Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 54342, 15 pages
doi:10.1155/2007/54342

Research Article

Scalable Video Coding with Interlayer Signal Decorrelation Techniques

Wenxian Yang, Gagan Rath, and Christine Guillemot

Institut de Recherche en Informatique et Systèmes Aléatoires, Institut National de Recherche en Informatique et en Automatique, 35042 Rennes Cedex, France

Received 12 September 2006; Accepted 20 February 2007

Recommended by Chia-Wen Lin

Scalability is one of the essential requirements in the compression of visual data for present-day multimedia communications and storage. The basic building block for providing the spatial scalability in the scalable video coding (SVC) standard is the well-known Laplacian pyramid (LP). An LP achieves the multiscale representation of the video as a base-layer signal at lower resolution together with several enhancement-layer signals at successive higher resolutions. In this paper, we propose to improve the coding performance of the enhancement layers through efficient interlayer decorrelation techniques. We first show that, with nonbiorthogonal upsampling and downsampling filters, the base layer and the enhancement layers are correlated. We investigate two structures to reduce this correlation. The first structure updates the base-layer signal by subtracting from it the low-frequency component of the enhancement layer signal. The second structure modifies the prediction in order that the low-frequency component in the new enhancement layer is diminished. The second structure is integrated in the JSVM 4.0 codec with suitable modifications in the prediction modes. Experimental results with some standard test sequences demonstrate coding gains up to 1 dB for I pictures and up to 0.7 dB for both I and P pictures.

1. INTRODUCTION

Scalable video coding (SVC) is currently being developed as an extension of the ITU-T Recommendation H.264 |ISO/IEC International Standard ISO/IEC 14496-10 advanced video [1]. It allows to adapt the bit rate of the transmitted stream to the network bandwidth, and/or the resolution of the transmitted stream to the resolution or rendering capability of the receiving device. In the current SVC reference software JSVM, spatial scalability is achieved using layers with different spatial resolutions. The higher-resolution signals, commonly known as enhancement layers, are represented as difference signals where the differencing is performed between the original high-resolution signals and predictions on a macroblock level. These predictions can be spatial (intraframe), temporal, or interlayer. The lower-base layer signal along with the associated interlayer-predicted enhancement layer signal constitutes the well-known Laplacian pyramid (LP) representation [2].

The Laplacian pyramid represents an image as an hierarchy of differential images of increasing resolution such that each level corresponds to a different band of image frequencies. The pyramid is generated from a Gaussian pyramid by taking the differences between its higher-resolution layers and the interpolations of the next lower-resolution layers. The difference layers, called detail signals, have typically much less entropy than the corresponding Gaussian pyramid layers. As a result, an LP requires much less bit rate than the associated Gaussian pyramid when encoded for transmission or storage. At the receiver, the decoder reconstructs the original signal by successively interpolating the lower-resolution signal and adding the detail layers up to the desired resolution.

In the context of scalable video coding, the LP structure can be more complex. The current SVC standard defines a three-layer scalable video structure (SD, CIF, and QCIF) where each layer has quarter the resolution of its upper layer. The standard defines an input video sequence as groups of pictures (GOPs) where each group contains one Intra (I) frame and may contain several forwardly predicted (P) and bidirectionally predicted (B) frames. The prediction in P and B frames can occur at the slice level, and the corresponding slices are known as predictive and bipredictive slices, respectively. The incorporation of motion compensation on each

layer renders the LP structure to represent either original signals or motion compensated residual signals at higher layers. For example, the I frames in upper resolution layers can have interlayer predictions (LP) applied to the original signal; the P and B frames, however, can have interlayer predictions (LP) applied to the motion compensated residual signals.

In the context of scalable video coding, the compression of the enhancement layers is an important issue. In the SVC standard, for the enhancement layer blocks coded with interlayer predictions, the decoder follows the standard LP reconstruction, that is, it interpolates the base layer and adds the enhancement layer to the interpolated signal. Do and Vetterli [3] have proposed to use a dual-frame-based reconstruction which has a better rate-distortion (R-D) performance. The dual-frame construction, however, requires biorthogonal upsampling and downsampling filters, which limits its application in SVC because of noticeable aliasing in lower-resolution layers. To improve upon this drawback, the authors in [4, 5] have proposed to add an update step for the base-layer signal at the LP encoder. This structure, however, necessitates not only an open loop LP structure but also the design of a new lowpass filter.

An alternative approach to improve the compression efficiency of enhancement layers is to employ better interlayer predictions. To that end, several techniques have already been proposed to the JVT [6–8]. In [6], optimal upsamplers are designed which depend on the downsampling filter, the quantization levels of the base layer, and the input video sequence. Later, a family of downsamplers is constructed to span a range of filter lengths, aliasing, and ringing characteristics available to an encoder [7], together with their corresponding upsamplers. In [8], the direction information of the base layer is used to improve the prediction for the macroblocks (MBs) with high-directional characteristics.

In this paper, we propose to improve the coding performance of the enhancement layers through efficient interlayer decorrelation techniques. We first show that, with nonbiorthogonal upsampling and downsampling filters, the base layer and the enhancement layers are correlated. We investigate two structures to reduce this correlation. The first structure updates the base-layer signal by subtracting from it the low-frequency component of the enhancement layer signal. The second structure modifies the prediction in order that the low-frequency component in the new enhancement layer is diminished. We present these structures both in the open-loop and in the closed-loop configurations. We analyze the reconstruction errors with both structures under reasonable assumptions regarding the statistical properties of the different quantization noises, and show that the second structure in the closed-loop configuration leads to an error that is dependent only on the quantization error of the enhancement layer. To improve the coding efficiency of the enhancement layer further, we use a recently proposed orthogonal transform in conjunction with the existing 4×4 transform. We incorporate the proposed prediction method in the JSVM software and present the results with respect to a current implementation.

The rest of the paper is organized as follows. In Section 2, we present a brief description of the classical Laplacian pyramid. Section 3 reviews the LP reconstruction structure and some of its recent improvements. Sections 4 and 5 describe the proposed decorrelation methods that result in either a reduced coarse signal or a reduced detail signal. In Section 6, we analyze the reconstruction errors that ensue from different decoding techniques. Section 7 touches upon the subject of transform coding of enhancement layers. Sections 8 and 9 present the details of the integration of the proposed method in the JSVM codec with necessary mode selection options and the results obtained with some standard test sequences. Finally, we draw conclusions alongside some future research perspectives in Section 10.

2. LAPLACIAN PYRAMID REPRESENTATION

The LP structure proposed by Burt and Adelson [2] is shown in Figure 1. For convenience of notation, let us consider an LP for 1D signals; the results can be carried over to the higher dimensions in a straightforward manner with separable filters. For an image, for example, the filtering operations can be performed first row-wise and then column-wise, each operation using 1D signals. For the sake of explanation, we will here consider an LP with only one level of decomposition. For multiple levels of decompositions, the results can be derived by repeating the operations on the lower-resolution layer. Considering an input signal $\mathbf{x}$ of N samples and dyadic downsampling, a coarse signal $\mathbf{c}$ can be derived as[1]

$$\mathbf{c} := H\mathbf{x}, \tag{1}$$

where H denotes the decimation filter matrix of dimension $(N/2) \times N$. H has the following general structure[2]:

$$H := \begin{bmatrix} \ddots & & & & & & & & & \\ \dots & 0 & h(L) & h(L-1) & \dots & h(1) & h(0) & 0 & 0 & \dots \\ \dots & \dots & 0 & 0 & h(L) & \dots & h(2) & h(1) & h(0) & \dots \\ & & & & & & & & & \ddots \end{bmatrix}. \tag{2}$$

The coefficients $h(n)$, $n = 0, 1, 2, \dots, L$, here denote the downsampling filter coefficients. The matrix structure above is a result of the filtering (i.e., convolution) and downsampling the filtered output by factor 2 (the elements of a row are right-shifted by 2 columns from the elements of the previous row). We assume an FIR filter having linear phase (i.e., symmetric). Repeated filtering and downsampling operations on the coarsest signal leads to the so-called Gaussian pyramid. The first level of LP is obtained by predicting the signal $\mathbf{x}$ based on the coarse signal $\mathbf{c}$. The prediction is made by upsampling the coarse signal with alternate zero

[1] We use the notation ":=" for "is derived as" or "is defined as."

[2] For a finite signal, because of the symmetric extension at the boundary, the columns of H and G matrices at the left and at the right are flipped.

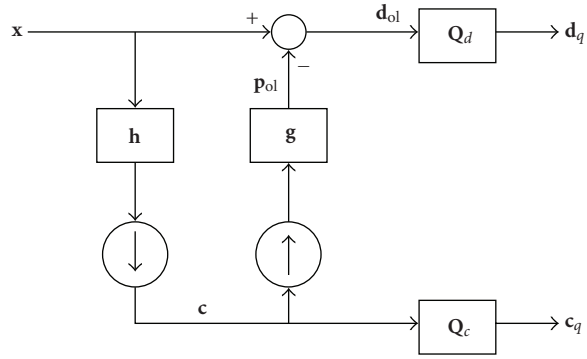


Figure 1: Open-loop Laplacian pyramid structure with one decomposition level.

