

Scalable Video Codec by Noncausal Prediction, Cascaded Vector Quantization, and Conditional Replenishment

Amir Asif, *Senior Member, IEEE*, and Maria Kouras

Abstract—In this paper, we describe a bandwidth adaptable, low bit-rate video coding scheme, referred to as scalable, noncausal prediction with vector quantization and conditional replenishment (SNP/VQR). Practical implementations of SNP/VQR are derived by exploiting the convergence and block banded properties of the state matrices. The resulting subblock SNP/VQR reduces the computational complexity and the storage requirements of the direct SNP/VQR by two orders of the linear magnitude of the frame dimensions. The subblock SNP/VQR is also capable of offering different quality of services (QoS) in the spatial and temporal domains. In the experiments, the block SNP/VQR compares favorably with the standard video codecs including the International Standards Organization (ISO) proposed MPEG4 and the International Telecommunication Union (ITU) proposed H.263, especially at low bit rates.

Index Terms—Conditional replenishment, Gauss Markov random process, noncausal prediction, quality of service, scalable video compression, vector quantization.

I. 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. Even after compression, streaming video data is large and consumes a great deal of server and network resources. Consider, for example, a *live* video stream, such as a sports event, simultaneously requested by multiple users through the Internet. If the data stream is delivered using unicast, i.e., via a dedicated channel for each request, the load on the network is roughly proportional to the number of users. A point-to-point unicast service is, therefore, not practical for such live transmissions with a wide audience. Instead, a more efficient approach is to multicast a single video stream over the network. Multicasting requires a scalable video coding scheme [1]–[8] such that the encoded bit stream can be parsed according to the bandwidth available at each receiver. Scalable video codecs offer other advantages including the capability of dynamically changing the spatial or temporal

resolution and are more resilient to bandwidth fluctuations [9], [10].

In this paper, we propose a new scalable, noncausal predictive video codec for transmissions at low bit rates. We are primarily interested in multimedia applications (such as live video conferencing or surveillance video) on wireless channels with transmission rates limited to 200 Kbps. At such low bit rates, the ISO standardized MPEG family and the ITU standardized H.xxx family exhibit several visual degradations that appear unacceptable on high resolution screens. This offers an opportunity for new video codecs capable of providing better visual quality at low bit rates.

Our video codec couples 3-D scalable, noncausal prediction (SNP) with vector quantization and conditional replenishment (VQR), and is referred to as SNP/VQR. The novelty and superior performance of SNP/VQR is due to the noncausal prediction paradigm. It contributes to the existing literature of the noncausal predictive codecs in the following ways.

- 1) The earlier applications of noncausal prediction [11], [12] are limited to two-dimensional (2-D) still image processing. In this paper, we develop a recursive framework for three-dimensional (3-D) noncausal prediction models and illustrate its usefulness in the context of video compression. Extension of 2-D noncausal prediction to compression of 3-D video sequences is challenging as the resulting implementation, referred to as the direct SNP/VQR, is computationally intensive. For an $(N_I \times N_J \times N_K)$ video sequence, the computational complexity of SNP/VQR is of $O(N_I^2 N_J^2 N_K^2)$, which precludes its application from real time communications.
- 2) Practical implementations of the SNP/VQR, referred to as the block SNP/VQR and the subblock SNP/VQR, are derived. The block SNP/VQR operates on each frame of the video and is designed by noting that the constituent blocks $\{L^{(k)}, F^{(k)}\}$ in the state matrix \mathcal{A} of the 3-D noncausal prediction model converge at a geometric rate. The block SNP/VQR provides a savings of $O(N_K)$ over the direct SNP/VQR. Further reduction in the complexity is achieved by observing that the subblocks $\{L_{l_1 l_2}^{(k)}, F_{l_1 l_2}^{(k)}\}$ constituting the state blocks $\{L^{(k)}, F^{(k)}\}$ converge to $\underline{0}$ along each block row l_1 . This leads to the subblock SNP/VQR, which exploits the block banded structure of the state blocks. The computational complexity of the subblock SNP/VQR is of $O(N_I N_J^2 N_K)$, a reduction of

Manuscript received January 13, 2004; revised March 11, 2005. This work was supported in part by the Natural Science and Engineering Research Council (NSERC) of Canada under Grant 229862. The associate editor coordinating the review of this manuscript and approving for publication was Dr. Chalapathy Neti.

A. Asif is with the Department of Computer Science and Engineering, York University, Toronto, ON M3J 1P3 Canada (e-mail: asif@cs.yorku.ca).

M. Kouras is with the Ministry of Health and Long-Term Care of Ontario, Toronto, ON M3J 1P3, Canada (e-mail: es203529@cs.yorku.ca).

Digital Object Identifier 10.1109/TMM.2005.861294

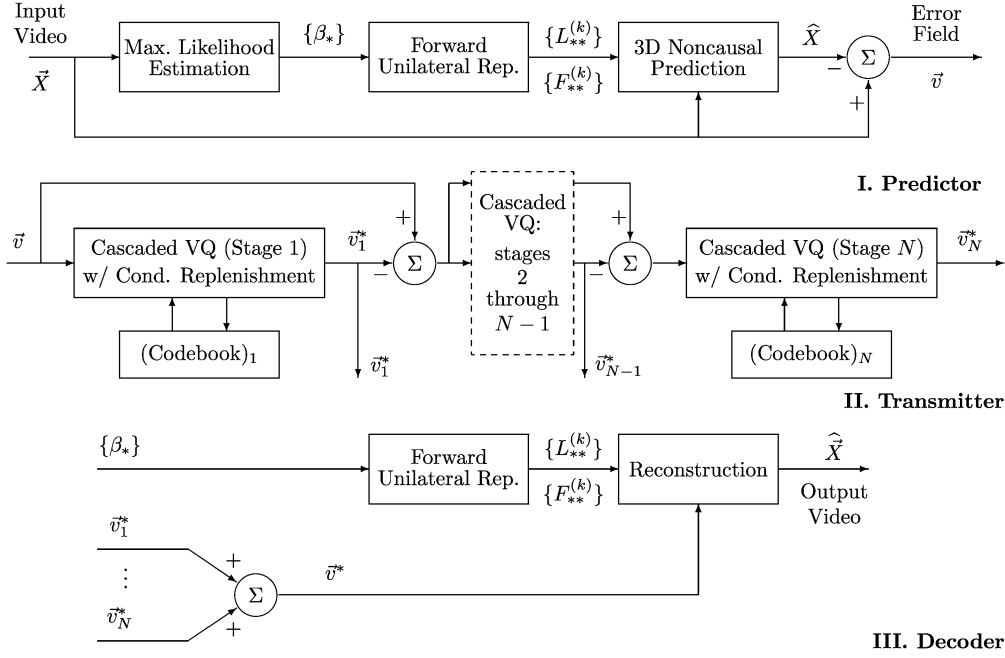


Fig. 1. Block diagram representation of the SNP/VQR video codec.

two orders of magnitude over the direct SNP/VQR. The subblock SNP/VQR also reduces the storage requirements by a factor of $N_I N_K$.

- 3) The subblock SNP/VQR is bandwidth scalable, where a subset of the compressed video stream provides the base quality. Additional enhancement layers build on the video quality. Scalability allows the subblock SNP/VQR to dynamically adjust to the bandwidth variations in the wireless network by transmitting a reduced number of layers at times of congestion.

