

# NONCAUSAL PREDICTIVE VIDEO CODEC OFFERING HIERARCHICAL QoS

*Maria Kouras and Amir Asif*

Computer Science York University  
North York, ON, Canada M3J 1P3  
Email: asif@cs.yorku.ca

## ABSTRACT

We present a scalable video coding scheme based on three dimensional (3D) Gauss Markov random process (GMrp) and replenishment extended vector quantization (VQ). The proposed video codec is capable of offering different quality of service (QoS) in the temporal and spatial domains. Compression results illustrate the superiority of the proposed scheme over the ITU standard, H.263, at high compression ratios.

## 1. INTRODUCTION

With the recent developments in the multimedia technology, streaming video has emerged as a popular data delivery format for many applications including video on demand, webcasts, and distance learning. In contrast to non-scalable video coding schemes, scalable codecs allow for multicasting video over heterogenous channels, provide different quality of services (QoS), and cope gracefully with the bandwidth fluctuations on the network.

In this paper, we propose a new scalable video codec for low bit rates based on noncausal prediction and vector quantization (VQ). Our scheme models the video sequence as a three dimensional (3D) Gauss Markov random process (GMrp) [1], which is used to generate a 3D error field considerable less correlated than the original video sequence. Cascaded VQ is then used to compress the error field. We apply extended replenishment to VQ [2] where the label of the current vector is encoded and transmitted only if it is different from the one at the same location in the previous frame. The proposed video codec provides for different QoS at the temporal and spatial levels. While the spatial QoS is a direct consequence of the cascaded VQ used to compress the prediction error, the temporal QoS is provided by decimating the video and using a different GMrp based video codec for the resulting streams. The proposed GMrp based video codec outperforms the ITU standard, H.263, both in terms of peak signal to noise ratio (PSNR) and perceived video quality in our simulations.

The paper is organized as follows. Section 2 reviews the 3D noncausal GMrp and develops a computationally efficient one-sided prediction model. Section 3 designs the GMrp based video codec with a note on its scalability in section 4. Section 5 compares the performance of the GMrp codec with H.263. Finally section 6 concludes the paper.

## 2. 3D NONCAUSAL GMRP

Our video codec uses a 3D noncausal Gauss Markov random process (GMrp) [1] to predict the intensity value of the reference pixel

based on its neighbouring pixels in the spatial and temporal domains. The intensity  $x(i, j, k)$  of a pixel at spatial location  $(i, j)$  in frame  $k$  is given by

$$\hat{x}(i, j, k) = \beta_v x(i-1, j, k) + \beta_v x(i+1, j, k) + \beta_h x(i, j-1, k) + \beta_h x(i, j+1, k) + \beta_t x(i, j, k-1) + \beta_t x(i, j, k+1) \quad (1)$$

where  $\beta_v$ ,  $\beta_h$  and  $\beta_t$  are respectively the vertical, horizontal, and temporal interactions. The prediction error is given by  $e(i, j, k) = x(i, j, k) - \hat{x}(i, j, k)$  for  $(1 \leq i \leq N_I)$ ,  $(1 \leq j \leq N_J)$ , and  $(1 \leq k \leq N_K)$ . Following lexicographic ordering, the pixels in row  $i$  of frame  $k$  are stacked into a  $(N_J \times 1)$  vector

$$X_{ik} = [x(i, 1, k) \ x(i, 2, k) \ \dots \ x(i, N_J, k)]^T, \quad (2)$$

which are arranged into a  $(N_I N_J \times 1)$  vector  $X_k$  as

$$X_k = [X_{1k}^T \ X_{2k}^T \ \dots \ X_{N_I k}^T]^T. \quad (3)$$

The frame vectors  $X_k$  are stacked one on top of the other, resulting into a  $(N_I N_J N_K \times 1)$  column vector

$$X = [X_1^T \ X_2^T \ \dots \ X_{N_K}^T]^T. \quad (4)$$

Using the vector notation defined in (2)-(4), the error field  $e(i, j, k)$  can be represented in matrix-vector form as  $\mathcal{A}X = e$  where  $\mathcal{A}$ , referred to as the potential matrix, has the following structure

$$\mathcal{A} = I_{N_K} \otimes A_1 + H_{N_K}^1 \otimes A_2 \quad (5)$$

with  $A_1 = I_{N_I} \otimes B + H_{N_I}^1 \otimes C$  and  $A_2 = I_{N_I} \otimes D$ . (6)