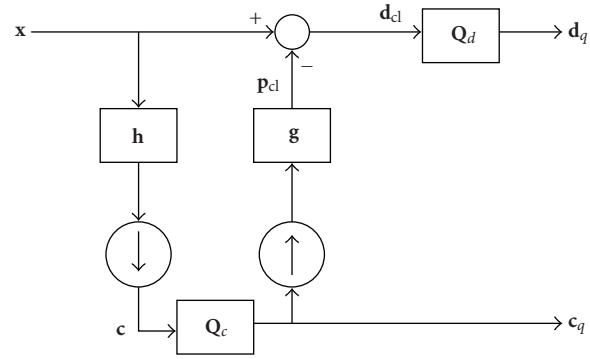


Figure 2: Closed-loop Laplacian pyramid structure with one decomposition level.

samples and then filtering the upsampled signal. In the SVC framework, the LP coefficients need to be quantized before being encoded. Depending on whether the quantizer for the low-resolution signal is inside or outside the prediction loop, there can be two different structures for the LP. The *open-loop* prediction structure with the quantizer outside the loop is shown in Figure 1. In this structure, the detail signal $\mathbf{d}_{\text{ol}}$ is given as

$$\mathbf{d}_{\text{ol}} := \mathbf{x} - G\mathbf{c} = (I_N - GH)\mathbf{x}, \tag{3}$$

where I_N denotes the identity matrix of order N and G denotes the interpolation filter matrix of dimension $N \times (N/2)$. G has the following general structure:

$$G := \begin{bmatrix} \ddots & & & & & & & & & \\ \cdots & 0 & g(0) & g(1) & g(2) & \cdots & g(M) & 0 & 0 & \cdots \\ \cdots & \cdots & 0 & 0 & g(0) & g(1) & \cdots & g(M-1) & g(M) & \cdots \\ & & & & & & & & & \ddots \end{bmatrix}^t. \tag{4}$$

The coefficients $g(n)$, $n = 0, 1, 2, \ldots, M$, here denote the upsampling filter coefficients and the superscript t denotes the matrix transpose operation. Like the decimation filter matrix, the interpolation filter matrix structure is a result of the upsampling by factor 2 and filtering. The down-shifting of a column by two rows from the previous column is due to the alternate zero elements in the upsampled signal. The filter is also assumed to be FIR and linear phase. Throughout the paper, we assume normalized downsampling and upsampling filters. That is,

$$\sum_n h(n) = 1, \qquad \sum_n g(n) = 2. \tag{5}$$

These normalization conditions guarantee that the coarse signals and the prediction signals have about the same dynamic range as the original signal.

The *closed-loop* configuration with the quantizer within the prediction loop is depicted in Figure 2. Here the quantized coarse signal is used to make the prediction for the higher-resolution signal. If $\mathbf{c}_q$ denotes the quantized low-resolution signal, the detail signal is obtained as

$$\mathbf{d}_{\text{cl}} := \mathbf{x} - G\mathbf{c}_q. \tag{6}$$

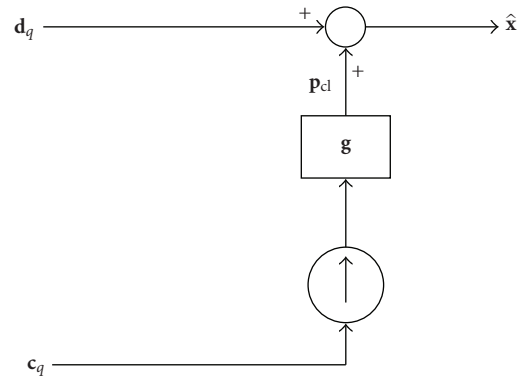


Figure 3: Standard reconstruction structure for LP.

Irrespective of the configuration, the coarse and the detail signals are encoded with suitable transforms and variable length coding (VLC) schemes before being transmitted to the decoder. In Figures 1 and 2, we use the same symbols $\mathbf{c}_q$ and $\mathbf{d}_q$ to denote the quantized coarse and detail signals. Clearly, given the same quantizers, both structures transmit the same coarse signal; however, the transmitted detail signals are different. We use the same symbol notation for the sake of simplicity and also because of the fact that their usual reconstruction structures (given in the next section) are identical. In JSVM, the closed-loop prediction structure is adopted because of its superior performance compared to the open-loop structure. Note that the coarse signal and the detail signals here refer, respectively, to the base layer and the interlayer-predicted enhancement layers in the JSVM.

3. LP DECODER STRUCTURES

The standard reconstruction method of an LP, either open-loop or closed-loop, is shown in Figure 3. First the decoded

coarse signal is upsampled and then filtered using the same interpolation filter as used at the encoder. This prediction signal is added to the decoded detail signal to estimate the original higher-resolution signal. Considering an LP with only one level of decomposition, we can reconstruct the original signal as

$$\hat{\mathbf{x}}_s := G\mathbf{c}_q + \mathbf{d}_q, \tag{7}$$

where $\mathbf{d}_q$ denotes the quantized or decoded detail signal. Observe that the prediction signal is identical to that for the closed-loop LP encoder (when there are no channel errors).

Because of its overcompleteness, an LP can be represented as a frame expansion as follows. Let K denote the resolution of the coarse signal $\mathbf{c}$. For dyadic downsampling, $K = N/2$. The coarse and the detail signals can be jointly expressed as

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{d}_{\text{ol}} \end{bmatrix} := \begin{bmatrix} H \\ I_N - GH \end{bmatrix} \mathbf{x} \equiv S\mathbf{x}, \tag{8}$$

where S denotes the matrix on the right-hand side having dimension $(N + K) \times N$. Since the LP is reversible for any combinations of the downsampling and upsampling filters $\mathbf{h}$ and $\mathbf{g}$, S has full-column rank. The rows of S constitute a frame and S can be called the frame operator or the analysis operator associated with the LP [3, 9, 10].

The usual reconstruction shown in (7) can be equivalently expressed using the reconstruction operator $[G \;\; I_N]$ as

$$\hat{\mathbf{x}}_s := \begin{bmatrix} G & I_N \end{bmatrix} \begin{bmatrix} \mathbf{c}_q \\ \mathbf{d}_q \end{bmatrix}. \tag{9}$$

It is trivial to prove that $[G \;\; I_N]S = I_N$. In [3], Do and Vetterli propose to reconstruct the original signal using the dual frame operator, which is $(S^tS)^{-1}S^t$. It can be shown that if the decimation and the interpolation filters are orthogonal, that is, $G^tG = HH^t = I_K$, $G = H^t$, the dual frame operator corresponding to the frame operator in (8) is $[G \;\; I_N - GH]$ [3]. If the filters are biorthogonal, that is, $HG = I_K$, the above reconstruction operator is still an inverse operator (i.e., it is a left-inverse of the analysis operator in (8)) even though it is not the dual-frame operator [3]. Therefore, with either orthogonal or biorthogonal filters, the original signal can be reconstructed as

$$\hat{\mathbf{x}}_f := \begin{bmatrix} G & I_N - GH \end{bmatrix} \begin{bmatrix} \mathbf{c}_q \\ \mathbf{d}_q \end{bmatrix} = G(\mathbf{c}_q - H\mathbf{d}_q) + \mathbf{d}_q. \tag{10}$$

The corresponding reconstruction structure is shown in Figure 4. It is easy to see that the above dual frame based reconstruction is identical to the standard reconstruction when the LP coefficients (both $\mathbf{c}$ and $\mathbf{d}_{\text{ol}}$) are not quantized.

The dual-frame-based reconstruction has the limitation that the decimation and the interpolation filters need to be at least biorthogonal. These filters, however, can lead to discernible and annoying aliasing in the coarse resolution signal. The authors in [11] try to alleviate this problem by proposing an update step at the encoder as shown in Figure 5(a). The

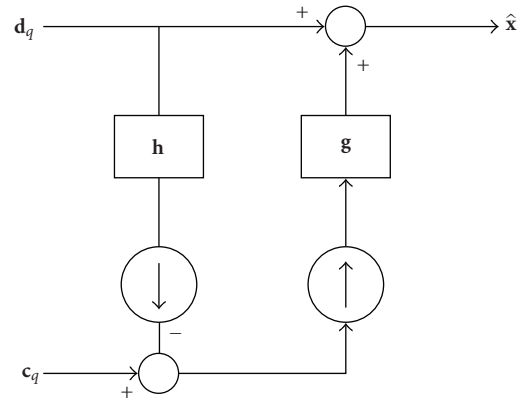


Figure 4: Frame-based reconstruction structure for LP.

detail signal $\mathbf{d}_{\text{ol}}$ undergoes a low-pass filtering and downsampling and this update signal is added to the coarse resolution signal $\mathbf{c}$ as follows:

$$\mathbf{c}_u := \mathbf{c} + F\mathbf{d}_{\text{ol}}, \tag{11}$$

where F denotes the update filter matrix. The update filter matrix has a similar structure as that of the decimation filter matrix except that the filter coefficients $h(n)$ are replaced by the update filter coefficients $f(n)$. The corresponding decoder, shown in Figure 5(b), has the same structure as the frame-based reconstruction in Figure 4 except that the decimation filter $\mathbf{h}$ is replaced by the update filter $\mathbf{f}$. Thus, the reconstructed signal is given as

$$\hat{\mathbf{x}}_u := G(\mathbf{c}_{uq} - F\mathbf{d}_q) + \mathbf{d}_q = G\mathbf{c}_{uq} + (I_N - GF)\mathbf{d}_q, \tag{12}$$

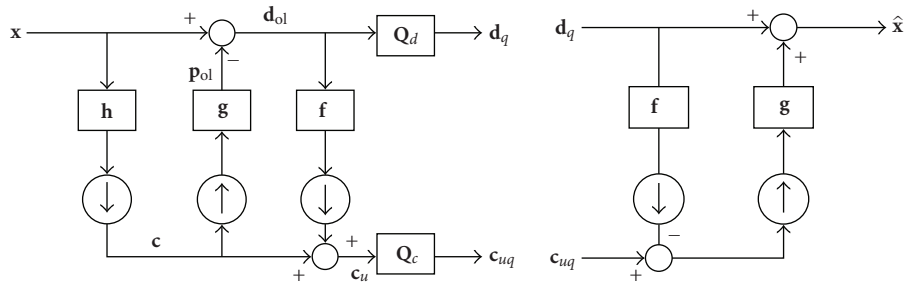
where $\mathbf{c}_{uq}$ denotes the quantized updated coarse signal.