We compare the results of SNP/VQR with the ISO standard MPEG4 [18] and the ITU standard, H.263 [17]. In our simulations, SNP/VQR compares favorably with these standards, both in terms of peak signal to noise ratio (PSNR) and perceived video quality, especially at low bit rates. The video sequences compressed with MPEG4 and H.263 show visible blocking, while the videos reconstructed with the subblock SNP/VQR display relatively minimal visual artifacts.

The paper is organized as follows. Section II introduces SNP/VQR and develops the unilateral backward and forward representations for noncausal prediction. Section III derives the practical implementations of the SNP/VQR, which recursively generate the error field by iterating along the rows of the video frames. Section IV reviews replenishment extended VQ and describes how SNP/VQR offers different QoS in the temporal and spatial domains. In Section V, we present our experimental results and compare the performance of SNP/VQR with the standard codecs, MPEG4 and H.263, at low bit rates. Finally, Section VI concludes the paper.

II. SNP/VQR VIDEO CODEC

As illustrated in Fig. 1, the compression procedure of SNP/VQR has a predictive component followed by a quantiza-

tion component. The predictive component, shown in part I of Fig. 1, has four stages. In the first stage, the vertical, horizontal, and temporal interactions $\{\beta_h, \beta_v, \beta_t\}$ are estimated from the input video. These interactions define the state or potential matrix \mathcal{A} used for prediction. The interaction parameters are also required at the decoder to reconstruct the video and constitute overhead information transmitted to the receiver. In the second stage, the 3-D noncausal model is transformed to an equivalent unilateral representation that uses block triangular matrices L and F for prediction. The recursive form is obtained by Cholesky factorization of the potential matrix \mathcal{A} . In the third stage, the unilateral prediction model is used to estimate each frame of the video. Stage four generates the uncorrelated error field \vec{v} by subtracting the values of the pixels in the predicted frame from the original pixel values. Since the noncausal prediction at the receiver is based on the reconstructed frames, we also use the reconstructed frames for prediction at the encoder.

To achieve high compression ratios, the error video \vec{v} is vector quantized using cascaded VQ. We apply conditional replenishment [14] at each stage of cascaded VQ where a vector quantized block is encoded and transmitted only if its VQ index is different from the corresponding index at the same location in the previous frame. Conditional Replenishment leads to considerable reduction in the number of code vectors transmitted. The vector quantization step is shown in part II of Fig. 1. Although not implemented in our codec, conditional replenishment can be coupled with motion compensation schemes to provide better tradeoffs between video quality and compression ratio than conditional replenishment alone.

The reconstruction of the video is performed by inverting the steps of the encoder in the reverse order as shown in part III of Fig. 1. The state variable representation at the decoder is obtained by converting the noncausal model to its equivalent re-

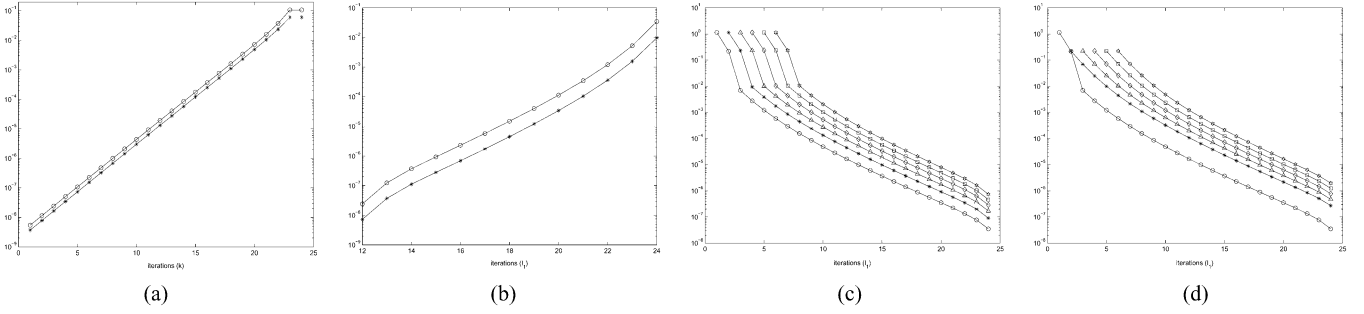


Fig. 2. Illustration of the convergence properties in the Cholesky blocks for $\{\beta_v = 0.156631, \beta_h = 0.166309, \beta_t = 0.167446\}$. (a) Plots of $\|L^{(k)} - L^\infty\|$ and $\|F^{(k)} - F^\infty\|$ versus frame k . (b) Plots of $\|L_{l_1+1}^\infty - L_{l_1}^\infty\|$ and $\|F_{l_1+1}^\infty - F_{l_1}^\infty\|$ versus block row l_1 . (c) Plots of $\|L_{l_1 l_2}^\infty\|$ for the last five rows l_1 and $(1 \leq l_2 \leq l_1)$ in L^∞ . (d) Plots of $\|F_{l_1 l_2}^\infty\|$ for the first five rows l_1 and $(l_1 \leq l_2 \leq N_I)$ in F^∞ .

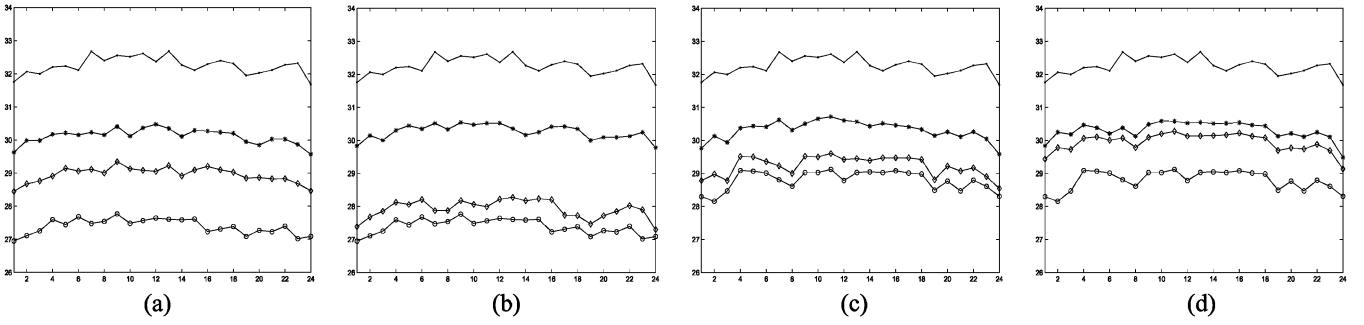


Fig. 3. Comparison of PSNRs for frames (1–24) of a test video sequence compressed with different versions of a 3-stage 6-bit cascaded VQ. In subplots (a)–(d), the PSNR of the video frame reconstructed from the first stage is shown with a “ \diamond ”, the PSNR of the video frame reconstructed from the first two stages is shown with a “ \circ ”, and the PSNR of the video frame reconstructed by combining the output of all three stages is shown with a “ $*$ ”. In each subplot, we also show the PSNR for the frames reconstructed using a single stage VQ with 6 bits (64 code vectors) quantization with a “ \bullet ”. Subplot 4(a) uses a 222-configuration for cascaded VQ, subplot 4(b) uses the 213-cascaded VQ, subplot 4(c) uses the 312-cascaded VQ, while subplot 4(d) uses the 321-cascaded VQ. The horizontal axis represents the frame index k .

cursive form exactly in the same way as done at the encoder. The quantized error field is used to reconstruct the video.

