperforms the rate-distortion optimized MPEG-4. Furthermore, it should be noted that the MPEG-4 optimization algorithm in [13] optimizes the PSNR for a given bit budget, without any explicit restrictions on shape distortion, whereas the proposed algorithm guarantees that the shape approximation will lie within the variable-width tolerance band.

IV. Conclusions

We have presented a framework for the joint encoding of shape and texture information. In order to accomplish that, a shape coding algorithm with an adaptive distortion measure was developed. The distortion measure should describe the relative importance of each part of the boundary and can depend on the corresponding texture information. In this paper, the width of the distortion band was defined to be inversely proportional to the texture gradient at each boundary pel. A number of other ways for doing this can be envisioned, such as the use of the curvature. We have shown that the variable-width distortion band is able to encode the shape using a lower number of bits while preserving its important features. In addition, if the object texture is encoded using Shape-Adaptive DCT, considerable savings in the total number of bits used for shape and texture are observed.

V. Acknowledgement

The authors would like to thank Haohong Wang of Northwestern University for providing comparison results using MPEG-4.

References

[1] L. P. Kondi, F. W. Meier, G. M. Schuster, and A. K. Katsaggelos, "Joint optimal object shape estimation and encoding," in *Proceedings SPIE Conf. on Visual Comm. and Image Proc.*, pp. 14–25, Jan. 1998.

[2] L. P. Kondi, G. Melnikov, and A. K. Katsaggelos, "Jointly optimal coding of texture and shape," in *Proceedings of the International Conference on Image Processing*, (Thessaloniki, Greece), 2001.

[3] A. K. Katsaggelos, L. P. Kondi, F. W. Meier, J. Ostermann, and G. M. Schuster, "MPEG-4 and rate-distortion-based shape-coding techniques," *Proceedings of the IEEE*, vol. 86, pp. 1126–1154, June 1998.

[4] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression, Optimal Video frame compression and Object boundary encoding.* Kluwer Academic Press, 1997.

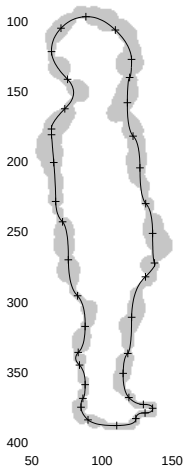

Fig. 1. A shape boundary along with its corresponding variable width tolerance band.

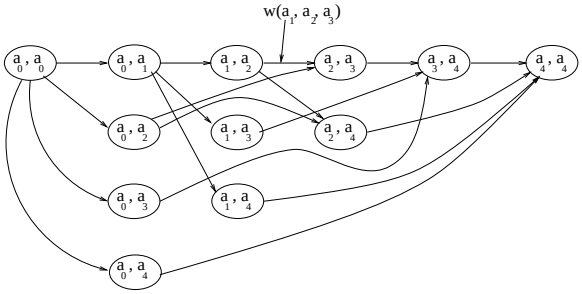

Fig. 2. An example of a Directed Acyclic Graph for the B-spline approximation case. One of these paths is the optimal path.

Fig. 3. Frame 0 of the “Kids” sequence.

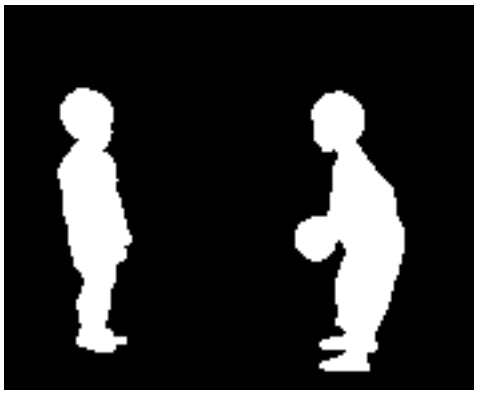

Fig. 4. Segmentation of frame 0 of the “Kids” sequence.

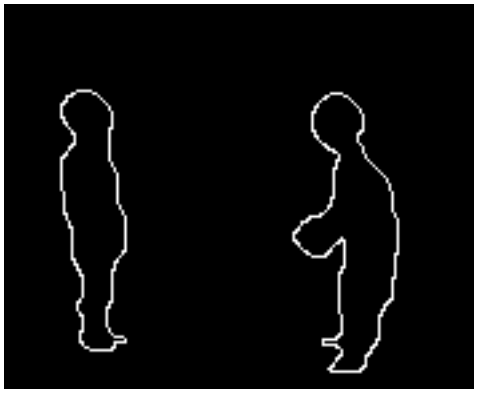

Fig. 5. Result of the variable width tolerance band.