Obviously, when the decimation and the update filters are identical, this reconstruction structure is the same as the dual frame-based reconstruction. This lifted pyramid is reversible for any set of filters $\mathbf{h}$, $\mathbf{g}$, and $\mathbf{f}$. In the special case, when the decimation and the update filters are identical and the decimation and the interpolation filters are biorthogonal, the update signal is equal to zero, and hence this improved pyramid is identical to the framed pyramid. Note that, like the framed pyramid, this improved pyramid is also open-loop. In the following, we propose some structures for LPs with nonbiorthogonal filters that are motivated from compression point of view. As we show below, they can be applied both in open-loop and closed-loop configurations.

4. IMPROVED OPEN-LOOP LP STRUCTURES

Consider first the open-loop configuration. When the upsampling and the downsampling filters are biorthogonal, $HG = I_K$ [3]. In this case, the detail signal obtained by the standard prediction does not contain any low-frequency component. This can be easily seen by downsampling the detail signal

$$H\mathbf{d}_{\text{ol}} = H(I_N - GH)\mathbf{x} = (H - HGH)\mathbf{x} = \mathbf{0}_{N/2\times 1}. \tag{13}$$



(a) Encoder (b) Decoder

Figure 5: Lifted-pyramid structure with an update step.

Therefore, the correlation between the coarse resolution signal $\mathbf{c}$ and the detail signal $\mathbf{d}_{\text{ol}}$ is equal to zero.

Biorthogonality is a constrained relationship between the downsampling and the upsampling filters: if the two filters are concatenated, the resulting filter is a half-band filter which is symmetric about the frequency $\pi/2$ [5, 12]. A sharp roll-off of the decimation filter will require that the upsampling filter has an overshoot close to the frequency $\pi/2$. This has a negative impact on the compression efficiency of enhancement layers. Therefore, the filters used in the JSVM are usually nonbiorthogonal. Throughout this paper, we assume nonbiorthogonal downsampling and upsampling filters for the LP as used in the JSVM.

Nonbiorthogonality, however, creates correlation between the low-resolution-coarse signal and the detail signal. This can be seen from the following equation:

$$H\mathbf{d}_{\text{ol}} = H(I_N - GH)\mathbf{x} = (I_K - HG)H\mathbf{x} = (I_K - HG)\mathbf{c}. \tag{14}$$

Since $HG \neq I_K$, the right-hand side, in general, is nonzero. The above equation can also be rewritten as

$$H\mathbf{d}_{\text{ol}} = (I_K - HG)\mathbf{c} = \mathbf{c} - H\mathbf{p}_{\text{ol}}, \tag{15}$$

where $\mathbf{p}_{\text{ol}}$ denotes the open-loop prediction. This shows that the low frequency component in the detail signal is equal to the difference between the coarse signal and the downsampled prediction signal. From compression point of view, this correlation is an undesired feature since it leads to higher bit rate.

4.1. Reduced-coarse signal

From (15), it is evident that the detail signal contains some nonzero low-frequency component. This component can be removed from the coarse signal since it can always be extracted from the detail signal at the receiver. Thus, we can update the coarse signal as

$$\mathbf{c}_r := \mathbf{c} - H\mathbf{d}_{\text{ol}} = \mathbf{c} - (I_K - HG)\mathbf{c} = HG\mathbf{c} = H\mathbf{p}_{\text{ol}}, \tag{16}$$

where $\mathbf{c}_r$ denotes the reduced coarse signal. Thus, the reduced coarse signal is equal to the filtered and downsampled prediction signal. We term the new signal as reduced-coarse signal since the operation of upsampling followed by downsampling can only lose information. The energy of the new coarse signal relative to the original coarse signal, however, depends on the signal itself, and can be bounded by the squares of the maximum and the minimum singular values of the operator HG.

The updated coarse signal and the detail signal are quantized at the desired bit rate and are transmitted. At the receiver, the decoder can estimate the coarse signal and the original signal at higher resolution as

$$\begin{aligned} \hat{\mathbf{c}} &:= \mathbf{c}_{rq} + H\mathbf{d}_q, \\ \hat{\mathbf{x}}_c &:= G\hat{\mathbf{c}} + \mathbf{d}_q = G\mathbf{c}_{rq} + (I_N + GH)\mathbf{d}_q, \end{aligned} \tag{17}$$

where $\mathbf{c}_{rq}$ denotes the quantized reduced coarse signal. Observe that, to reconstruct the lower resolution coarse signal $\mathbf{c}$, the receiver needs to have received the higher resolution detail signal $\mathbf{d}$. Therefore, the above algorithm is not suitable for SVC application. The alternative approach would be to design the decimation and interpolation filters such that the updated coarse signal $\mathbf{c}_r$, instead of $\mathbf{c}$, has the desired low-resolution quality. This approach does not require the availability of the detail signal to reconstruct the desired coarse signal, which is $\mathbf{c}_r$. However, the filters need to be designed differently from those used in the JSVM. Notice that this method is similar to the open-loop LP structure of Flierl and Vandergheynst [4] with the update filter equal to the downsampling filter and the addition (subtraction) operation at the encoder (decoder) replaced by the subtraction (addition). They propose to follow the second approach through the appropriate design of the three filters so that the equivalent downsampling filter has the desired frequency response. Therefore, in their approach, the low-resolution signal can be reconstructed without the higher-resolution detail signal.

4.2. Reduced-detail signal

The second method to reduce the correlation is to keep the low-resolution signal intact, but to remove the low-frequency part from the detail signal. As we see in (15), this part could be always computed by the decoder once it had received the low resolution signal $\mathbf{c}$. The encoder thus can update the

detail signal as

$$\mathbf{d}_r := \mathbf{d}_{\text{ol}} - GH\mathbf{d}_{\text{ol}} = (I_N - GH)\mathbf{d}_{\text{ol}}, \tag{18}$$

where $\mathbf{d}_r$ denotes the reduced detail signal. From the expression on the right-hand side, we observe that the reduced detail signal is nothing but the "detail of the detail," that is, the detail signal for the LP representation of the original detail signal. We term the new detail as the reduced detail signal, since the removal of the coarse component from the original detail signal tends to reduce its energy. By substituting the value of the low-frequency component from (15) and the detail signal from (3), we get

$$\mathbf{d}_r = \mathbf{x} - G\mathbf{c} - G(I_K - HG)\mathbf{c} = \mathbf{x} - (2I_N - GH)G\mathbf{c}. \tag{19}$$

Thus, the updated detail signal can be obtained in one step through an improved prediction which is given as

$$\mathbf{p}_r := (2I_N - GH)G\mathbf{c} = (2I_N - GH)\mathbf{p}_{\text{ol}}. \tag{20}$$

The coarse signal and the improved detail signal are quantized at the desired bit rate and are transmitted. At the receiver, the decoder first estimates the prediction and then reconstructs the original signal as

$$\begin{aligned} \hat{\mathbf{p}}_r &:= (2I_N - GH)G\mathbf{c}_q, \\ \hat{\mathbf{x}}_d &:= \hat{\mathbf{p}}_r + \mathbf{d}_{rq} = (2I_N - GH)G\mathbf{c}_q + \mathbf{d}_{rq}, \end{aligned} \tag{21}$$

where $\mathbf{d}_{rq}$ denotes the quantized reduced detail signal. Note that the correlation between the newly obtained detail signal and the coarse signal is still nonzero because of the nonbiorthogonality. However, it can be shown that the new correlation is less than the original correlation. Since the detail signal undergoes quantization after transform coding, and the downsampling and upsampling operations increase the complexity, we do not iterate the above operation further.

The above method has the advantage that it suits the SVC application. We do not require the higher-resolution enhancement layer signal to decode the low-resolution base layer signal without redesigning the JSVM filters. However, the above method still suffers from the problem of the open-loop, that is, the error at the higher resolution depends on the quantization error of the coarse layer. In the following, we present the above two methods in the closed-loop mode. As we will see later, only the second method will lead to the reconstruction error which is independent of the quantization error of the coarse layer.

5. IMPROVED CLOSED-LOOP LP STRUCTURES

The purpose of the closed-loop prediction in the classical LP structure is to avoid the mismatch between the predictions at the encoder and at the decoder. This is achieved by interpolating the quantized or decoded coarse resolution signal as the prediction. Since the predictions at the encoder and the decoder are identical, the reconstruction error is solely dependent on the quantization error of the detail signal. Further, this also implies that the reconstruction error is bounded by the quantization step size of the detail layer. In the following, we use the same notations as with the open-loop configuration in order to avoid introducing further notations, but their meanings should be clear from the considered configuration.

5.1. Reduced-coarse signal

In the closed-loop configuration, the encoder updates the coarse signal based on the quantized detail signal. Thus, the reduced detail signal is obtained as

$$\mathbf{c}_r := \mathbf{c} - H\mathbf{d}_q. \tag{22}$$

As in the open-loop configuration, the updated coarse signal is quantized at the desired bit rate and is transmitted. At the receiver, the decoder estimates the coarse signal and the original signal at higher resolution using (17). Observe that, because of the quantized detail signal inside the update loop, the update signal at the encoder and the decoder are identical. This update signal can be expressed as

$$H\mathbf{d}_q = H(\mathbf{d}_{\text{ol}} + \mathbf{q}_d) = (I_K - HG)\mathbf{c} + H\mathbf{q}_d, \tag{23}$$

where $\mathbf{q}_d$ represents the quantization noise of the detail signal. Here we have assumed an additive quantization noise model. If the quantization noise is assumed to be highpass, the second term on the right-hand side almost vanishes. Therefore the update signal is almost the same as that in the case of the open-loop configuration. As a consequence, there will not be much difference in the reconstruction error compared to that with the open-loop structure.