We now explain the prediction step in more details, highlighting the computational issues involved in prediction.

A. Noncausal Prediction

The video is modeled as a 3-D noncausal Gauss Markov random process (GMrp) [15], [16] where we use a bilateral neighborhood of pixels to make a linear prediction of the current pixel value. This implies that the video is represented by a bilateral autoregressive linear model driven by a correlated input noise. This is the 3-D extension of the minimum mean square error (MMSE) representation derived by Woods [15] and, for a first order video, is given by

$$\begin{aligned} & x(i, j, k) - \beta_v(x(i+1, j, k) + x(i-1, j, k)) \\ & - \beta_h(x(i, j+1, k) + x(i, j-1, k)) \\ & - \beta_t(x(i, j, k+1) + x(i, j, k-1)) \\ & = e(i, j, k) \end{aligned} \quad (1)$$

where $x(i, j, k)$ is the pixel intensity at spatial location (i, j) and frame k , and similarly for the remaining pixels. The input error field is denoted by $e(i, j, k)$ and β_v , β_h , and β_t are the model parameters, referred to as the vertical, horizontal, and temporal interactions. As customary in video processing, we use the row-major order to vectorize video $x(i, j, k)$ into vector \vec{X}

and the error field $e(i, j, k)$ into vector \vec{e} . The pixels in frame k are, therefore, arranged in a column vector

$$X^{(k)} = \begin{bmatrix} X_1^{(k)T} & X_2^{(k)T} & \dots & X_{N_I}^{(k)T} \end{bmatrix}^T \quad (2)$$

where $X_i^{(k)}$ represents pixels in row i of frame k . Vectors $X^{(k)}$ are then stacked in order such that

$$\vec{X} = [X^{(1)T} \quad X^{(2)T} \quad \dots \quad X^{(N_k)T}]^T. \quad (3)$$

The same procedure is used to generate $e_i^{(k)}$ representing the error field on row i of frame k , and vector $e^{(k)}$ constituting the error field of frame k . Stacking all $N_I N_J N_K$ equations corresponding to the 3-D video lattice, the MMSE representation of the 3-D GMrp is written in three alternative forms summarized by Theorem 1.

Theorem 1: The following are three equivalent representations for a 3-D, first order, noncausal GMrp with zero Dirichlet boundary conditions.

1) Bilateral Representation:

$$\mathcal{A}\vec{X} = \vec{e} \quad (4)$$

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

$$\begin{aligned} & A_1 = I_{N_I} \otimes B + H_{N_I}^1 \otimes C, \\ & \text{and } A_2 = I_{N_I} \otimes D. \end{aligned} \quad (6)$$

In (6) and (7), the symbols I_{N_K} and I_{N_I} are identity matrices, while $H_{N_K}^I$ and $H_{N_I}^I$ are Toeplitz matrices that have zeros everywhere except for the first upper and lower diagonals, which are composed of all ones. The subscript denotes the order of the matrix. The operator \otimes represent the Kronecker product and the constituent blocks B , C , and D are given by

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

2) Forward Unilateral Representation:

$$L^{(1)} X^{(1)} = v^{(1)} \quad (8)$$

$$F^{(k)} X^{(k-1)} + L^{(k)} X^{(k)} = v^{(k)}, \quad (2 \leq k \leq N_K) \quad (9)$$

where $v^{(k)}$ represents the row-ordered pixels in frame k of the 3-D whitened error field $v(i, j, k)$ obtained from the transformation $\vec{v} = \mathcal{L}^{-T} \vec{e}$. The forward Cholesky factor \mathcal{L} is derived from the *upper/lower* Cholesky decomposition $\mathcal{A} = \mathcal{L}^T \mathcal{L}$ and has the following structure

$$\mathcal{L} = \begin{bmatrix} L^{(1)} & \underline{0} & \cdot & \cdot & \underline{0} \\ F^{(2)} & L^{(2)} & \underline{0} & \cdot & \underline{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \underline{0} & \cdot & F^{(N_K-1)} & L^{(N_K-1)} & \underline{0} \\ \underline{0} & \cdot & \underline{0} & F^{(N_K)} & L^{(N_K)} \end{bmatrix}. \quad (10)$$

The forward Cholesky blocks $L^{(k)}$'s in \mathcal{L} are lower triangular and $F^{(k)}$'s are upper triangular.

3) Backward Unilateral Representation:

$$U^{(N_K)} X^{(N_K)} = w^{(N_K)} \quad (11)$$

$$U^{(k)} X^{(k)} + \Theta^{(k)} X^{(k+1)} = w^{(k)} \quad (12)$$

for $(N_K - 1 \geq k \geq 1)$. The vector $w^{(k)}$ represents the row-ordered pixels in frame k of the 3-D whitened error field $\vec{w} = U^{-T} \vec{e}$. The backward Cholesky factor \mathcal{U} is obtained from the *lower/upper* Cholesky decomposition $\mathcal{A} = \mathcal{U}^T \mathcal{U}$ and has the following structure

$$\mathcal{U} = \begin{bmatrix} U^{(1)} & \Theta^{(1)} & \underline{0} & \cdot & \underline{0} \\ \underline{0} & U^{(2)} & \Theta^{(2)} & \cdot & \underline{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \underline{0} & \cdot & \underline{0} & U^{(N_K-1)} & \Theta^{(N_K-1)} \\ \underline{0} & \cdot & \cdot & \underline{0} & U^{(N_K)} \end{bmatrix}. \quad (13)$$

The backward Cholesky blocks $U^{(k)}$'s are upper triangular and $\Theta^{(k)}$'s are lower triangular. ■

Given that the covariance of the random vector \vec{X} is $\sigma^2 \mathcal{A}$, it is straightforward to sure that the error fields \vec{v} and \vec{w} are white. In other words, we have completely uncorrelated the 3-D video field with the unilateral representations. Theoretically, complete decorrelation of the video sequence leads to higher compression with the reconstructed sequences exhibiting improved perceived quality and a higher PSNR.

To obtain the forward Cholesky blocks $\{L^{(k)}, F^{(k)}\}$'s in the forward unilateral representation, we expand $\mathcal{A} = \mathcal{L}^T \mathcal{L}$

in terms of the constituent blocks, leading to the following relationships:

$$\left. \begin{aligned} L^{(N_K)} &= \text{chol}(A_1) \\ F^{(N_K)} &= (L^{(N_K)})^{-T} A_2 \end{aligned} \right\} (k=1) \quad (14)$$

$$\left. \begin{aligned} L^{(k)} &= \text{chol}(A_1 - (F^{(k+1)})^T F^{(k+1)}) \\ F^{(k)} &= (L^{(k)})^{-T} A_2. \end{aligned} \right\} \quad (15)$$

for $(N_K - 1 \geq k \geq 1)$

where the notation $L_{N_K} = \text{chol}(A_1)$ indicate that the matrix L_{N_K} is the Cholesky factor of A_1 . Likewise the backward Cholesky blocks $\{U^{(k)}, \Theta^{(k)}\}$'s in the backward unilateral representation are obtained by expanding $\mathcal{A} = \mathcal{U}^T \mathcal{U}$ in terms of the constituent blocks. The resulting expressions are

$$\left. \begin{aligned} U^{(1)} &= \text{chol}(A_1) \\ \Theta^{(1)} &= (U^{(1)})^{-T} A_2 \end{aligned} \right\} (k=1) \quad (16)$$

$$\left. \begin{aligned} U^{(k)} &= \text{chol}(A_1 - (\Theta^{(k-1)})^T \Theta^{(k-1)}) \\ \Theta^{(k)} &= (U^{(k)})^{-T} A_2. \end{aligned} \right\} \quad (17)$$

for $(N_K - 1 \geq k \geq 1)$.