In (5)-(6), the symbol  $\otimes$  denotes the Kronecker product. The notation  $I_{N_I}$  represents the identity matrix and  $H_{N_K}^1$  denotes the Toeplitz matrix with zeros everywhere except for the first upper and first lower diagonals which consist of 1's. The subscripts denote the order of the matrices. The constituent blocks are

$$B = -\beta_h H_{N_J}^1 + I_{N_J}, \ C = -\beta_v I_{N_J}, \ \text{and} \ D = -\beta_t I_{N_J}. \quad (7)$$

To derive one-sided representations [3], we take the Cholesky factor of the potential matrix  $\mathcal{A} = \mathcal{U}^T \mathcal{U}$ . Since  $\mathcal{A}$  is block tridiagonal, therefore  $\mathcal{U}$  is upper triangular with only the main and first upper block diagonals being nonzero. Expanding  $\mathcal{A}X = e$  in terms of the diagonal  $\{U_k\}$  and upper diagonal blocks  $\{\Theta_k\}$ , gives

$$U_k X_k + \Theta_k X_{k+1} = \vec{w}_k, \ \text{for} \ (1 \leq k \leq N_K - 1) \quad (8)$$

$$U_{N_K} X_{N_K} = \vec{w}_{N_K} \quad (9)$$

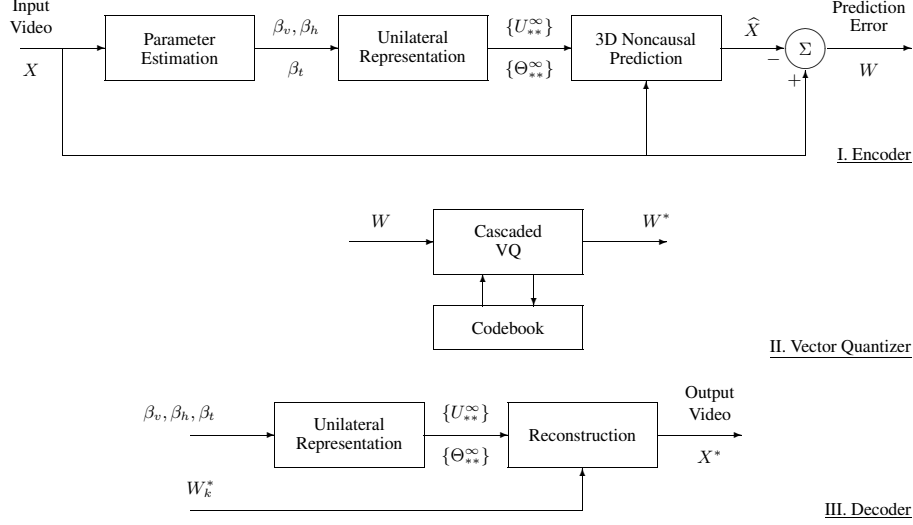


Fig. 1. Proposed video codec based on the 1st order 3D noncausal GMp.

where  $\vec{w} = U^{-T}e$  and is white [3]. The Cholesky blocks  $\{U_k, \Theta_k\}$  are obtained by expanding  $\mathcal{A} = U^T U$  in terms of the constituent subblocks as

$$U_1 = \text{chol}(A_1) \quad \text{and} \quad \Theta_1 = U_1^{-T} A_2 \quad (10)$$

$$U_k = \text{chol}(A_1 - \Theta_{k-1}^T \Theta_{k-1}) \quad \text{and} \quad \Theta_k = U_k^{-T} A_2 \quad (11)$$

for  $(2 \leq k \leq N_k)$ . The one-sided regression models (8)-(9) are computationally impractical to implement even for a reduced video format like QCIF. To derive practical implementations, we approximate the Cholesky blocks with  $L$ -block banded matrices. Before presenting the approximation, we comment first on the structure of the Cholesky factors. In this paper, we state the properties leaving the analytical proofs for an extended publication.