5.2. Reduced-detail signal

In the closed-loop configuration, the new prediction will be based on the quantized- or decoded-coarse signal. Thus, the new detail signal is obtained as

$$\mathbf{d}_r := \mathbf{x} - (2I_N - GH)G\mathbf{c}_q. \tag{24}$$

The improved detail signal is quantized at the desired bit rate and is transmitted. At the receiver, the decoder first computes the prediction and then reconstructs the original signal using (21). Because the decoder also uses the decoded-coarse signal for prediction, there is no mismatch between the predictions made at the encoder and the decoder.

In the closed-loop prediction, the quality of the prediction depends on the quantization parameter of the coarse signal. If the quantization parameter is high, the detail signal can have larger energy, which implies higher bit rate. The same is true for the proposed closed-loop structures.

Because of the compatibility with the SVC architecture, here we will consider only the last method, that is, the closed-loop improved prediction, for integration in the JSVM. The two configurations with reduced-coarse signal can be incorporated in the SVC architecture provided the filters are designed such that the reduced coarse signal has the desired quality without aliasing. We will not address this problem here since the filter design for SVC is a separate problem.

6. RECONSTRUCTION ERROR ANALYSIS

Here we will assume that there is no channel noise, or equivalently all the channel errors have been successfully corrected by forward error correction schemes. Thus, the reconstruction error at the receiver is solely due to the quantization noise. In the following, we analyze the error performance of the two methods in both the open-loop and the closed-loop configurations.

6.1. Open-loop LP structures

As before, we will consider an LP with only one level of decomposition. For the sake of simplicity of analysis, we will assume that the coarse and the detail signals are scalar quantized. The quantization step sizes are small enough so that the corresponding quantization noise components can be assumed to be zero-mean, white, and uncorrelated. Further, since in the open-loop the coarse signal and the detail signal are quantized independently, their quantization noises can be assumed to be uncorrelated.

6.1.1. LP with standard reconstruction

Let $\mathbf{q}_c$ and $\mathbf{q}_d$ denote the quantization noises for the coarse signal and the detail signal, respectively. Assuming the quantization noise to be additive, we can write

$$\mathbf{c}_q = \mathbf{c} + \mathbf{q}_c, \qquad \mathbf{d}_q = \mathbf{d}_{\text{ol}} + \mathbf{q}_d. \tag{25}$$

Because of the afore-mentioned white-noise assumptions,

$$\mathbb{E}(\mathbf{q}_c\mathbf{q}_c^t) = \sigma_c^2 I_K, \qquad \mathbb{E}(\mathbf{q}_d\mathbf{q}_d^t) = \sigma_d^2 I_N, \tag{26}$$

where σ_c^2 and σ_d^2 denote the variances of the coarse and the detail signal components, respectively, and $\mathbb{E}$ denotes the mathematical expectation. Further, because of the assumption of zero cross-correlation between the coarse signal and the detail signal,

$$\mathbb{E}(G\mathbf{q}_c\mathbf{q}_d^t) = \mathbb{E}(\mathbf{q}_d\mathbf{q}_c^t G^t) = \mathbf{0}_{N\times N}. \tag{27}$$

Referring to (7) and (3), the reconstruction error can be expressed as

$$\mathbf{e}_s := \hat{\mathbf{x}}_s - \mathbf{x} = (G\mathbf{c}_q + \mathbf{d}_q) - (G\mathbf{c} + \mathbf{d}_{\text{ol}}) = G\mathbf{q}_c + \mathbf{q}_d. \tag{28}$$

Thus, the mean square error with the standard reconstruction is given as:

$$\begin{aligned}\text{MSE}_s &:= \frac{1}{N}\mathbb{E}||\mathbf{e}_s||^2 = \frac{1}{N}\mathbb{E}(\mathbf{e}_s^t\mathbf{e}_s)\\ &= \frac{1}{N}\mathbb{E}\left((G\mathbf{q}_c + \mathbf{q}_d)^t(G\mathbf{q}_c + \mathbf{q}_d)\right)\\ &= \frac{1}{N}\sigma_c^2\,\text{tr}\,(G^tG) + \sigma_d^2,\end{aligned} \tag{29}$$

where the last expression follows from the assumptions stated above. Here $\text{tr}(\cdot)$ denotes the trace of the matrix. We see that the reconstruction error is a function of the quantization error of both the coarse signal and the detail signal. Therefore, in a multiple-level LP, the reconstruction error at any level contributes to the reconstruction error at all higher-resolution levels. Observe that the reconstruction error is also a function of the upsampling filter matrix G. In practice, the quantization noise of the coarse signal is dependent on the coarse signal itself, and therefore, the reconstruction error is also an indirect function of the downsampling filter matrix H.

6.1.2. LP with frame reconstruction

Let $\mathbf{q}_{cu}$ denote the quantization noise of the updated coarse signal. Therefore, we can write $\mathbf{c}_{uq} = \mathbf{c}_u + \mathbf{q}_{cu}$. Referring to (12) for the frame reconstruction with an update, and using (3) and (11), the reconstruction error can be expressed as

$$\mathbf{e}_u := \hat{\mathbf{x}}_u - \mathbf{x} = G\mathbf{q}_{cu} + (I_N - GF)\mathbf{q}_d. \tag{30}$$

Let us assume that $\mathbf{q}_{cu}$ has similar statistical properties as that of $\mathbf{q}_c$, that is, its components are white and uncorrelated with variance σ_{cu}^2, and they are uncorrelated with the components of $\mathbf{q}_d$. Using similar steps as for the standard reconstruction, the mean square error expression can be obtained as

$$\begin{aligned}\text{MSE}_u := \frac{1}{N}\mathbb{E}||\mathbf{e}_u||^2 &= \frac{1}{N}\sigma_{cu}^2\,\text{tr}\,(G^tG)\\ &+ \frac{1}{N}\sigma_d^2\,\text{tr}\,\left((I_N - GF)^t(I_N - GF)\right).\end{aligned} \tag{31}$$

In the special case when $f(n) = h(n)$, $F = H$, and therefore

$$\text{MSE}_u = \frac{1}{N}\sigma_{cu}^2\,\text{tr}\,(G^tG) + \frac{1}{N}\sigma_d^2\,\text{tr}\,\left((I_N - GH)^t(I_N - GH)\right). \tag{32}$$

If the upsampling filter $g(n)$ and the update filter $f(n)$ turn out to be biorthogonal, $FG = I_K$. In that case, the mean square error can be simplified as

$$\begin{aligned}\text{MSE}_u &= \frac{1}{N}\sigma_{cu}^2\,\text{tr}\,(G^tG) + \sigma_d^2\left(1 - \frac{K}{N}\right)\\ &= \frac{1}{N}\sigma_{cu}^2\,\text{tr}\,(G^tG) + \frac{\sigma_d^2}{2}, \quad (\because K = N/2).\end{aligned} \tag{33}$$

6.1.3. Reduced-coarse signal

We will assume that the quantization noise of the reduced-coarse signal has similar statistical properties as of the original coarse signal. Even though the update signal depends on the detail signal, for simplicity we will assume that the quantization noise of the reduced coarse and the detail signals are uncorrelated. Let $\mathbf{q}_{cr}$ denote the quantization noise of the reduced-coarse signal. Therefore, $\mathbf{c}_{rq} = \mathbf{c}_r + \mathbf{q}_{cr}$. Referring to (17), (3), and (16), the reconstruction error can be expressed as

$$\mathbf{e}_c := \hat{\mathbf{x}}_c - \mathbf{x} = G\mathbf{q}_{cr} + (I_N + GH)\mathbf{q}_d. \tag{34}$$

Thus, the mean square error can be derived as

$$\begin{aligned}\text{MSE}_c := \frac{1}{N}\mathbb{E}||\mathbf{e}_c||^2 &= \frac{1}{N}\sigma_{cr}^2\,\text{tr}\,(G^tG)\\ &+ \frac{1}{N}\sigma_d^2\,\text{tr}\,\left((I_N + GH)^t(I_N + GH)\right),\end{aligned} \tag{35}$$

where σ_{cr}^2 denotes the variance of the quantization noise $\mathbf{q}_{cr}$.

6.1.4. Reduced-detail signal

We will assume that the quantization noise of the reduced-detail signal has similar statistical properties as of the original detail signal. Further, the quantization noises of the coarse and the detail signals can be assumed to be uncorrelated. Let $\mathbf{q}_{dr}$ denote the quantization noise of the reduced detail signal. Therefore, $\mathbf{d}_{rq} = \mathbf{d}_r + \mathbf{q}_{dr}$. Referring to (21) and (19), the reconstruction error can be expressed as

$$\mathbf{e}_d := \hat{\mathbf{x}}_d - \mathbf{x} = (2I_N - GH)G\mathbf{q}_c + \mathbf{q}_{dr}. \tag{36}$$

Thus, the mean square error can be derived as

$$\begin{aligned}\mathrm{MSE}_d &:= \frac{1}{N}\mathbb{E}||\mathbf{e}_d||^2 \\ &= \frac{1}{N}\sigma_c^2 \,\mathrm{tr}\,(G^t(2I_N - GH)^t(2I_N - GH)G) + \sigma_{dr}^2,\end{aligned} \tag{37}$$

where σ_{dr}^2 denotes the variance of the quantization noise $\mathbf{q}_{dr}$. We observe that, for both structures, the reconstruction error at any level of LP is dependent on the reconstruction errors on the lower resolution layers.

