IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing
September 13-15, 1999, Copenhagen, Denmark
Electronic Proceedings

Context based optimal shape coding

Gerry Melnikov, Aggelos K. Katsaggelos and Guido M. Schuster*

Northwestern University
Electrical and Computer Engineering Dept
Evanston, Illinois 60208, USA

Email: gerrym@ece.nwu.edu , aggk@ece.nwu.edu

***3COM**
Advanced Technologies Research Center
Mount Prospect, Illinois 60056, USA

Email: Guido_Schuster@3com.com

Abstract

This paper investigates how context-based DPCM techniques can be used in conjunction with an operationally rate-distortion (ORD) optimized shape coder. Object contours are approximated by connected second-order spline segments, each defined by three consecutive control points, and, in the case of inter mode, by segments of the motion-compensated reference contours. In scalable intra mode and non-scalable inter mode shape coding, consecutive control points are encoded predictively, using an angle and run framework, with respect to corresponding contexts in the reference frame and the base layer, respectively. We employ a novel criterion for selecting global object motion vectors in the inter mode, which further improves efficiency.

Table of Contents

1 Introduction

With emergence of new multimedia applications and the adopted MPEG-4 standard, object-oriented video coding, and hence efficient object shape representation, has attracted considerable attention.

We have previously proposed optimal approximations of a given boundary based on curves of different orders and for various distortion metrics, processing each frame independently (intra mode) and without scalability [2,6].

While there are clearly many cases, particularly in heterogeneous or error-prone environments, when the property of bitstream scalability is desirable from the transmission point of view, it has generally been thought as yet another constraint placed on the encoder, and, hence, not expected to outperform non-scalable techniques in the ORD sense. Generally, the issue is not resolved since there has been no theoretical results linking the rate distortion function to scalability. Section ``Scalable Shape Coding'' deals with the question of how redundancy between layers of a scalable shape coder can be exploited through the use of contexts. Scalability algorithms based on base layer error compensation can be found in [1].

Inter mode shape coding is another challenging application where context-based DPCM can be used to decorrelate similar contours. Methods proposed for MPEG-4 and reviewed in [2] are suboptimal and while achieving good compression in the intra mode, suffer from poor coding efficiency in the inter mode. In [3] the first attempt was made to apply the vertex-based ORD optimal contour coding using contexts to the inter mode, in order to take into account the temporal contour redundancies present in typical video sequences. In this paper this approach is extended by employing a novel criterion for global object-based motion vector selection which fits naturally into the chosen code structure. Furthermore, the encoder adaptively switches between context and tracking modes [3] to even better capitalize on contour similarity.

2 Shape Coding in the Inter Mode

In this paper we solve the problem of encoding temporally correlated contours optimally in the ORD sense within the chosen vertex-based framework. Contours are approximated by connected 2^{nd}-order B-spline segments, each defined by 3 consecutive control points, (p_{u-1},p_u,p_{u+1}). Admissible control point band with a prescribed labeling scheme [2] is introduced to decrease complexity and avoid cycles.

2.1 Temporal Redundancy

Although object boundaries between frames are clearly correlated, efforts to gain coding efficiency based on this apparent redundancy have, so far, met relatively small success [2]. With the existing context-based methods, one global motion vector, minimizing the number of mismatched pixels, is employed per contour to align corresponding objects in two consecutive frames. These algorithms differ from their intra mode counterparts only in the choice of neighboring pixels serving as contexts, i.e., the context is now a spatio-temporal neighborhood. The main disadvantage of this type of approach is its pixel-based nature, which suffers from mis-alignments due to motion and noise, introduced during the frame acquisition and segmentation processes. This makes predictions of whether a given pixel is on or off the boundary highly unreliable.

2.2 Context based DPCM

In this work we reduce the effects of contour noise by utilizing temporal context for the predictive encoding of the (*angle, run*) symbols [3], instead of the underlying pixels. Thus, instead of using temporal context to predict boundary pixel locations, we use them to estimate the current (*angle, run*) symbol, which defines the location of the next consecutive control point. The underlying assumption is that these symbols are affected by noise to a much smaller extent than the original boundary.