The values of the interaction parameters β_v , β_h , and β_t as well as the noise variance σ^2 are obtained by maximization of the likelihood function

$$P(\vec{X}) = \frac{|\mathcal{A}|^{1/2}}{(2\pi\sigma^2)^{N_I N_J N_K / 2}} \exp \left\{ -\frac{1}{2\sigma^2} \vec{X}^T \mathcal{A} \vec{X} \right\}. \quad (18)$$

See [12] for details. Theorem 1 provides two different block implementations for 3-D noncausal prediction of the video pixels. The forward unilateral representation is more suitable for live streaming applications since the video is transformed in its natural order $(1 \leq k \leq N_K)$ into a 3-D uncorrelated error field. In the subsequent discussion, we focus on the forward unilateral representation.

III. PRACTICAL SNP/VQR CODEC

The forward unilateral regression model (8)–(9) is computationally impractical to implement even for a reduced video format like QCIF with a frame size of (144×176) pixels. For a QCIF video sequence, the linear dimension of the Cholesky blocks $(L^{(k)}, F^{(k)})$ in the forward regression model is roughly of $O(2.5 \times 10^4)$. Storage and matrix operations at such high dimensions are clearly not feasible. To derive practical implementations, we approximate the Cholesky blocks by an M -block banded structure. Before presenting the subblock implementation, we comment first on the structure of the Cholesky blocks, which provide intuitive justification for the block banded approximation.

A. Cholesky Factors

The structure of the Cholesky blocks $\{L^{(k)}, F^{(k)}\}$ is illustrated through a simple example. A first order Dirichlet field with $\beta_v = 0.156631$, $\beta_h = 0.166309$, and $\beta_t = 0.167446$ is defined on a 3-D $(24 \times 24 \times 24)$ lattice, i.e., $N_I = N_J = N_K = 24$. The Cholesky blocks $\{L^{(k)}, F^{(k)}\}$ are computed using relationships (14)–(15). The following observations are made from the computed values of the Cholesky blocks.

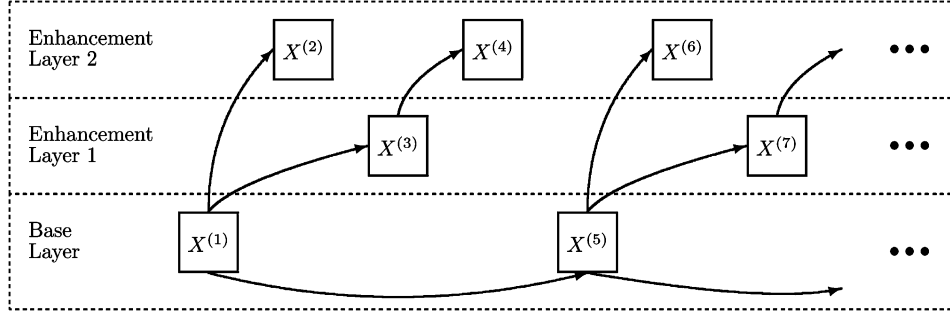


Fig. 4. Reconfiguration of video frames to achieve temporal scalability. Starting from the first frame $X^{(1)}$, the base layer encodes every fourth frame using the block SNP/VQR. Enhancement layer 1 starts with the third frame $X^{(3)}$ and encodes every fourth frame using frames reconstructed from the base layer in the 3-D GMp predictive model. Enhancement layer 2 encodes the remaining frames of the video sequence and uses the frames reconstructed from the first enhancement and base layers during its prediction.

Property 1: In Fig. 2(a), we plot the norm of the differences $\|L^{(k)} - L^\infty\|$ and $\|F^{(k)} - F^\infty\|$ for $(N_k \leq k \leq 1)$. The plots in Fig. 2(a) highlight the rapid geometric convergence of the sequences $L^{(k)}$ and $F^{(k)}$ in a small number of iterations k . This result has been proved theoretically in [16].

Property 2: In Fig. 2(b), we plot the norm of the differences $\|F_{l_1+1l_1+1}^\infty - F_{l_1l_1}^\infty\|$ of the constituent subblocks in F^∞ . The plot is shown as a solid line marked with the symbol “*”. Likewise, the norm of the differences $\|L_{l_1+1l_1+1}^\infty - L_{l_1l_1}^\infty\|$ for consecutive subblocks on the main diagonal in L^∞ is plotted as a solid line marked with the symbol “o”. We observe that the constituent subblocks in F^∞ and L^∞ themselves converge along the block diagonals. Similar convergences were observed for other subblocks in F^∞ and L^∞ along diagonals outside the main block diagonal.

Property 3: In Fig. 2(c) and (d), the norms of the subblocks $\{L_{l_1l_2}^\infty\}$ and $\{F_{l_1l_2}^\infty\}$ along a block row l_1 are plotted. Interestingly, we note that the subblocks constituting the Cholesky blocks $\{L^\infty, F^\infty\}$ converge to $\underline{0}$ along block row l_1 on the respective nonzero side of the main diagonal. This illustrates that a block banded approximation to the Cholesky blocks $\{L^\infty, F^\infty\}$ is reasonable.

Based on observations 1–3, we approximate the Cholesky blocks $F^{(k)}$ in the Cholesky factor \mathcal{L} by M_1 block banded upper triangular matrices [21] as follows

$$F^{(k)} = \begin{bmatrix} \underline{0} & & & \\ & F_{l_1l_2}^{(k)} & & \\ & & \ddots & \\ & & & \underline{0} \end{bmatrix}, \quad (19)$$

for $0 \leq (l_2 - l_1) \leq M_1$. Similarly, the Cholesky blocks $L^{(k)}$ in the Cholesky factor \mathcal{L} are approximated by M_2 block banded lower triangular matrices as follows

$$L^{(k)} = \begin{bmatrix} & & & \underline{0} \\ & L_{l_1l_2}^{(k)} & & \\ & & \ddots & \\ \underline{0} & & & \end{bmatrix}, \quad (20)$$

for $0 \leq (l_1 - l_2) \leq M_2$. Note that the upper triangular structure of $F^{(k)}$ and the lower triangular structure of $L^{(k)}$ follow directly from the forward unilateral representation, (8)–(9), and are not approximations. The only approximation made in (19)–(20) is to impose a block banded structure on the nonzero triangular portion of the Cholesky blocks. When coupled with the one sided forward representation of (8)–(9), the block banded approximations considerably simplify the prediction model. The simplified model is considered next.

B. Subblock Implementation

To derive the subblock SNP/VQR, we expand (8) and (9) at the subblock level with $L^{(k)}$ and $F^{(k)}$ approximated by the M -block banded approximations given in (19) and (20). Here we present the resulting expressions for $M_1 = M_2 = 3$.