#### Structure of the Cholesky Factors:

1. The Cholesky blocks  $U_k$  and  $\Theta_k$  converge geometrically to steady state values  $(U^\infty, \Theta^\infty)$  as  $k$  increases.
2. Partitioning the upper triangular Cholesky block  $U_k$  in terms of subblocks  $\{U_{ij}^{(k)}\}$  and the lower triangular Cholesky block  $\Theta_k$  in terms of subblocks  $\{\Theta_{ij}^{(k)}\}$ , it is observed that the off-diagonal subblocks  $\{U_{ij}^{(k)}, \Theta_{ij}^{(k)}\}$  converge to  $\underline{0}$  along block row  $i$  as we move away from the main diagonal.
3. The subblocks  $\{U_{ij}^{(k)}\}$  and  $\{\Theta_{ij}^{(k)}\}$  converge to steady state values along  $|j - i|$  block diagonals.

Based on observations 1-3, we approximate the Cholesky blocks  $U_k$  in the Cholesky factor  $U$  by a bidiagonal upper triangular block matrix with subblocks  $U_{ii}^{(k)}$  on the main block diagonal and subblocks  $U_{i+1,i}^{(k)}$  on the first upper block diagonal. Similarly, the Cholesky blocks  $\Theta_k$  in the Cholesky factor  $U$  are approximated by a bidiagonal lower triangular block matrix with blocks  $\Theta_{ii}^{(k)}$  on the main block diagonal and blocks  $\Theta_{i+1,i}^{(k)}$  as illustrated below.

$$U_k = \begin{bmatrix} U_{11}^{(k)} & U_{12}^{(k)} & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & U_{N_I-1, N_I-1}^{(k)} & U_{N_I-1, N_I}^{(k)} \\ & & & & U_{N_I, N_I}^{(k)} \end{bmatrix} \quad (12)$$

$$\text{and } \Theta_k = \begin{bmatrix} \Theta_{11}^{(k)} & & & & \\ \Theta_{21}^{(k)} & \Theta_{22}^{(k)} & & & \\ & \ddots & \ddots & & \\ & & & \Theta_{N_I, N_I-1}^{(k)} & \Theta_{N_I, N_I}^{(k)} \end{bmatrix}. \quad (13)$$

Based on the above block banded approximation, the one sided backward representation of (8)-(9) is simplified considerably.

**Reduced One-sided Prediction Model:** Expanding (8)-(9) in terms of the constituent blocks (12)-(13), the error signal  $\vec{w}$  is represented in terms of row vector  $\vec{w}_{ik}$  corresponding to the  $i$ 'th row in frame  $k$  of the video sequence allowing for row by row computation of  $\vec{w}$ .

$$\text{Frame } (k = N_K): U_{N_I, N_I}^{(N_K)} X_{N_I, N_K} = \vec{w}_{N_I, N_K} \quad (14)$$

$$U_{ii}^{(N_K)} X_{i, N_K} + U_{i+1, i}^{(N_K)} X_{i+1, N_K} = \vec{w}_{i, N_K} \quad (15)$$

for  $(N_I - 1) \leq i \leq 1$

Frame  $(N_K - 1 \geq k \geq 1)$ :

$$U_{N_I, N_I}^{(k)} X_{N_I, k} + \Theta_{N_I, N_I-1}^{(k)} X_{N_I-1, k+1} + \Theta_{N_I, N_I}^{(k)} X_{N_I, k+1} = \vec{w}_{N_I, k} \quad (16)$$