Figure 1 depicts a hypothetical context window in the reference frame after motion compensation, which is centered on a pixel, denoted by O in the admissible control point band. It is used to extract both the most likely direction and the most likely length of the vector pointing from that pixel to the next potential control point. That is, if an actual control point is located at the current position, this context provides an estimate of the location where the next control point is most likely to be. The context for the angle component is obtained by selecting the direction in which most of the transitions between consecutive boundary pixels in the reference frame occur. This corresponds to the North-West (NW) direction in Fig. 1, which is pointed to 6 times. This estimate of the direction, being a statistical average, is very robust to contour noise. Similarly, the run length component of the context is obtained by selecting the most frequently occurring distance between consecutive control points in the context window. In this example, a run length of 4, occurring 2 times, is selected.

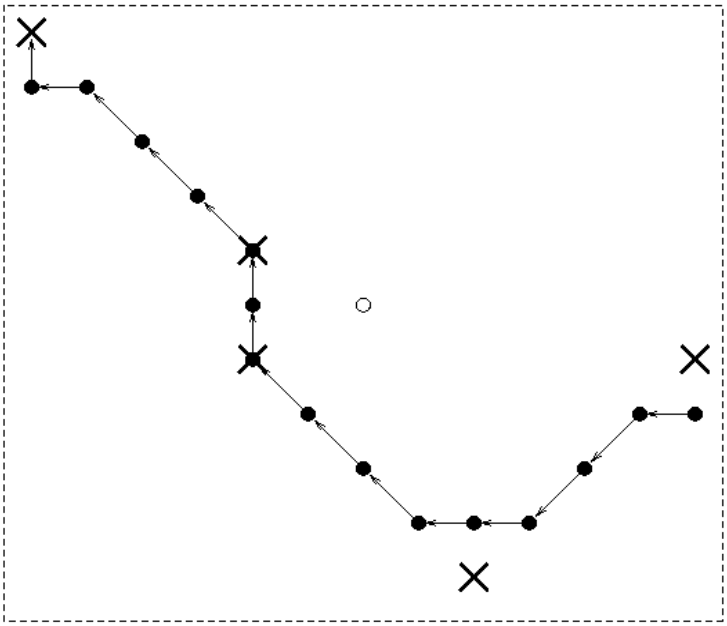

Figure 1: Control points (X marks) and boundary pixels (circles) in a temporal context window. Context: NW direction, run of 4.

Having computed contexts as described above, we employ a spatially adaptive VLC scheme that assigns shorter codewords to combined (angle, run) symbols that are close to ($angle_{context}$, $run_{context}$). Figure 2 shows a toy example of this idea for the *angle* component. With vector lengths proportional to the probabilities of the corresponding directions, Fig. 2A applies to the case when the direction context is known (NW), and Fig. 2B to the case there are too few reference pixels (based on a predetermined threshold) for its meaningful computation, in which case the algorithm reverts to the intra mode ([3]).