Frame ($k = 1$): $\forall (1 \leq l_1 \leq N_I)$,

$$\sum_{\tau=\max(1, l_1-3)}^{l_1} L_{l_1\tau}^{(1)} X_\tau^{(1)} = v_{l_1}^{(1)}. \quad (21)$$

Frame ($2 \leq k \leq N_K$): $\forall (1 \leq l_1 \leq N_I)$,

$$\sum_{\tau=l_1}^{\min(l_1+3, N_I)} F_{l_1\tau}^{(k)} X_\tau^{(k-1)} + \sum_{\tau=\max(1, l_1-3)}^{l_1} L_{l_1\tau}^{(k)} X_\tau^{(k)} = v_{l_1}^{(k)}. \quad (22)$$

To compute the Cholesky subblocks $\{F_{l_1l_2}^{(k)}, L_{l_1l_2}^{(k)}\}$, (14) and (15) are expanded in terms of the block banded structure defined in (19) and (20). For $M_1 = M_2 = 3$, the simplified expressions are given by (23)–(26), where term $\bar{\delta}_{l_1N_I}$ equals 1 if $l_1 \neq N_I$. Otherwise, $\bar{\delta}_{l_1N_I} = 0$. Since the Cholesky subblocks $\{L_{l_1l_2}^{(k)}, F_{l_1l_2}^{(k)}\}$ in (23)–(26) converge to a steady state, therefore, only a limited number of these subblocks are computed. The steps involved in computing the steady state values are as follows.

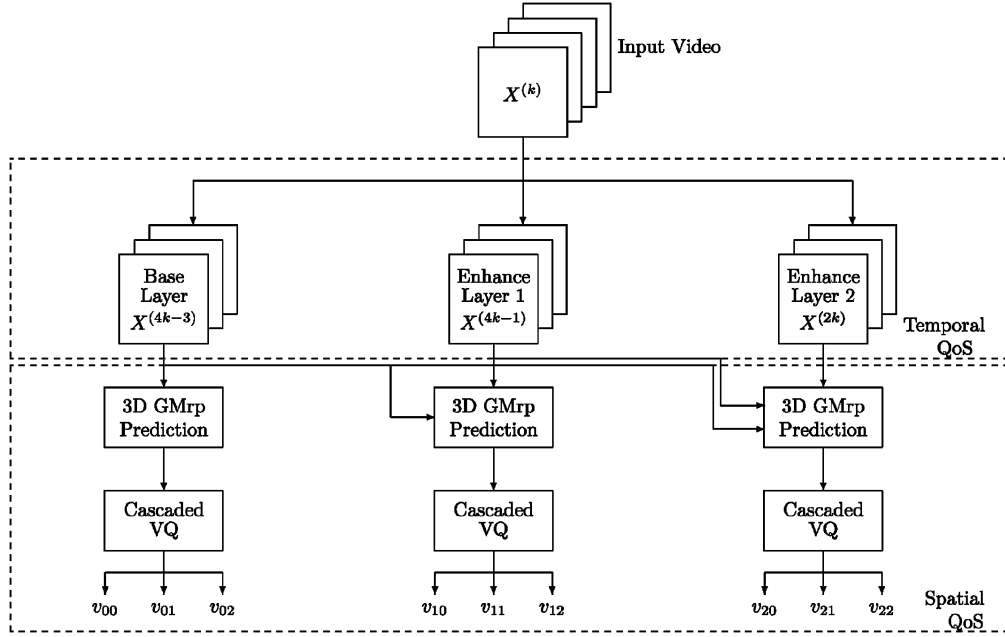


Fig. 5. Block diagram representation for SNP/VQR video codec offering different quality of services (QoS). A subscriber accessing the video using the Bronze service receives only one stream v_{00} . A subscriber using the Silver service receives streams (v_{00}, v_{01}) and (v_{10}, v_{11}) . A subscriber using the Gold service receives all the streams. Other QoS are also possible using different combinations of the temporal and spatial feeds.

of spatial QoS. For the best spatial quality video, the outputs of all three stages of VQ are transmitted to the receiver. For intermediate quality, the outputs of the first two stages are used by the receiver. Finally, for the lowest spatial quality, the output of the first stage is transmitted to the receiver.

Determination of the optimal distribution of the number of code vectors between different stages of cascaded VQ is a computationally intensive problem [22]. The computational complexity of such a search exceeds that of an exhaustive search. To determine a good distribution, we ran an experiment for a 3-stage, 6-bit vector quantizer. In Fig. 3, we include a selective subset of the possible distributions and plot the PSNRs of the video sequences reconstructed from a 3-stage VQ. We observe that the performance of the “321”-bit is closest to the performance of a single stage VQ. In the subblock implementation of SNP/VQR, we use a 3-stage, 6-bit vector quantizer with a “321”-bit distribution between the three stages.

Temporal Scalability: is achieved by representing the video sequence $X^{(k)}$, ($1 \leq k \leq N_K$), in three layers. Fig. 4 shows a possible configuration. The base layer consists of every fourth frame $X^{(4k-3)}$ and is encoded with the subblock SNP/VQR using a lower frame rate. The first enhancement layer encodes frames $X^{(4k-1)}$ and uses the frames reconstructed from the base layer during the prediction step. The difference between the actual frames $X^{(4k-1)}$ and the corresponding predicted frames is computed and quantized using a 3-stage “321”-VQ enhanced with conditional replenishment. Finally, the second enhancement layer encodes the remaining frames $X^{(2k)}$ with its prediction based on the frames reconstructed from the first enhancement and base layers.

Choice of Service: We propose three quality of services (QoS) (Gold, Silver, and Bronze) in the subblock SNP/VQR by combining the spatial and temporal feeds discussed earlier. The

procedure is illustrated in Fig. 5. The error field obtained from the three temporal layers is encoded using a 3-stage, 6-bit VQ with a “321” distribution between the three stages of the VQ. Other QoS are possible by adjusting the number of stages and code vectors at each stage in the cascaded VQ and by using a different temporal decimation scheme.

Bronze Service: uses the base layer $X^{(4k-3)}$ of the temporal feed compressed with the first stage of the 321-cascaded VQ. The frame rate is, therefore, one-fourth of the original video. Using a (4×4) block size in VQ, the compression ratio in the spatial domain is $((4 \times 4 \times 8)/3)$ or 43, leading to an overall compression ratio of $(43 \times 4) = 172$. Additional compression is achieved by conditional replenishment, not included in the above calculation for computing the compression ratio. For a video sequence with QCIF resolution (144×176) and a frame rate of 30 fps, a compression ratio of 172 corresponds to a transmission rate of about 35 Kbps. The bronze service is suitable for dial-up networks using V.90 modems that support a channel capacity of 56 Kbps.

Silver Service: uses the first enhancement layer $X^{(4k-1)}$ and the base layer $X^{(4k-3)}$ of the temporal feed compressed with the first two stages of the 321-cascaded VQ. The frame rate is, therefore, one-half of the streaming video with the spatial compression given by $((4 \times 4 \times 8)/5)$ or 25.4. Compared to the original sequence, the compression ratio of the silver service is $(25.4 \times 2) = 50.8$. For the QCIF resolution, this implies a transmission rate of 120 Kbps. Additional compression can be achieved using the conditional replenishment during VQ.

Gold Service: uses all three layers of the temporal feed compressed with a 3-stage 321-cascaded VQ. The frame rate is the same as that of the input video. The spatial compression ratio is $((4 \times 4 \times 8)/6)$ or 21.3, leading to a transmission rate of 286 Kbps for a video sequence with QCIF resolution.