$$U_{ii}^{(k)} X_{i, k} + U_{i+1, i}^{(k)} X_{i+1, k} + \Theta_{i, i-1}^{(k)} X_{i-1, k+1} + \Theta_{ii}^{(k)} X_{i, k+1} = \vec{w}_{i, k} \quad (17)$$

$$U_{11}^{(k)} X_{1, k} + U_{12}^{(k)} X_{2, k} + \Theta_{11}^{(k)} X_{1, k+1} = \vec{w}_{1, k} \quad (18)$$

In computing the error image, the memory requirements for storing the equivalent one-sided recursive representation is drastically reduced because we replace  $U^{(k)}$  and  $\Theta^{(k)}$  by their steady state values  $U^\infty$  and  $\Theta^\infty$ . These steady state values are further approximated by the bidiagonal approximation using (12)-(13). Thus we only have to store the pairs  $(U_{ii}^\infty, U_{i+1, i}^\infty)$  and  $(\Theta_{ii}^\infty, \Theta_{i+1, i}^\infty)$  for a small number of rows  $i$  until convergence is achieved.

### 3. VIDEO CODEC

The compression and reconstruction steps in the video codec are summarized in fig.1. The encoder involves the following steps:

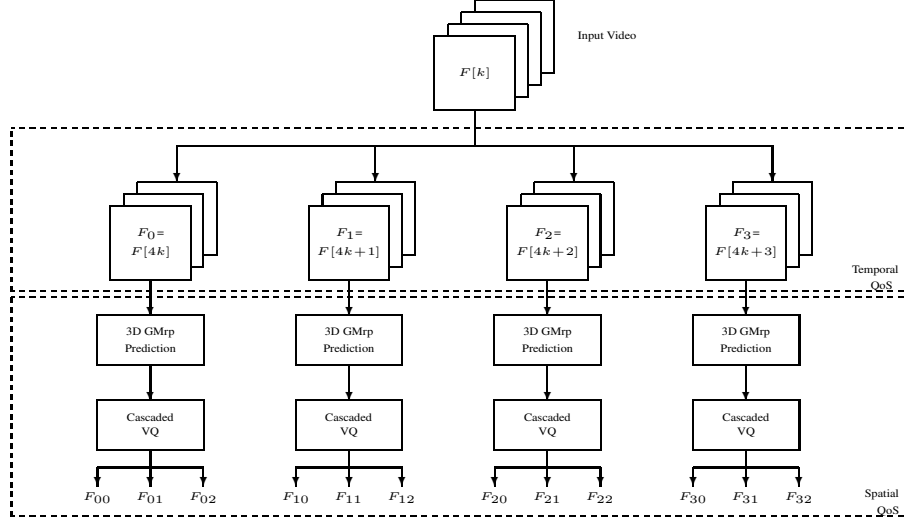


Fig. 2. Block diagram representation for the proposed video codec offering different quality of services (QoS)..

1. **Parameter Estimation:** The vertical, horizontal, and temporal interactions ( $\beta_v, \beta_h, \beta_t$ ) are estimated from the video and also transmitted to the receiver as overhead.
2. **Unilateral Representation:** Based on the values of the field interactions, the steady state values of the regressors ( $U_{ii}^\infty, U_{ii+1}^\infty$ ) and ( $\Theta_{ii}^\infty, \Theta_{ii+1}^\infty$ ) are computed.
3. **Noncausal GMrp Prediction:** Using the unilateral prediction model, (14)-(18), the error video  $\vec{w}$  is computed for each frame  $k$ , a row  $i$  at each iteration.
4. **Vector Quantization (VQ):** To achieve high compression, the error video  $\vec{w}$  is vector quantized using cascaded VQ preceded by quadtree mean removal. A global codebook based on a set of training video sequences is generated for VQ. We use the extended replenishment VQ scheme proposed in [2] where a label of the current vector is transmitted only if it is different from the one at the same location in the previous frame. This results in significant compression over the conventional VQ [4].