6.2. Closed-loop LP structures

Let $\mathbf{q}_c$ and $\mathbf{q}_d$ denote the quantization noises for the coarse signal and the detail signal, respectively. Assuming the quantization noise to be additive, we can write

$$\mathbf{c}_q = \mathbf{c} + \mathbf{q}_c, \qquad \mathbf{d}_q = \mathbf{d}_{\mathrm{cl}} + \mathbf{q}_d. \tag{38}$$

We use the same notations for the errors and mean square errors as in the open-loop configurations in order to avoid introducing further symbols. We will further assume that the quantization noises have similar statistical properties as in the case of open-loop configurations.

6.2.1. LP with standard reconstruction

Referring to (7) for standard reconstruction, and using (6) and (38), the reconstruction error can be expressed as

$$\mathbf{e}_s := \hat{\mathbf{x}}_s - \mathbf{x} = (G\mathbf{c}_q + \mathbf{d}_q) - (G\mathbf{c}_q + \mathbf{d}_{\mathrm{cl}}) = \mathbf{q}_d. \tag{39}$$

Thus, the mean square error with the standard reconstruction can be computed as follows:

$$\mathrm{MSE}_s := \frac{1}{N}\mathbb{E}||\mathbf{e}_s||^2 = \sigma_d^2. \tag{40}$$

We see that the reconstruction error is equal to the quantization error of the detail signal. This is true even if we have an LP with multiple layers.

6.2.2. Reduced coarse signal

Referring to (17), (3), and (22), the reconstruction error can be expressed as

$$\mathbf{e}_c := \hat{\mathbf{x}}_c - \mathbf{x} = G\mathbf{q}_{cr} + \mathbf{q}_d. \tag{41}$$

The mean square error thus can be derived as

$$\mathrm{MSE}_c := \frac{1}{N}\mathbb{E}||\mathbf{e}_c||^2 = \frac{1}{N}\sigma_{cr}^2 \,\mathrm{tr}\,(G^tG) + \sigma_d^2. \tag{42}$$

We see that the mean square error has a similar form to that of the standard reconstruction in the open-loop structure. Since the aim of updating is to reduce the energy, the encoding of the updated signal would have better rate-distortion performance. This would imply effectively better rate-distortion performance for the original signal at the higher resolution. It is evident that, like the open-loop structures, the error is dependent on the quantization noise of the lower-base layer.

6.2.3. Reduced-detail signal

Referring to (21) and (24), the reconstruction error can be expressed as

$$\mathbf{e}_d := \hat{\mathbf{x}}_d - \mathbf{x} = \mathbf{q}_{dr}. \tag{43}$$

The error thus depends only on the quantization noise of the reduced detail layer. The mean square error can be derived as

$$\mathrm{MSE}_d := \frac{1}{N}\mathbb{E}||\mathbf{e}_d||^2 = \sigma_{dr}^2. \tag{44}$$

The aim of the improved prediction is to reduce the energy of the detail signal. Following the results in information theory [13], this would result in a better rate-distortion performance for the encoding of the enhancement layer. This implies that, for a given bit rate, the improved prediction would result in less distortion. Comparing (40) and (44), this would mean that $\sigma_{dr}^2 < \sigma_d^2$.

7. TRANSFORM CODING OF ENHANCEMENT LAYER

In practice, the detail signal undergoes an orthogonal transform before being quantized and entropy coded. The transform aims to remove the spatial correlation in the detail signal coefficients and to compact its energy in fewer number of coefficients. The current SVC standard, for this purpose, uses a 4×4 integer transform, which is an approximation of the discrete cosine transform (DCT) applied over a block size of 4×4. The DCT, however, may not be the optimal transform since the detail signal contains more high frequency components. A closer look at (3) reveals that the detail signal has certain inherent structure. Most of its energy is concentrated along certain directions which are decided by the downsampling and the upsampling filters. These directions can be found out by the singular value decomposition [14] of $I_N - GH$ as follows:

$$I_N - GH \equiv U\Sigma V^t, \tag{45}$$

where U and V are $N \times N$ orthogonal matrices and Σ is an $N \times N$ diagonal matrix. In [15], we have shown that, in open-loop configuration with biorthogonal upsampling and downsampling filters, either the U matrix or the V matrix applied on the detail signal leads to a critical representation

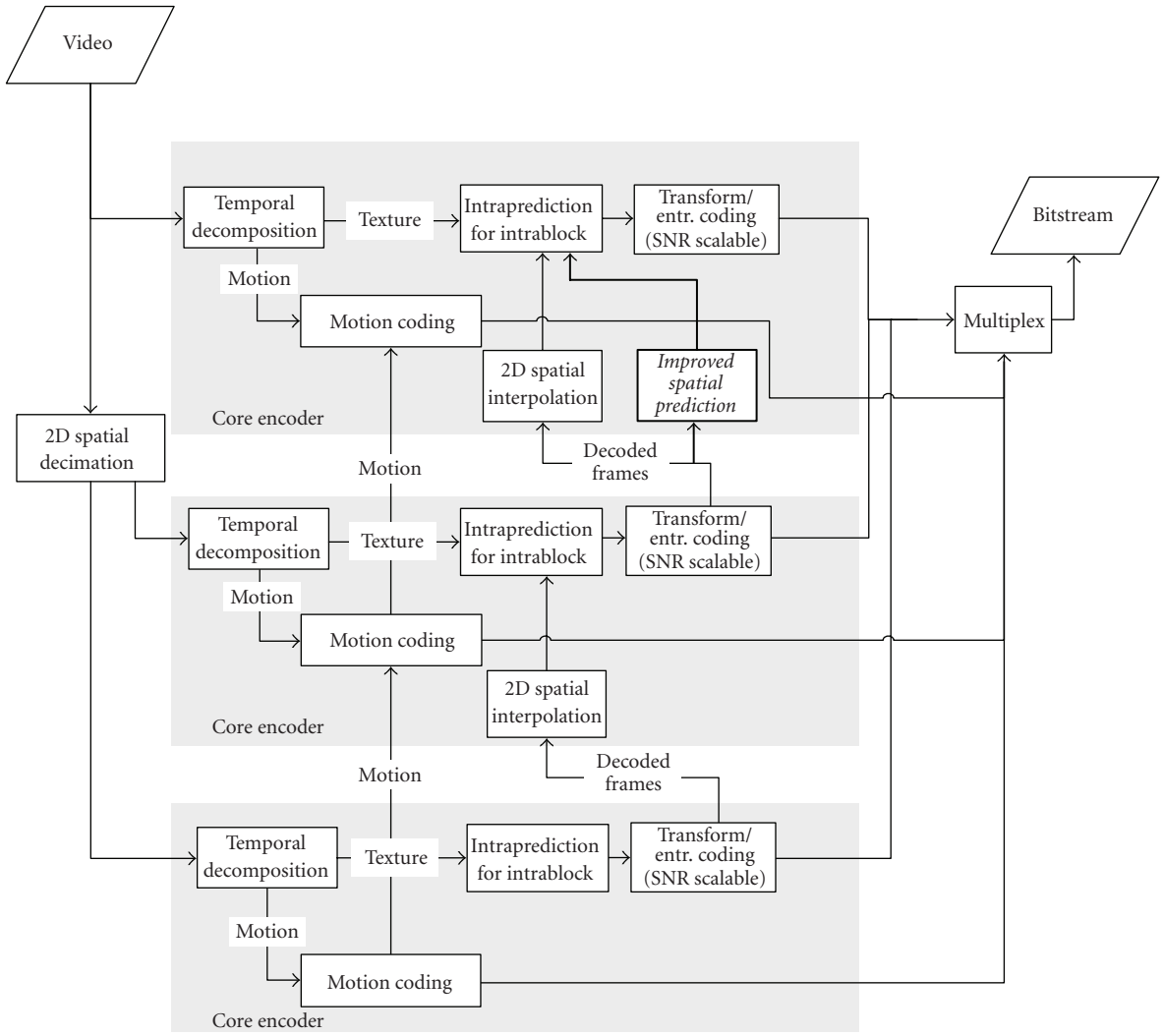


FIGURE 6: Improved scalable encoder using a multiscale pyramid with 3 levels of spatial scalability [1]. The proposed algorithm is embedded in the "improved spatial prediction" module for the spatial intraprediction of the SD layer from the CIF layer for I and P frames.

of the LP. We refer to these matrices as the U-transform and the V-transform, respectively. The 4×4 integer transform applied in the JSVM is referred to as the DCT hereafter.

Under the closed-loop configuration, the above structure is somewhat weakened. The introduction of the quantization noise in the prediction loop destroys the redundancy structure of the LP. Nevertheless, the above matrices are orthogonal and can always be applied to the original detail or the newly-obtained detail signal. The decoder can use the transpose of these matrices for the inverse transformation. Experimental results presented in [15] showed that the V-transform had a slightly better R-D performance than the U-transform. Therefore, for the actual implementation with JSVM, we consider only the V-transform.

8. IMPLEMENTATION WITH JSVM

Figure 6 depicts the structure of the improved JSVM encoder with the proposed spatial prediction module. The original JSVM encoder is described in [1]. The encoder supports quality, temporal, and spatial scalabilities. A quality base layer residual provides minimum reconstruction quality at each spatial layer. This quality base layer can be encoded into an AVC compliant stream if no interayer prediction is applied. Quality enhancement layers are additionally encoded and can be chosen to either provide coarse or fine grain quality (SNR) scalability. To achieve temporal scalability, hierarchical B pictures are employed. The concept of hierarchical B pictures provides a fully predictive structure that is already provided with AVC. Alternatively, motion compensated temporal filtering (MCTF) can be used as a nonnormative encoder configuration for temporal scalability.