TABLE I
DISTRIBUTION OF CODE VECTORS BETWEEN THE THREE LAYERS OF SNP/VQR WITH A 3-STAGE, 321-CASCADED VQ

Layer	Number of Vectors in Cascaded VQ							Replenishment overhead in bits	
	Total	Dropped			Transmitted			Raw	Comp.
		Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3		
Base	31,680	24,192	24,515	—	7,488	7,165	31,680	63,360	14,096
Enhancement # 1	31,680	24,136	24,459	—	7,544	7,221	31,680	63,360	14,213
Enhancement # 2	63,360	48,612	47,337	—	14,748	16,023	63,360	126,720	28,283
Total	126,720	92,672	96,311	—	29,780	30,409	126,720	255,440	56,592

TABLE II
COMPARISON OF PSNR FOR TEST VIDEO SEQUENCES RECONSTRUCTED AT 30 FRAMES PER SECOND WITH MPEG4 AND THE BLOCK SNP/VQR AT DIFFERENT BPS REPRESENTATIONS. SOME PSNR VALUES FOR MPEG4 ARE MISSING BECAUSE OF THE INABILITY OF THE MPEG4 IMPLEMENTATION [24] TO ACHIEVE THE CORRESPONDING TRANSMISSION RATES WITHOUT REDUCING THE FRAME RATE

Foreman			Highway			News		
Rate (kbps)	PSNR (dB)		Rate (kbps)	PSNR (dB)		Rate (kbps)	PSNR (dB)	
	MPEG4	SNP/VQR		MPEG4	SNP/VQR		MPEG4	SNP/VQR
160	32.61	32.75	148	35.94	35.95	191	36.21	36.25
125	31.60	31.84	100	35.43	35.48	119	32.95	33.68
73	28.64	30.32	59	33.43	34.06	75	30.11	32.57
50	26.76	28.98	27	29.81	31.66	44	26.14	29.68
36	—	27.79	20	—	29.50	33	—	28.29

V. EXPERIMENTS

The experiments presented here are designed to make two major points. First, we show that reasonably good quality is obtained at low bit rates using SNP/VQR and these results are superior, in terms of visual quality, to those obtained at similar bit rates using the ITU standard H.263 and the ISO standard MPEG4. Since we are interested in comparing the overall performance, we use the Gold service of SNP/VQR in our comparison with the two standards. The H.263 and MPEG4 encoders have many optional features and their performances vary from one implementation to another depending on how many of the available features are selected. For H.263, we use the baseline implementation available at [23] that incorporates half-pixel motion compensation, 3-D variable length coding of DCT coefficients, and coding of overhead information such as macroblock control data and coded block patterns. Optional features like unrestricted motion vectors, syntax based arithmetic coding, and advanced prediction mode are not implemented. The MPEG4 codec is downloaded from [24] and is also a baseline implementation. In addition to the perceived quality, we use peak signal to reconstructed noise ratio (PSNR) as the quantitative measure of video quality to compare the performance of the three codecs.

In the second set of experiments, we seek to compare different quality of services (Gold, Silver, and Bronze) described in Fig. 5. The reconstructed sequences are compressed both temporally and spatially in the second set of experiments. We illustrate how much improvement is possible in the perceived quality when a client moves to a higher class of service from a lower class. In our simulations, we use four test sequences: *carphone*, *foreman*, *highway*, and *news* that have a QCIF resolution of (144×176) pixels per frame with a display rate of 30 frames/s. *Carphone* and *foreman* are typical “talking head” sequences with limited movement, while *news* and *highway* have relatively

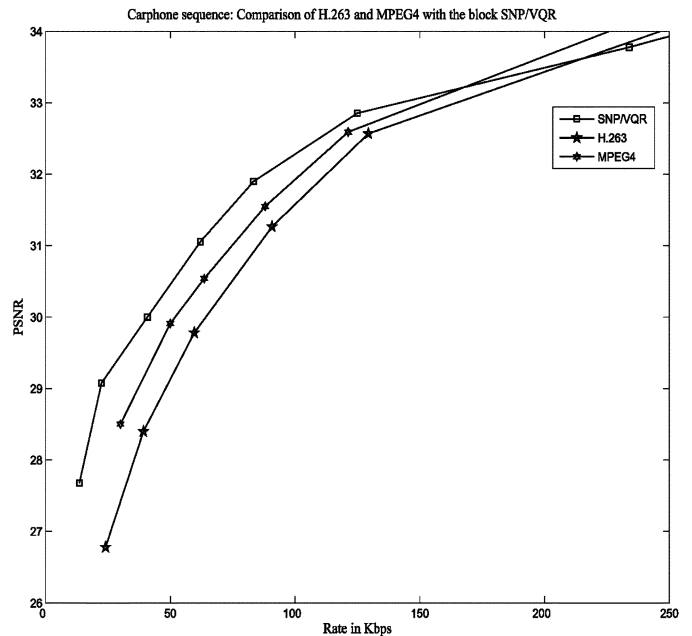


Fig. 6. Comparison of the mean PSNR for the texticarphone sequence compressed using H.263, MPEG4, and SNP/VQR. In these results, SNP/VQR is configured at the Gold service with $\beta_h = 0.166309$, $\beta_v = 0.156631$, and $\beta_t = 0.167446$.

higher background motion. We are primarily motivated with video conferencing or surveillance applications over a wireless channel. In such applications, the camera motion is limited with the effective transmission rate constrained to below 200 kbps. The test sequences are good representatives of the applications under consideration. Extension of SNP/VQR to video sequences with abrupt scene changes requires recalculation of the model interaction parameters $\{\beta_v, \beta_h, \beta_t\}$. We propose two strategies for dealing with such situations. Strategy 1 involves a passive



Fig. 7. Frame 21 extracted from the carphone sequence compressed to different bit rates using H.263, MPEG4, and the block SNP/VQR. Frame (a) is the original frame. The frames in the left most column are compressed using H.263, the frames in the middle column are compressed using MPEG4, while the frames in the right most column are compressed using the block SNP/VQR. The bit rate for frames (b), (c), and (d) is 90 Kbps ($CR = 67$), the bit rate for frames (e), (f), and (g) is 39 Kbps ($CR = 155$), and the bit rate for frames (h), (i), and (j) is 24 Kbps ($CR = 253$).

technique of refreshing the model interaction parameters after every consecutive N frames in the video sequence. This strategy is similar to the one employed in the standard codecs where the reference frame (I-picture) is refreshed periodically. Strategy 2 determines if the scene has changed in the video sequence. Only when the scene changes substantially, the interaction parameters are recalculated.