The reconstruction of the video is performed by inverting steps 1 to 4 in the reverse order as illustrated in the decoder of fig. 1.

#### 4. QUALITY OF SERVICE

Our video codec provides for different quality of services (QoS) [5] at the temporal and spatial levels. We illustrate the process in fig. 2 where for simplicity we assumed a 3-stage cascaded VQ to compress the 3D error sequence.

**Spatial QoS:** is a direct consequence of the cascaded VQ used to compress the prediction error. For the best spatial quality video, the output of all stages ( $n = 3$ ) of VQ is transmitted to the receiver. For intermediate spatial qualities, the output of a reduced number of stages is transmitted.

**Temporal QoS:** is provided by offering different refresh rates to the receiver. In our experimental setup, we decimate the video in four streams. Each stream contains every fourth frame. Stream 1 contains frames 1, 5, 9, and so on till the end of the video feed. Stream 2 contains frames 2, 6, 10, and so on. Stream 3 consists

of frame 3, 7, 11, and so on, while stream 4 consists of frames 4, 8, 12, and so on till the end of frames. Each stream is compressed separately using a different GMrp based video codec. For the best temporal quality, the prediction error from all four streams is transmitted to the receiver. The receiver will run four different decoders and interleave the decoded frames before displaying them. For intermediate temporal quality, the prediction error from streams 1 and 3 (or streams 2 and 4) is transmitted to the receiver. In this case, the receiver runs two decoders and interleave the two decoded streams. For the lowest temporal quality, the prediction error for any one of the four streams (stream 1, 2, 3, or 4) is transmitted. The receiver runs a single GMrp decoder and displays the decoded feed.

**Choice of Service:** We propose three quality of services (Gold, Silver and Bronze) by combining the spatial and temporal feeds. We explain the services with reference to fig. 3.

**Bronze Service:** is based on feed  $F_{00}$ . The frame rate is therefore one-fourth of the original video.

**Silver Service:** uses feeds  $F_{00}, F_{01}$  and  $F_{20}, F_{21}$ . The frame rate is one-half of the streaming video.

**Gold Service:** uses all feeds  $F_{00}$  to  $F_{32}$ .

#### 5. EXPERIMENTS

The experiments presented here are designed to make two major points. First, we show that reasonably good quality can be obtained at low bit rates using the proposed compression procedure and these results are superior to those obtained at similar bit rates using the ITU standard, H.263 baseline system. In the second set of experiments, we seek to compare the performance of the different quality of services presented in the paper. The reconstructed sequences in the second set of experiments are therefore compressed to different compression ratios, both temporally and spatially. We illustrate how much improvement is obtained as a client moves from a lower class of service to a higher one.

Fig. 3 shows the peak signal to noise ratio (PSNR) computed from the ‘‘Salesman sequence’’ after its encoding by the proposed system and the H.263 standard at different compression ratios.

Here we used the Gold QoS for the proposed codec with a single stage VQ as we are interested in comparing the overall quality of the proposed scheme with H.263. Fig. 3 indicates that the proposed system has a significantly higher PSNR than the H.263 standard especially at high compression ratios (CR). To provide subjective evaluation of the sequences, a representative frame reconstructed from the proposed codec and H.263 is included in fig. 4. The frame compressed using GMrp codec exhibits good visual quality with most of the detail of the original being retained, e.g., the structure of the eye and eyebrows. Moreover, there is little blocking visible in the frame despite the fact that VQ is prone to introducing blocking at low bit rates. Similar observations were made for other frames in the test sequence.