The encoder is based on a layered approach to achieve spatial scalability. It provides a downsampling stage that generates the lower-resolution signals for lower layers. Each spatial resolution (except the base layer, which is AVC coded) includes refinement of the motion and texture information, and the core encoder block for each layer basically consists

Table 1: Average number of MBs for mode selection over 8 intraframes for CITY SD at different QPs.

QP		Spatial intra	$\mathbf{d}$		$\mathbf{d}'$	
QCIF/CIF	SD		DCT	V-trans.	DCT	V-trans.
18,18	30	21.25	189.875	76.125	739.125	557.625
	36	13.625	182.125	18.25	988.75	381.25
	42	2.125	179.625	2.75	1289.875	109.625
	48	0	176.375	0.625	1370.125	36.875
24,24	30	33.125	465.875	164.75	578.375	341.875
	36	16	384.5	67.5	763	353
	42	2	383	7.75	1075.375	115.875
	48	0	384.375	0.25	1161.75	37.625

of an AVC encoder. The spatial resolution hierarchy is highly redundant. As shown in Figure 6, the redundancy between adjacent spatial layers is exploited by different interlayer prediction mechanisms for motion parameters as well as for texture data. For the texture data, the prediction mechanism amounts to computing a difference signal between the original higher-resolution signal and the interpolated version of the coded and decoded signal at the lower-spatial resolution.

In our implementation, we aim to improve the coding performance by exploiting the redundancy of the Laplacian pyramid structure adopted for spatial scalability. To that end, we modify only the interlayer texture prediction module keeping the other modules same as in the original JSVM. Furthermore, the original downsampling and upsampling filters are maintained. This means that the improved prediction in (21) is obtained with the existing JSVM filters H and G. The Fidelity Range Extension (FRExt) of SVC supports the high profiles and adds more coding efficiency without a significant amount of implementation complexity. The new features in FRExt include an adaptive transform block-size and perceptual quantization scaling matrices. Our proposed method also applies to FRExt, as will be discussed later. Through theoretical analysis, improved interlayer motion and residual prediction can also be achieved, and this remains a future work.

As we have mentioned earlier, in the current JSVM software, the interlayer prediction is implemented in the closed-loop mode. For each macroblock (MB), the selection of prediction modes (interlayer, spatial-intra, temporal, etc.) is based on a rate-distortion optimization (RDO) procedure. However, the closed-loop structure does not guarantee an improved rate-distortion performance either with the modified prediction or with the V-transform; the performance can vary depending on the local signal statistics. Thus, to apply the proposed method in SVC, we propose three additional MB modes employing the improved prediction and the V-transform besides the existing interlayer prediction mode. The three proposed MB modes are (i) existing interlayer prediction followed by V-transform ($\mathbf{d}$ + V-transform), (ii) improved prediction followed by DCT ($\mathbf{d}'$ + DCT), (iii) improved prediction followed by V-transform ($\mathbf{d}'$ + V-transform). We refer to the existing mode, interlayer prediction followed by DCT, as "$\mathbf{d}$ + DCT." The three proposed modes are applied for encoding the SD layer by prediction from the CIF layer.

The mode selection statistics over several frames are shown in Table 1 for intraframes. These statistics are obtained by including all the modes in the original JSVM software together with the three proposed modes and running over 8 intraframes of the CITY video sequences. The improved prediction and the V-transform are applied only to the SD layer while the QCIF and CIF layers are encoded using the existing modes. The table shows the number of macroblocks undergoing different modes for different QP values of QCIF, CIF, and SD layers. Note that the size of a macroblock is 16×16 and the total number of macroblocks in an SD image (with the resolution of 704×576) is equal to 1584. Thus, in Table 1, the entries (number of macroblocks) in each row add up to 1584.

From Table 1, first we observe that majority of macroblocks choose the improved prediction irrespective of the transform method followed, and especially at high QP values of SD. This demonstrates that the proposed interlayer prediction successfully reduces the redundancy and energy in the detail signal.

Second, the number of blocks following the V-transform is significant at low QPs of SD. However, the number of blocks selecting the V-transform is always less than that of blocks selecting the DCT. One reason is that the rate-distortion in the current implementation is optimized w.r.t. DCT. The rate-distortion optimization in mode selection plays an important role to the overall coding performance. In general video encoders, the mode that minimizes the coding cost, which is defined as

$$f \equiv R + \lambda D, \tag{46}$$

will be selected. Here R is the bitrate for coding the MB mode syntax as well as the residual data and D is the corresponding distortion. The optimal Lagrange multiplier λ should be selected such that line f is tangent with the R-D curve, and is defined as

$$\lambda \equiv 0.85 \times 2^{\min(52,QP)/3-4} \tag{47}$$

Table 2: Definition of macroblock modes for I and P frames in JSVM and proposed encoding scheme.

For I frames:

JSVM	Spatial-intra	Intra_4 × 4, Intra_8 × 8
	Interlayer texture	(**d** + DCT)_4 × 4, (**d** + DCT)_8 × 8
Proposed	Interlayer texture	(**d** + DCT)_4 × 4, (**d** + DCT)_8 × 8 (**d** + V-trans)_4 × 4, (**d** + V-trans)_8 × 8 (**d**′ + DCT)_4 × 4, (**d**′ + DCT)_8 × 8 (**d**′ + V-trans)_4 × 4, (**d**′ + V-trans)_8 × 8

For P frames:

JSVM	Spatial-intra	Intra_4 × 4, Intra_8 × 8
	Temporal	Skip, Inter_16 × 16, Inter_16 × 8, Inter_8 × 16, Inter_8 × 8
	Interlayer texture	(**d** + DCT)_4 × 4, (**d** + DCT)_8 × 8
	Interlayer MV/resi.	IntraBLSkip, Inter_4, Inter_8, Inter_16
Proposed	Temporal	Skip, Inter_16 × 16, Inter_16 × 8, Inter_8 × 16, Inter_8 × 8
	Interlayer texture	(**d** + DCT)_4 × 4, (**d** + DCT)_8 × 8 (**d** + V-trans)_4 × 4, (**d** + V-trans)_8 × 8 (**d**′ + DCT)_4 × 4, (**d**′ + DCT)_8 × 8 (**d**′ + V-trans)_4 × 4, (**d**′ + V-trans)_8 × 8
	Interlayer MV/resi.	IntraBLSkip, Inter_4, Inter_8, Inter_16

empirically in the current JSVM implementation. However, this λ is defined according to the DCT of the data to be encoded, and does not optimize the R-D performance of the V-transform. Still we notice that the number of MBs selecting the V-transform is significant with lower QPs of the SD layer. The improvement of applying the V-transform in our proposed method remains a future work.

Overall, the proposed modes seem to be the chosen ones, especially for low QP values of CIF and QCIF layers, that is, better base layer qualities. It is also clear that the number of MBs selecting the spatial intra mode is much smaller than the number of MBs selecting the interlayer prediction modes. Thus, we propose to suppress the spatial intra mode and include the other three interlayer prediction modes. More specifically, the MB modes used in original JSVM and the proposed encoding scheme for I and P frames are defined as in Table 2. Note that all the 8×8 modes are valid only when FREext is enabled.

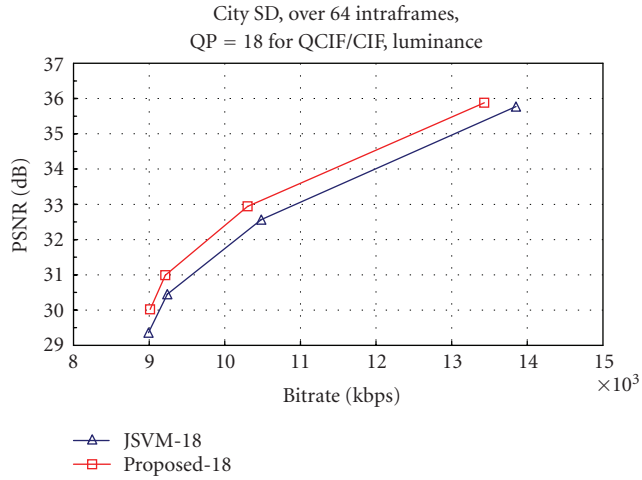
Note that the V-transform is always applied over macroblocks of size 16×16 for the luma component and of size 8×8 for the chroma components. Over a macroblock of size 16×16 (luma) or 8×8 (chroma), the order of complexity is about the same as that of the existing 4×4 transform except that the operations use floating-point numbers. In the proposed modes adopting the V-transform, that is, **d** + V-transform_4×4, **d**+V-transform_8×8, **d**′+V-transform_4×4, **d**′ + V-transform_8 × 8, the suffix (4×4 or 8×8) refers to the block sizes for zigzag scanning of the transform coefficients.

Accordingly, the syntax for coding MB modes is also modified. Two extra flags *BLDetailFlag* and *BLTransformFlag* are needed in the syntax for signaling the additional MB modes. *BLDetailFlag* defines whether the original spatial prediction or the improved prediction is selected, and *BLTransformFlag* selects between DCT and V-transform. These two flags are encoded using the context adaptive binary arithmetic coding (CABAC). Note that, since the spatial intramode, which includes several submodes, is disabled, the number of syntax bits of our proposed encoder remains similar to that of the original JSVM.