Before continuing on with the comparison, we explain how we compute the transmitted bit rate for SNP/VQR. As an example, we choose to compress the first 80 frames of the 8-bit monochrome carphone sequence with SNP/VQR configured for the Gold service using a 6-bit 3-stage VQ with a '321'-bit distribution between the three stages. The total number of bits in the carphone sequence is given by $(144 \times 176) \times 80 \times 8$, or roughly

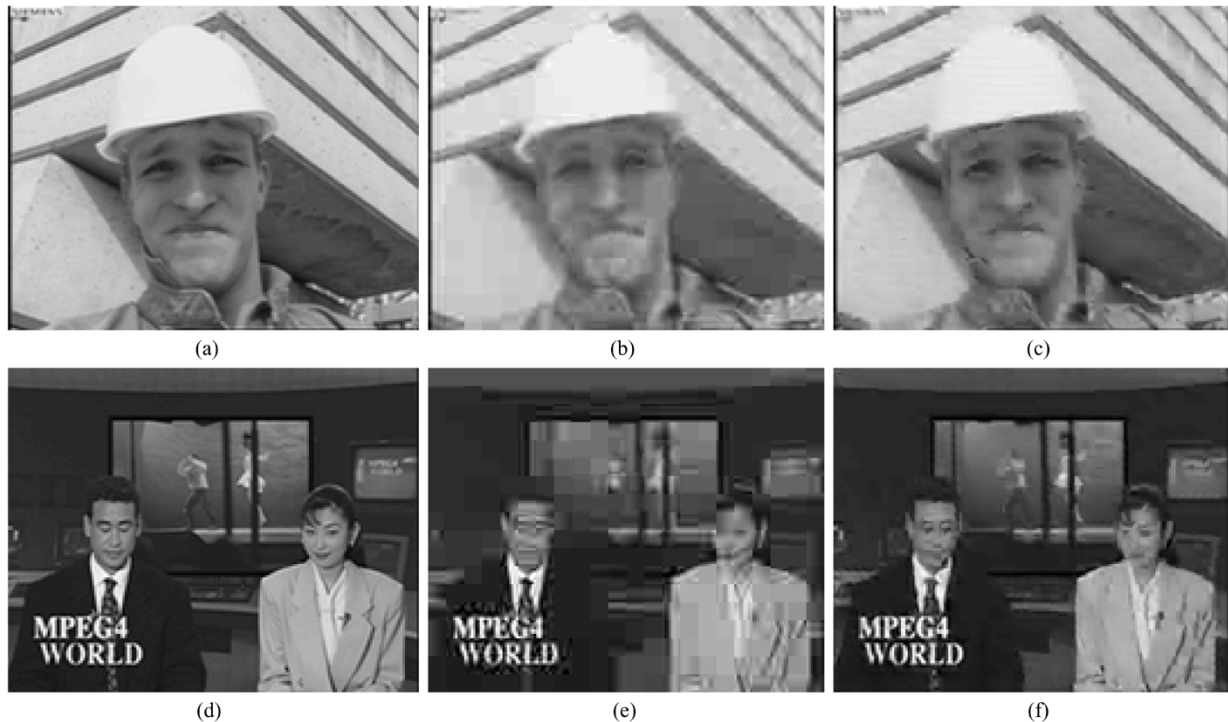


Fig. 8. Comparison of the block SNP/VQR with MPEG4 for the foreman and news sequences. (a) Original frame no. 41 of the foreman sequence. (b) Frame no. 41 compressed to 72 kbps using MPEG4. (c) Same as (b) except the frame is compressed with SNP/VQR. (d) Original frame no. 41 of the news sequence. (e) Frame no. 41 compressed to 41 kbps using MPEG4. (f) Same as (e) except the frame is compressed with SNP/VQR. In (a)–(d), the frames are compressed spatially. The frame rate of the compressed sequences is 30 fps.

16.22 Mbits, leading to a transmission rate of 6.08 Mbps at 30 frames per second. Using a (4×4) block in cascaded VQ, the total number of vectors from the test sequence is 126 760. Not all of these vectors are transmitted at each stage of cascaded VQ. A significant number of vectors are dropped due to conditional replenishment, which leads to additional compression. Conditional replenishment, however, introduces some overhead. Each stage of VQ requires one additional bit to convey the status of the quantized vector to the receiver, i.e., to code whether the vector is being transmitted or not. The amount of this overhead in bits is given by the number of input vectors at each stage of cascaded VQ. This overhead is reduced by using an entropy coding scheme like Lempel–Ziv algorithm. It may be noted that the entropy coding scheme is only applied to the overhead resulting from conditional replenishment. The data from VQ is transmitted without any additional coding. This is in accordance with H.263 (and MPEG4) where the discrete cosine transform (DCT) coefficients in H.263 (or the wavelet coefficients in MPEG4) are quantized and entropy coded before transmission.

Table I shows the number of vectors dropped due to conditional replenishment at the first two stages of the 3-stage cascaded VQ. We do not apply conditional replenishment at the third stage of cascaded VQ. Since the code book at the third stage is of size 2, 1 bit is used to represent each code vector at the third stage. The savings obtained from conditional replenishment at the third stage is, therefore, compensated by the associated overhead. Based on Table I, the total number of bits used to represent the video sequence is given by

$$(29,780 \times 3 + 30,409 \times 2 + 126,720 \times 1 + 56,592)$$

or, 333 470 bits. The compression ratio (CR) for SNP/VQR is therefore given by $16.22 \times 10^6 / 333,470 = 48.64$, which provides an average bit rate of 125 Kbps.

Fig. 6 plots the mean PSNR computed for the carphone sequence after its encoding with the H.263, MPEG4, and SNP/VQR codecs as a function of the transmission rate. In SNP/VQR, we vary the total number of bits in cascaded VQ such that the resulting bit rates range from 25 Kbps to 250 Kbps. The codebooks used at different bit rates are universal in the sense that these are generated from a set of training sequences that do not include the four test sequences used in the comparative study. Fig. 6 illustrates that SNP/VQR produces sequences with higher PSNR values than H.263 for bit rates below 200 Kbps and MPEG4 for bit rates below 175 kbps. At 125 Kbps, for example, the video sequence compressed with SNP/VQR has a PSNR value that is roughly 0.5 dB higher than the video sequence obtained at the same bit rate from H.263, while the improvement over MPEG4 is given by 0.25 dB. At a transmission rate of around 50 Kbps, the block SNP/VQR provides improvements of about 1.5 dB over H.263 and 0.75 dB over MPEG4. Table II provides additional PSNR comparisons between the block SNP/VQR and MPEG4 for three more test sequences. Our earlier observations are validated with the results listed in Table II.

A higher PSNR does not necessarily imply a superior quality reconstructed video, because the perceived video quality is highly dependent on the human visual system (HVS). To provide subjective evaluation of the sequences, representative frames are extracted from the sequences compressed with H.263, MPEG4, and SNP/VQR. Fig. 7 illustrates the perceived



Fig. 9. Comparison of different quality of services. (a) Original frame no. 60 and (b)–(d) frame 60 obtained as part of the video sequence compressed with (b) the Gold service to a bit rate of 125 Kbps ($CR = 48.6$). (c) the Silver service at a bit rate of 31 Kbps ($CR = 196$). (d) the Bronze service at a bit rate of 7.75 Kbps ($CR = 784$). In Silver and Bronze services, gains due to temporal compression are included in the compression ratios.

differences between frame 21 of the carphone sequence reconstructed using H.263, MPEG4, and SNP/VQR at three different bit rates. In Fig. 7, the frames compressed with SNP/VQR exhibit better visual quality with more details retained, e.g., the structure of the eyes and eyebrows, and the tower seen through the car's window. Moreover, there is little blocking visible in the frames compressed with SNP/VQR despite the fact that VQ is prone to introducing blocking at low bit rates. Similar observations are made for the other test sequences compressed with MPEG4 and SNP/VQR as highlighted in Fig. 8.