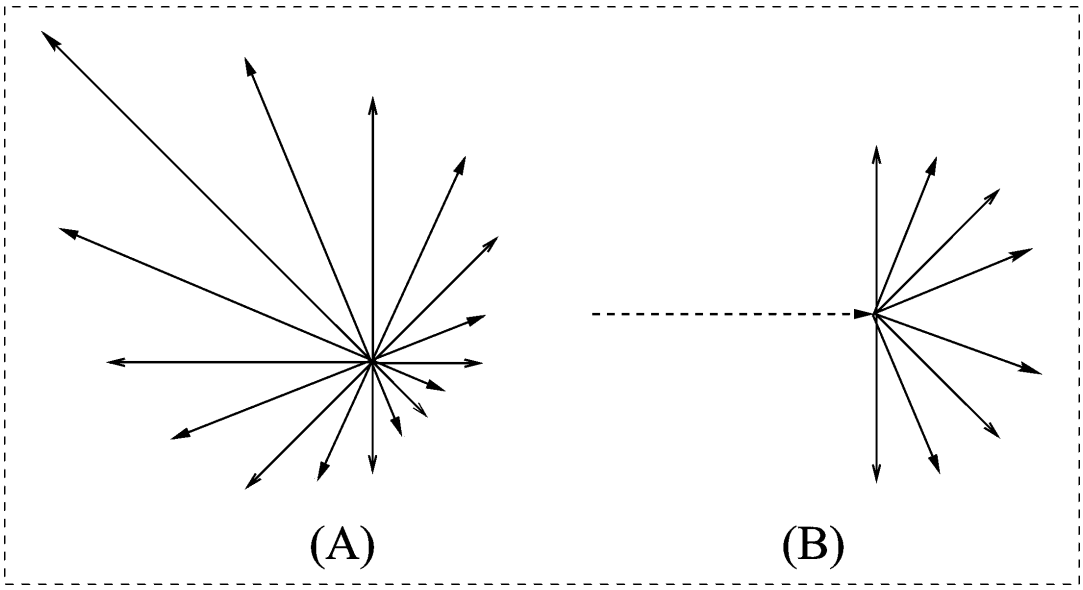

Figure 2: Typical probability assignment for the direction with NW context (A), with no context (B).

With non-homogeneous and non-rigid motion it is often the case that certain contour segments are well approximated by the motion-compensated reference frame, while others are not. For this reason we include, in the source alphabet, symbols representing the tracking of pixels in the reference boundary. Thus, based on the chosen tradeoff between the rate and the distortion, the encoder may select to approximate stretches of the contour under consideration by following the reference contour for n pixels, with each value of n corresponding to one symbol. Since the previously reconstructed frame is available at the decoder, the next control point can be unambiguously determined by following the reference contour for a specified number of pixels.

2.3 Motion Vector Selection

The problem of selecting a suitable global motion vector between two binary masks is a non-trivial one. Approaches evaluated by MPEG-4 [2] use a single global motion vector per object minimizing the number of pixels in error between the current and the reference objects. With non-rigid motion, this criterion often spreads the error pixels all around the boundary, which is inconsistent with the objective of tracking the reference contour or utilizing context where possible. To better capitalize on the proposed code structure, the following criterion is used to choose a global motion vector:

$$\overline{M} = \arg\min_{i,j} \sum_{(i,j)\in E_{\overline{M}}} \{(i-\bar{i})^2 + (j-\bar{j})^2\},$$

where the summation is over the set of error pixels $E_{\overline{M}}$ whose center of mass is at $(\bar{i},\bar{j})$, under motion compensation of the reference contour by $\overline{M}$. Application of the proposed criterion has the effect of pushing the majority of error pixels to one side of the contour, while accurately approximating the rest. This, together with the tracking mode and the use of context, makes the encoder efficient with respect to non-rigid object motion and contour noise.

3 Scalable Shape Coding

Similar to the previous section, where *angle* and *run* contexts were used as reference for DPCM, here we use contexts to de-correlate base and enhancement layer contours in a scalable shape coder. The base layer is encoded by the intra mode ORD optimal shape coder [5], based on a specified rate distortion tradeoff λ_b. Then the enhancement layer, coded by the algorithm described in section ``Shape Coding in the Inter Mode'' with $\lambda_e < \lambda_b$, provides a more accurate approximation of the boundary, given the reconstructed base layer as reference. The tested hypothesis is that, perhaps the raw base layer approximation encoded with very few bits will serve as a good predictor for the *angle* component in the enhancement layer, thus saving bits in its DPCM description. A typical base layer approximation is shown in Fig. 3.

Figure 3: Shape approximation with the base layer only (R=73 bits, D=917 pels).

In order to make a fair RD comparison between a scalable and a non-scalable coder, we employ an iterative procedure to arrive at a locally most efficient set of VLCs [4,3] for the single layer approach, as well as, one for each scalable layer, since layer statistics differ.

While the encoding of all layers can be based on the same segment distortion metric, the base layer metric has no bearing on the contour distortion after decoding of the enhancement layer. In fact, *angle* contexts are the only information provided by the base layer, and, hence, it is desirable that the error between the original and the base layer context planes be as small as possible. For this reason, in the base layer, we employ a distortion metric that is based on the sum of absolute differences between those two context planes, evaluated at the pixels in error. This process can also be thought of as weighing the error pixels by the error in context.