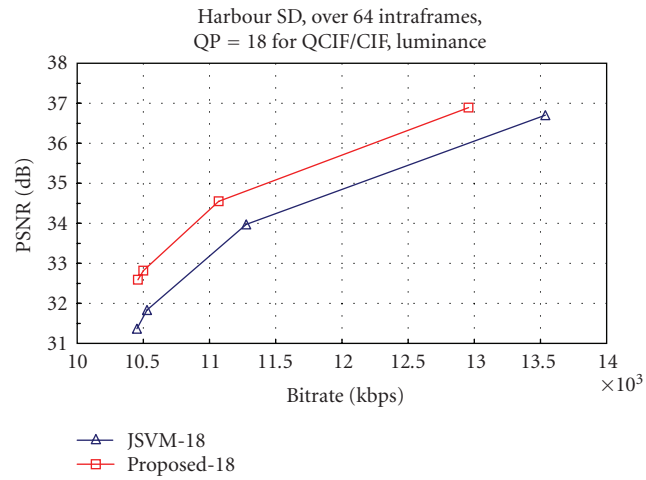
The zigzag scanning, quantization, and entropy coding methods of the transformed coefficients remain unchanged, that is, those techniques as adopted in JSVM are also applied to the transformed coefficients of the MBs selecting the proposed modes. However, the quantizer in JSVM is designed to be used in conjunction with the integer DCT transform, and a multiplication factor MF is incorporated in the quantization. To quantize the coefficients obtained by V-transform directly, this multiplication factor is removed.

9. EXPERIMENTAL RESULTS AND ANALYSIS

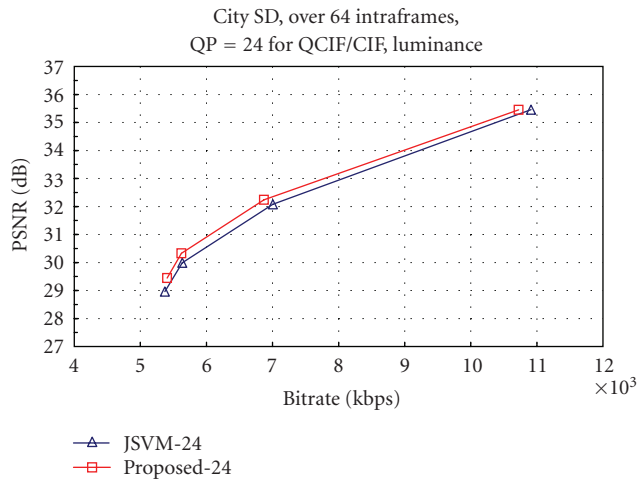
The proposed scheme is tested using standard video sequences CITY and HARBOUR, and the anchor results are obtained by JSVM 4.0. In the encoding of 3 spatial layers, that is, QCIF, CIF, and SD, the proposed method is only applied between the CIF layer and the SD layer. Thus, only the coding



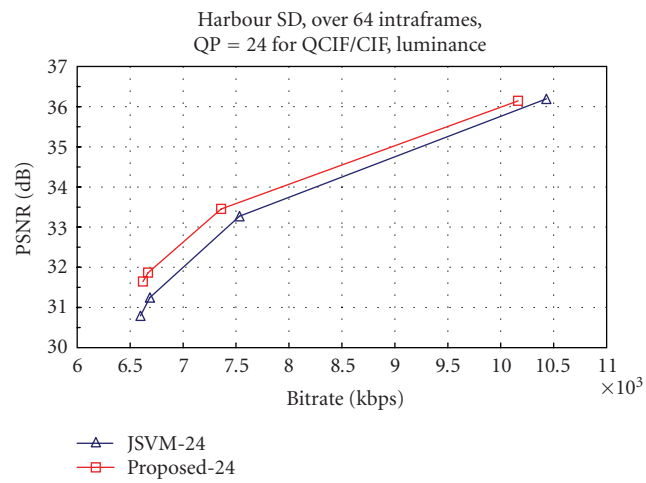
(a) CITY, QP = 18



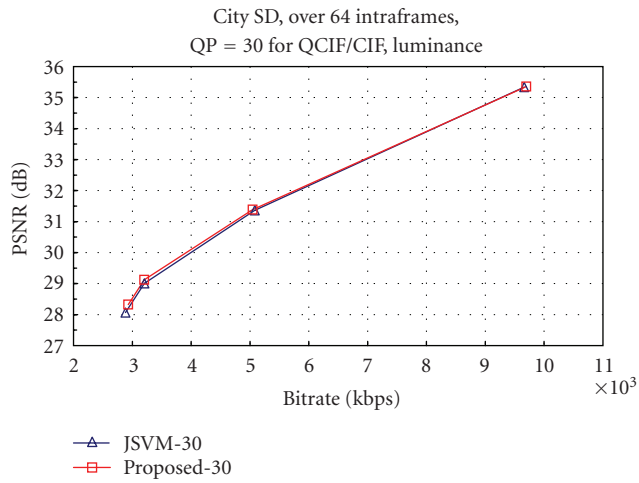
(b) HARBOUR, QP = 18



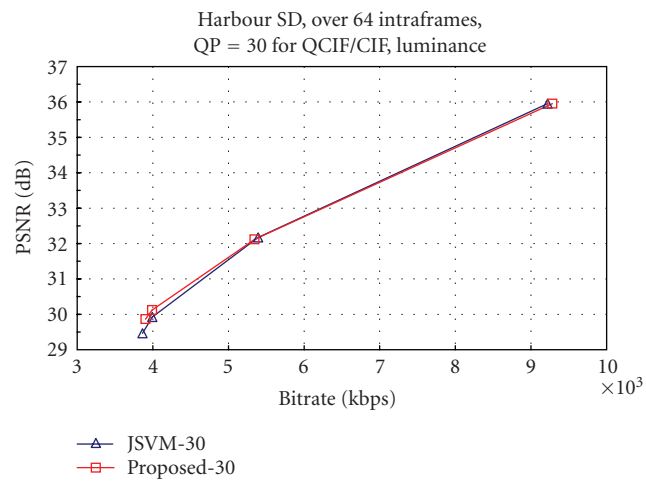
(c) CITY, QP = 24



(d) HARBOUR, QP = 24

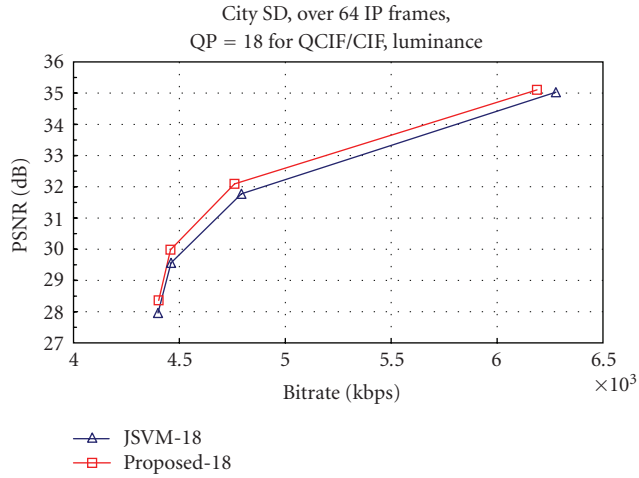


(e) CITY, QP = 30

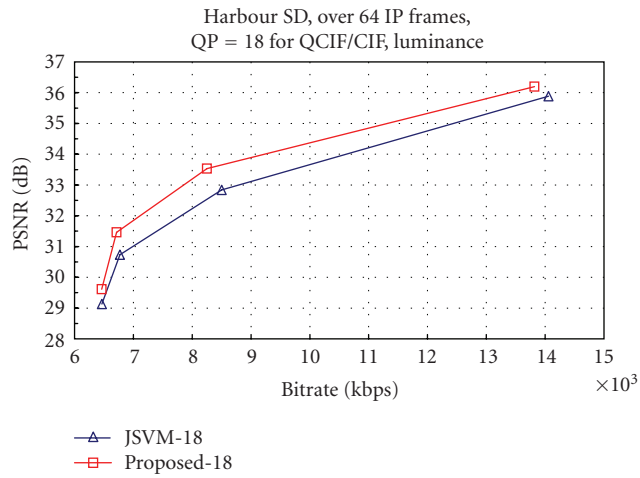


(f) HARBOUR, QP = 30

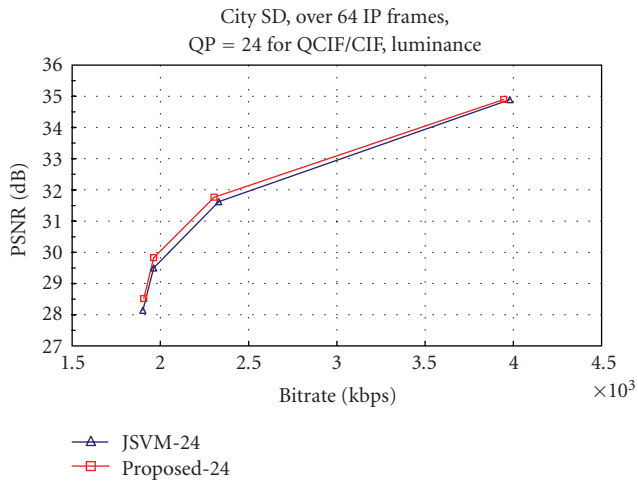
Figure 7: PSNR-rate curves for the luminance component of (a) CITY and (b) HARBOUR SD 30 Hz over 64 intraframes, when QPs for QCIF/CIF are 18, 24, 30.