In the second set of experiments, we compare the improvement in the visual quality obtained by switching from a lower quality of service to a higher service. Fig. 5 illustrates a representative frame from the Salesman sequence compressed using the Gold, Silver, and Bronze services. Not depicted in the frame is the difference in the frame rates offered with each service. The Gold service is offered at the original frame rate of 30 fps while the Silver service skips every second frame and the Bronze service displays only the fourth frame in the sequence. The frame rate for Silver service is 15 fps while the frame rate for Bronze service is 7.5 fps. We used the 321-cascaded VQ to compress the error image obtained after GMrp prediction. The overall compression ratio for the Gold service is 149 compared to the compression ratios of 271 and 424 respectively for the Silver and Bronze services. As expected there is a noticeable difference in the visual quality between the three services because of the difference in the compression ratios. However, the frame compressed using the Bronze service is intelligible and comprehensible despite the high compression ratio.

## 6. CONCLUSION

The paper presents a new procedure for compression of video sequences based on modeling the video using a 3D Gauss Markov random process (GMrp) and quadtree cascaded vector quantization (VQ). The proposed video codec outperforms the ITU H.263 video compression standard at high compression ratios and is capable of offering different quality of services (QoS) both in the temporal and spatial domains.

## 7. REFERENCES

- [1] S. M. Schweizer and J. M. F. Moura, "Hyperspectral Imagery: Clutter Adaptation in Anomaly Detection," *IEEE Transactions on Information Theory*, vol. 46(5), pp. 1855-1871, Aug. 2000.
- [2] M. Goldberg and H. Sun, "Image Sequence Coding using Vector Quantization," *IEEE Transactions on Communications*, vol. COM-34, pp. 703-710, Mar. 1986.
- [3] A. Asif, "A Fast Implementation of the RTS Smoother for Image Deblurring," submitted to *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP 2004.
- [4] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. COM-28, pp. 232-240, Jan. 1980.
- [5] L. Xu, "Resource-Efficient Delivery of On-Demand Streaming Data Using UEP Codes," *IEEE Transactions on Communications*, vol. 51(1), pp. 63-71, Jan. 2003.

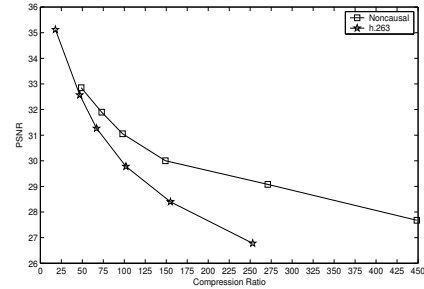


Fig. 3. Comparison of PSNR between the proposed noncausal codec and H.263 at different compression ratios.



Fig. 4. Selected frames from the "Salesman" sequence. The frame at the top is the original, the bottom left is compressed using H.263 (CR = 253), and the bottom right is compressed using the proposed GMrp codec (CR = 271).

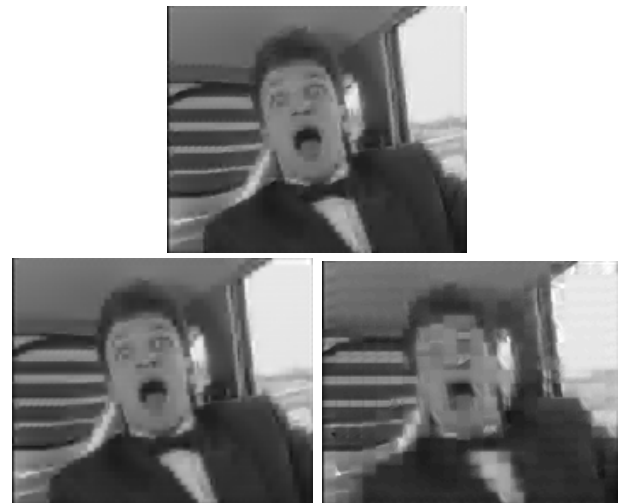


Fig. 5. Reconstructed frames from the Gmrp codec at three different QoS. The top frame is obtained using the Gold service, the bottom left using the Silver service, and the bottom right using the Bronze service.