No prediction is employed for the *run* component of the enhancement layer since it is not expected to be any correlation between control points of contours encoded with different values of λ. Rate distortion performance of the scalable coder is measured in terms of $R_b + R_e$ versus D_e, where R_b and R_e are the base and the enhancement layer rates, respectively. A particular realization of the tuplet (λ_b , λ_e) results in a point in the RD plane, which is not guaranteed to be on the convex hull. While there exist no theoretical guidelines for selecting λ_b and λ_e, resulting in a hull point, and in order to observe a (possible) empirical relationship between λ_b^{hull} and λ_e^{hull}, we use an exhaustive search (in the $\lambda_b \times \lambda_e$ space) to extract the convex hull.

4 Results

Figure 4 shows the ORD curves of the proposed algorithm for the SIF sequence ``kids'' at convergence of the VLC optimization algorithm described in [3]. The distortion axis d_n represents the average of the D_{MPEG4}'s defined as the ratio of the number of error pixels to the number of object interior pixels in a frame ([2]) for one frame, over 100 frames. As the figure demonstrates, our result compares favorably with both the baseline and the vertex-based algorithms (reviewed in [2]) in the inter mode across most of the range of bitrates. In the very low distortion region ($d_n \leq 0.006$) of operation, however, the proposed algorithm requires more bits

than both the baseline and the vertex-based methods. This is due to the fact that for near-lossless boundary encoding the chosen code structure (direction plus run) is inefficient.

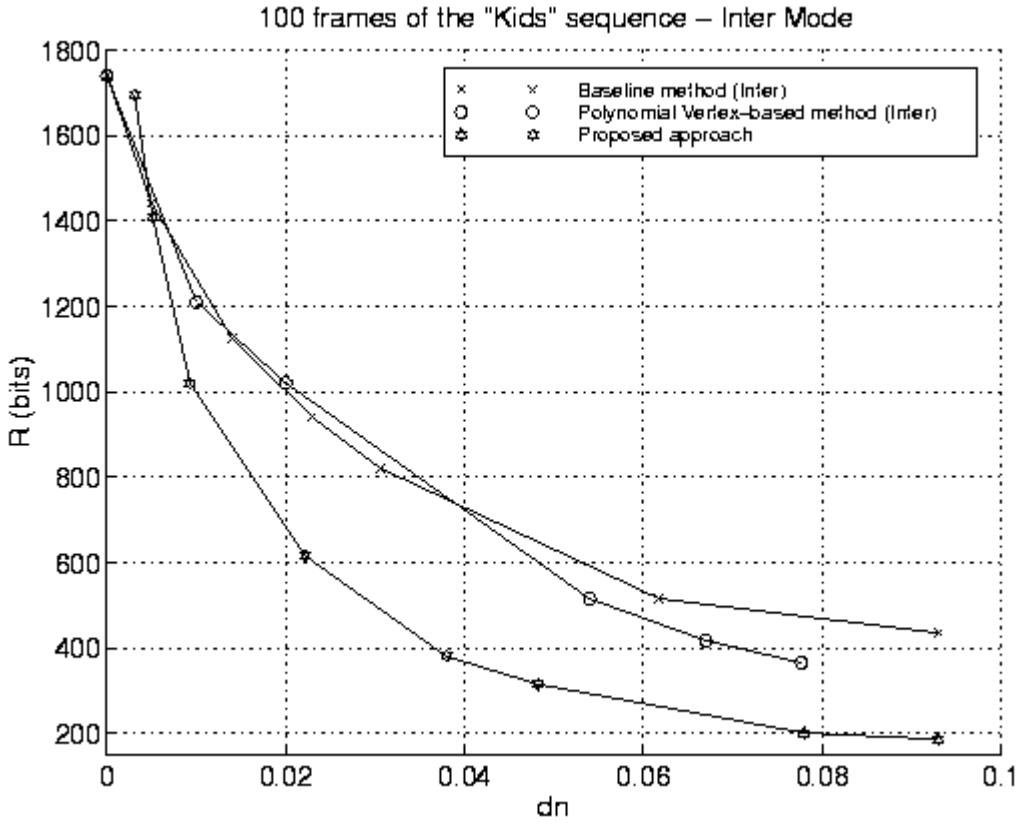

Figure 4: Rate-Distortion curves.

Rate distortion performance of a scalable shape coder is compared to its non-scalable counterpart at convergence of VLC optimization for just one object in Fig. 5.