In the second set of experiments, we illustrate the improvement in the visual quality obtained by switching SNP/VQR from a lower QoS to a higher service. Fig. 9 shows frame 60 from the carphone sequence compressed using the Gold, Silver, and Bronze services. Not depicted in the frames is the difference in the frame rates offered with each service. The Gold service uses the original frame rate of 30 fps while the Silver service displays every second frame and the Bronze service displays every fourth frame in the sequence. Hence, the frame rate for Silver service is 15 fps while the frame rate for Bronze service is 7.5 fps. Taking the spatial compression achieved with the 321-cascaded VQ in consideration, the overall compression ratio for the Gold service is 48.6 compared to the compression ratios of 196 and 784 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 offers better visual quality than the frame reconstructed from H.263 and MPEG4 at the same compression ratios.

VI. SUMMARY

This paper presents a noncausal predictive codec SNP/VQR, which couples 3-D noncausal prediction with conditional replenishment extended cascaded VQ. Extension of 2-D noncausal prediction to the compression of 3-D video sequences is challenging because of the high computational complexity of SNP/VQR. We present a practical implementation of the SNP/VQR, which, in comparison with the direct SNP/VQR, provides computational savings of two orders of magnitude of the linear dimension of the video. The computational complexity of the block SNP/VQR is comparable to that of the standard codecs. In our simulations, SNP/VQR compares favorably with the ITU H.263 and ISO MPEG4 video compression standards especially at low bit rates. SNP/VQR is also capable of offering different quality of services (QoS) both in the temporal and spatial domains. The paper considers three QoS, referred to as Gold, Bronze and Silver services. The offered services are hierarchical such that any higher QoS can be derived from a lower QoS by transmitting additional enhancement layers. This feature is especially useful for delivering live streaming video over heterogeneous multicast networks.

REFERENCES

- [1] N. Shacham, "Multipoint communication by hierarchically encoded data," in *Proc. IEEE Infocom*, vol. 3, Florence, Italy, May 1992, pp. 2107–2114.
- [2] S. Deering, "Internet multicasting routing: State of the art and open research questions," in *Multimedia Integrated Conferencing for Europe*, Oct. 1993.

- [3] S. McCanne, M. Vetterli, and V. Jacobson, "Low-complexity video coding for receiver driven layered multicast," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 983–1001, Aug. 1997.
- [4] M. van der Schaar and H. Radha, "A hybrid temporal-SNR fine-granular scalability for internet video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 318–331, Mar. 2001.
- [5] I. Sodagar, H.-J. Lee, P. Hatrak, and Y.-Q. Zhang, "Scalable wavelet coding for synthetic/natural hybrid images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 244–254, Mar. 1999.
- [6] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 332–344, Mar. 2001.
- [7] M. Kouras and A. Asif, "Noncausal predictive video codec offering hierarchical QoS," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada, May 17–21, 2004.
- [8] W. Tan and A. Zakhor, "Video multicast using layered FEC and scalable compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 373–386, Mar. 2001.
- [9] L. Vicisano, L. Rizzo, and J. Crowcroft, "TCP-like congestion control for layered multicast data transfer," in *Proc. IEEE Infocom '98*, vol. 3, San Francisco, CA, Mar. 1998, pp. 996–1003.
- [10] X. Li, S. Paul, and M. Ammar, "Layered video multicast with retransmissions (LVMR): Evaluation of hierarchical rate control," in *Proc. IEEE Infocom '98*, vol. 3, San Francisco, CA, Mar. 1998, pp. 1062–1072.
- [11] A. Asif and J. M. Moura, "Image codec by noncausal prediction, residual mean removal, and cascaded VQ," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 1, pp. 42–55, Feb. 1996.
- [12] A. Asif, "Fast rauch-tung-striebl smoother based image restoration for noncausal images," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 371–375, Mar. 2004.
- [13] S. M. Schweizer and J. M. F. Moura, "Hyperspectral imagery: Clutter adaptation in anomaly detection," *IEEE Trans. Inform. Theory*, vol. 46, no. 8, pp. 1855–1871, Aug. 2000.
- [14] M. Goldberg and H. Sun, "Image sequence coding using vector quantization," *IEEE Trans. Commun.*, vol. COM-34, pp. 792–800, 1986.
- [15] T. J. Woods, "Two dimensional discrete markovian fields," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 2, pp. 232–240, Mar. 1972.
- [16] J. M. F. Moura and N. Balram, "Recursive structure of noncausal Gauss-Markov random fields," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 334–354, Mar. 1992.
- [17] "ITU-T Recommendation H.263 V.2: Video Coding for Low Bit Rate Communication," ITU Telecommunication Standardization Sector of ITU, 1998.
- [18] *ISO/IEC. IS 14496-1: Information Technology—Coding of Audio-Visual Objects*, 1999.
- [19] H. Andrew and B. Hunt, *Digital Image Restoration*. Englewood Cliffs, NJ: Prentice-Hall, 1977, pp. 211–220.
- [20] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 232–240, Jan. 1980.
- [21] A. Asif and J. M. F. Moura, "Block matrices with L-block banded inverse: Inversion algorithms," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 630–642, Feb. 2005.
- [22] C. F. Barnes and R. L. Frost, "Vector quantizers with direct sum codebooks," *IEEE Trans. Inform. Theory*, vol. 39, no. 2, pp. 565–580, Mar. 1993.
- [23] T. Tanakitparpa. H.263 Video Codec. [Online]. Available: <http://www.angelfire.com/in/H261/h263capture.html>
- [24] FFmpeg Multimedia System. [Online]. Available: <http://ffmpeg.sourceforge.net/index.php>



Amir Asif (M'97–SM'02) received the B.S. degree in electrical engineering with the highest honors and distinction in 1990 from the University of Engineering and Technology, Lahore, Pakistan, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University (CMU), Pittsburgh, PA in 1993 and 1996, respectively.

He has been an Associate Professor of computer science and engineering at York University, North York, ON, Canada, since 2002. Prior to this, he was on the faculty of CMU, where he was a Research Engineer from 1997 to 1999 and the Technical University of British Columbia, Vancouver, BC, Canada, where he was an Assistant/Associate Professor from 1999 to 2002. His research interests lie in statistical signal processing (one and multidimensional), image and video processing, multimedia communications, and genomic signal processing. He has authored over 35 technical contributions and a textbook, including invited ones, published in international journals and conference proceedings.

Dr. Asif has been a Technical Associate Editor for the IEEE SIGNAL PROCESSING LETTERS since 2002. He has organized two IEEE conferences on signal processing theory and applications and served on the technical committees of several international conferences. He has received several distinguishing awards including the 2004 Excellence in Teaching Award from the Faculty of Science and Engineering and the 2003 Mildred Baptist teaching excellence award from the Department of Computer Science and Engineering. He is a member of the Professional Engineering Society of Ontario.



Maria Kouras was born in Moscow, Russia, in 1983 and emigrated to Canada in 1997. She received the B.Sc. degree (with honors) in computer science from York University, Toronto, ON, Canada, in 2004. She is planning to pursue the M.S. degree in 2006.

She is currently with the Ministry of Health and Long-Term Care of Ontario, Toronto. Among her research interests are image and video processing and multimedia communications.

Ms. Kouras was an Natural Science and Engineering Research Council (NSERC) summer scholarship recipient for 2003.