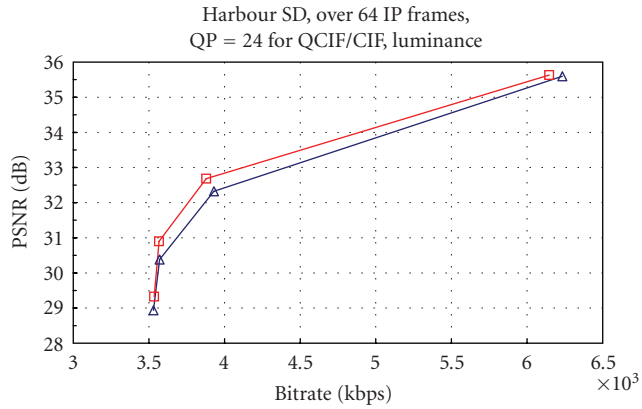
(a) CITY, QP = 18



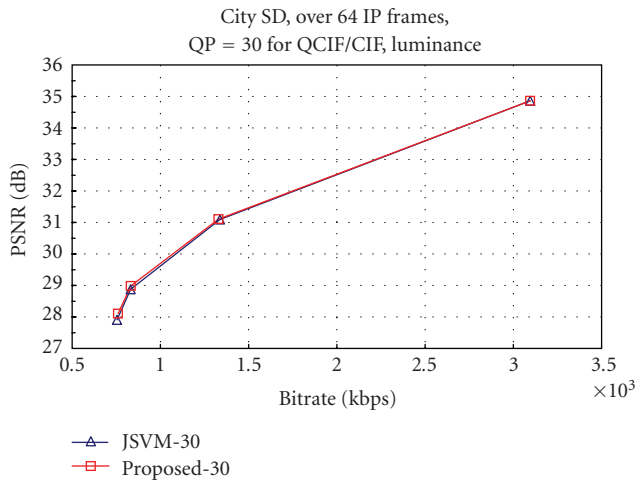
(b) HARBOUR, QP = 18



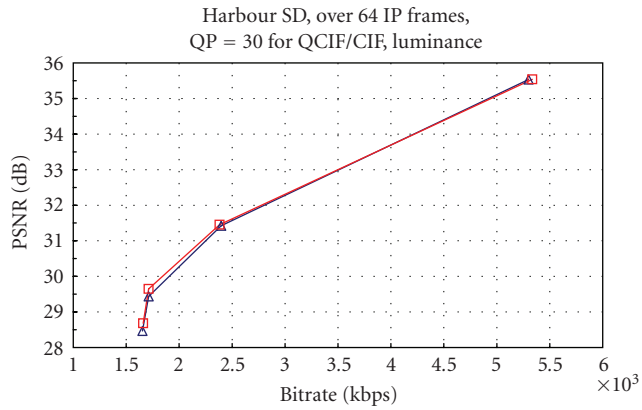
(c) CITY, QP = 24



(d) HARBOUR, QP = 24



(e) CITY, QP = 30



(f) HARBOUR, QP = 30

Figure 8: PSNR-rate curves for the luminance component of (a) CITY and (b) HARBOUR SD 30 Hz over 64 I and P frames, with GOP size = 1 and Intraperiod = 8, when QPs for QCIF/CIF are 18, 24, 30.

results of the SD layer are presented. Since FGS layers are not involved in our experiments, we set both QPs for QCIF/CIF to 18, 24, and 30, which approximately correspond to the base-layer quality with the initial QP 36, 42, and 48 plus three FGS layers. First we test the proposed method using 64 intraframes. Then we test the proposed method using the GOP structure defined as *GOPSize = 1* and *IntraPeriod = 8*, which means one I frame followed by 7 P frames for every 8 frames. Other parameters in the configuration files are listed as follows: FRExt: off for QCIF layer, on for CIF/SD layers; Loop Filter: on; Update Step: 0; Adaptive QP: 1; Inter Layer Pred: 0 for QCIF layer, 2 for CIF/SD layers; Number of FGS layers: 0. Results for all Intraframes are shown in Figure 7, and the results with P frames are shown in Figure 8. We observe that the proposed improved prediction works well for small QP values for CIF/QCIF layers. As the QP is increased, the prediction becomes less efficient. With QP equal to 18, PSNR gain up to 1 dB can be achieved with all intraframes and a gain up to 0.7 dB gain can be achieved with Intra and inter P frames (with HARBOUR sequence).

We must note here that, for all the simulations, we did not modify the entropy coding that follows the transform (DCT or V-transform). In the current JSVM software, it is implemented as context adaptive variable length coding (CAVLC). The current zigzag scan and the coding scheme are optimized for the DCT; therefore, we expect better results if the scanning and encoding of the V-transformed coefficients are modified so as to suit the characteristic of the V-transform. This is a subject of research and we will not pursue it in this paper.

10. CONCLUSIONS

In this paper, we have proposed two improved Laplacian pyramid structures for scalable video coding. The proposed structures exploited the inherent redundancy of the underlying Laplacian pyramid with nonbiorthogonal filters by rendering the enhancement-layer signal less correlated with the base-layer. The first structure updated the base-layer signal by subtracting from it the low-frequency component of the enhancement-layer signal. The second structure modified the prediction with a view to reducing the low-frequency component in the enhancement layer. The corresponding decoder structures were accordingly modified in order to reconstruct the signal at both resolution levels. The simplicity of the structures is reflected by the fact that they did not require to modify the current upsampling filter, nor did they require to design additional filters. Moreover, the structures could be implemented both in the open-loop and in the closed-loop configurations.

We studied the distortion performances of both structures in the open-loop and in the closed-loop configurations. Open-loop structures used unquantized continuous-valued update signals whereas the closed-loop signals used the decoded quantized signals for the purpose. It was demonstrated that only the second structure (with reduced enhancement layer) in the closed-loop configuration leads to a reconstruction error that is dependent on the quantization error of the enhancement layer, but not on the reconstruction error of the lower-resolution layers. Out of the two structures, it was also the only structure that was compatible with the current JSVM architecture.

Along with a recently proposed transform for the enhancement layer, the proposed structure was integrated with JSVM in the SD layer. Based on the experimental results, the macroblock modes in I and P frames were redesigned. Results with test sequences demonstrated that the proposed scheme achieves better R-D performance compared to the original prediction modes. The performance improvement was significant in the case of low-base layer QP suggesting potential application of the proposed method in high-quality scalable video coding.

For the present JSVM integration, there are still some open issues such as the optimization of the VLC for the V-transform, the choice of the λ parameter in rate-distortion optimized mode selection, the optimization of the FGS, and so forth. Further research results along these directions can provide us the complete picture on the true coding performance of the proposed method. Experimental results demonstrate that interlayer prediction is the dominant mode in I frames, and the stationary regions in P and B frames. Thus, the proposed method could have a significant impact on the overall coding performance of still sequences or sequences having low-motion level.

REFERENCES

[1] JVT, "Joint scalable video model JSVM-4," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6)*, Nice, France, October 2005.

[2] P. J. Burt and E. H. Adelson, "The Lapacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[3] M. N. Do and M. Vetterli, "Framing pyramids," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2329–2342, 2003.

[4] M. Flierl and P. Vandergheynst, "An improved pyramid for spatially scalable video coding," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 2, pp. 878–881, Genova, Italy, September 2005.

[5] D. Santa-Cruz, J. Reichel, and F. Ziliani, "Opening the Laplacian pyramid for video coding," in *Proceedings of IEEE International Conference on Image Processing (ICIP '05)*, vol. 3, pp. 672–675, Genova, Italy, September 2005.

[6] A. Segall, "Study of upsampling/down-sampling for spatial scalability," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6)*, Nice, France, October 2005.

[7] A. Segall, "Upsampling and down-sampling for spatial scalability," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6)*, Bangkok, Thailand, January 2006.

[8] C. K. Kim, D. Y. Suh, and G. H. Park, "Directional filtering for upsampling according to direction information of the spatially lower layer," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6)*, Bangkok, Thailand, January 2006.

[9] V. K. Goyal, J. Kovačević, and J. A. Kelner, "Quantized frame expansions with erasures," *Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 203–233, 2001.

[10] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, Pa, USA, 1992.

[11] M. Flierl and P. Vandergheynst, "Inter-resolution transform for spatially scalable video coding," in *Proceedings of Picture Coding Symposium (PCS '04)*, pp. 243–247, San Francisco, Calif, USA, December 2004.

[12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.

[14] G. Strang, *Linear Algebra and Its Applications*, Brooks Cole Publishers, Florence, Ky, USA, 3rd edition, 1988.

[15] G. Rath and C. Guillemot, "Compressing the Laplacian pyramid," in *Proceedings of the 8th IEEE Workshop on Multimedia Signal Processing (MMSP '06)*, pp. 75–79, Victoria, BC, Canada, October 2006.

Wenxian Yang received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2001 and the Ph.D. degree in Computer Engineering from Nanyang Technological University, Singapore, in 2006. In 2004, she was with Microsoft Research Asia, Beijing, China for internship. From 2005 to 2006, she was a Postdoctoral Researcher in the French National Institute for Research in Computer Science and Control (INRIA-IRISA), France. She is now a Postdoctoral Fellow in The Chinese University of Hong Kong. Her research interests include video compression, 3D video compression and processing.

Gagan Rath received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology at Kharagpur in 1990 and the M.E. and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science in Bangalore in 1993 and 1999. He is currently a Research Scientist at INRIA in France. His research interests include signal processing for communications, distributed video coding, scalable video coding, and joint source and channel coding.

Christine Guillemot is currently Directeur de Recherche at INRIA, in charge of the TEMICS research group dealing with image modelling, processing, video communication, and watermarking. She holds the Ph.D. degree from Ecole Nationale Superieure des Telecommunications (ENST) Paris. From 1985 to October 1997, she has been with FRANCE TELECOM/CNET, where she has been involved in various projects in the domain of coding for TV, HDTV and multimedia applications, and coordinated a few (e.g., the European RACE-HAMLET project). From January 1990 to mid 1991, she has worked at Bellcore, NJ, USA, as a visiting scientist. Her research interests are signal and image processing, video coding, and joint source and channel coding for video transmission over the Internet and over wireless networks. She has served as Associate Editor for IEEE Trans. on Image Processing (2000–2003), and for IEEE Trans. on Circuits and Systems for Video Technology (2004–2006). She is a member of the IEEE IMDSP and of the IEEE MMSP technical committees.