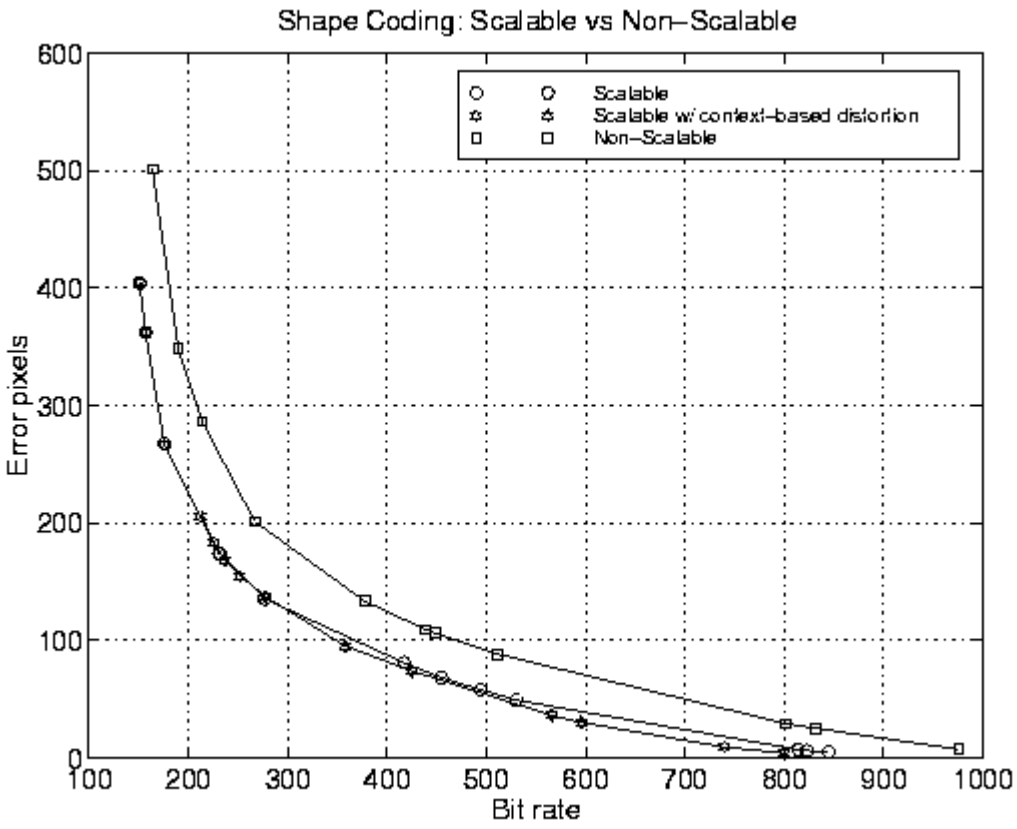

Figure 5: Rate-Distortion performance of the scalable coder.

Clearly, for one object, the scalable approach provides greater RD efficiency. The use of context-based distortion for the base layer resulted in further gains. However, more experiments are needed to draw similar conclusions, based on multiple objects.

5 References

[1] C. Jordan and T. Ebrahimi, ``Scalable vertex-based shape coding -S4h results", ISO/IEC/JTC1/SC29/WG11 MPEG96/2034 , Bristol, April 1997

[2] A. K. Katsaggelos, L. Kondi, F. W. Meier, J. Ostermann, G.M. Schuster, ``MPEG-4 and Rate Distortion Based Shape Coding Techniques", *Proc. IEEE*, pp. 1126-1154, June 1998.

[3] G. Melnikov, G. M. Schuster, A. K. Katsaggelos, ``Inter Mode Vertex-based Optimal Shape Coding", *Proc. ICASSP99*, Mar. 1999.

[4] D. Saupe, ``Optimal Piecewise Linear Image Coding", *Proc. SPIE Conf. on Visual Comm. and Image Proc.*, vol. 3309, pp. 747-760, 1997.

[5] G. M. Schuster and A. K. Katsaggelos, ``An optimal polygonal boundary encoding scheme in the rate distortion sense,'' *IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 13-26, Jan. 1998.

[6] G. M. Schuster, G. Melnikov, and A. K. Katsaggelos, ``Optimal Shape Coding Techniques,'' *IEEE Signal Processing Magazine*, pp. 91-108, Nov. 1998.