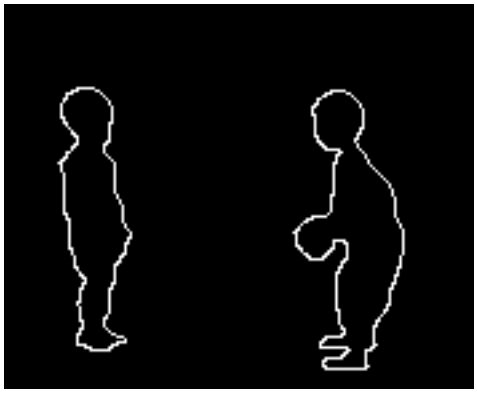

Fig. 6. Result of the fixed width tolerance band algorithm ($D_{max} = 1$).

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	467	4623	5090	29.86
3 pels	323	4428	4751	29.98
Variable	345	4557	4902	29.93

TABLE I

RESULTS OF SHAPE CODING USING B-SPLINES AND A QP OF 20 ("KIDS" FRAME 0).

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	528	4690	5218	30.09
3 pels	304	4502	4806	30.14
Variable + bias of 0.5	384	4364	4748	30.36

TABLE II

RESULTS OF ENCODING USING STRAIGHT LINES AND A QP OF 20 ("KIDS" FRAME 0).

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	528	8227	8755	34.28
3 pels	304	7998	8302	34.34
Variable + bias of 0.5	384	7849	8233	34.46

TABLE III

RESULTS OF ENCODING USING STRAIGHT LINES AND A QP OF 10 ("KIDS" FRAME 0).

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	340	5327	5667	28.96
3 pels	252	5240	5492	29.05
Variable	303	5286	5589	28.99

TABLE IV

RESULTS OF SHAPE CODING USING B-SPLINES AND A QP OF 20 ("BREAM" FRAME 0).

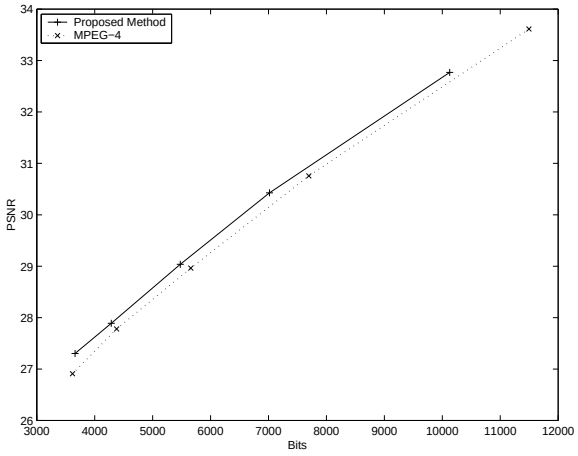

Fig. 7. Comparison between the proposed algorithm and MPEG-4 (frame 0 of "Bream" sequence).

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	343	5286	5629	29.01
3 pels	235	5249	5484	29.02
Variable + bias of 0.5	304	5173	5477	29.04

TABLE V

RESULTS OF ENCODING USING STRAIGHT LINES AND A QP OF 20 ("BREAM" FRAME 0).

[5] G. M. Schuster and A. K. Katsaggelos, "An optimal boundary encoding scheme in the rate distortion sense," *IEEE Transactions on Image Processing*, Jan. 1998.

[6] G. Melnikov, G. M. Schuster, and A. K. Katsaggelos, "Shape coding using temporal correlation and joint VLC optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 744–754, Aug. 2000.

[7] G. M. Schuster, G. Melnikov, and A. K. Katsaggelos, "Operationally optimal vertex-based shape coding," *IEEE Signal Processing Magazine*, vol. 15, pp. 91–108, Nov. 1998.

[8] R. Nygaard, G. Melnikov, and A. K. Katsaggelos, "A rate distortion optimal ECG coding algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 48, pp. 28–40, Jan. 2001.

[9] K. J. Kim, C. W. Lim, M. G. Kang, and K. T. Park, "Adaptive approximation bounds for vertex based contour encoding," *IEEE Transactions on Image Processing*, vol. 8, pp. 1142–1147, Aug. 1999.

[10] T. Sikora and B. Makai, "Shape-adaptive DCT for generic coding of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, pp. 59–62, Feb. 1995.

[11] A. K. Jain, *Fundamentals of digital image processing*. Prentice-Hall, 1989.

[12] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to algorithms*. McGraw-Hill Book Company, 1991.

[13] H. Wang, G. M. Schuster, and A. K. Katsaggelos, "Operational rate-distortion optimal bit allocation between shape and

Tolerance Band	Shape Bits	Intensity Bits	Total Bits	PSNR (dB)
1 pel	343	9968	10311	32.71
3 pels	235	10025	10260	32.76
Variable + bias of 0.5	304	9823	10127	32.77

TABLE VI

RESULTS OF ENCODING USING STRAIGHT LINES AND A QP OF 10 ("BREAM" FRAME 0).

texture for mpeg-4 video coding," in *Proceedings of the International Conference on Multimedia and Expo (ICME)*, (Baltimore, MD), July 2003.