EURASIP Journal on Applied Signal Processing 2004:12, 1931–1942

Stochastic Modeling of the Spatiotemporal Wavelet Coefficients and Applications to Quality Enhancement and Error Concealment

Georgia Feideropoulou

Signal and Image Processing Department, ENST, 46 rue Barrault, 75634 Paris Cedex 13, France
Email: feiderop@tsi.enst.fr

Béatrice Pesquet-Popescu

Signal and Image Processing Department, ENST, 46 rue Barrault, 75634 Paris Cedex 13, France
Email: pesquet@tsi.enst.fr

Received 1 September 2003; Revised 9 January 2004

We extend a stochastic model of hierarchical dependencies between wavelet coefficients of still images to the spatiotemporal decomposition of video sequences, obtained by a motion-compensated 2D+t wavelet decomposition. We propose new estimators for the parameters of this model which provide better statistical performances. Based on this model, we deduce an optimal predictor of missing samples in the spatiotemporal wavelet domain and use it in two applications: quality enhancement and error concealment of scalable video transmitted over packet networks. Simulation results show significant quality improvement achieved by this technique with different packetization strategies for a scalable video bit stream.

Keywords and phrases: wavelets, spatiotemporal decompositions, stochastic modeling, hierarchical dependencies, video quality, scalability.

1. INTRODUCTION

Video coding schemes involving motion-compensated spatiotemporal (2D + t) wavelet decompositions [1, 2, 3] have been recently shown to provide very high coding efficiency and to enable complete spatiotemporal, SNR, and complexity scalability [4, 5, 6]. Apart from the flexibility introduced by the scalability of the bit stream, an increased robustness in error-prone environments is possible. Unequal error protection of such kind of bit streams is easily achievable, due to the inherent priority of data. These features make scalable video methods desirable for video transmission over heterogeneous networks, involving, in particular, packet losses. In most cases, however, if packets are lost, an error concealment method needs to be applied. This is usually done after the inverse transformation, that is, in the spatiotemporal domain.

There exists a plethora of error concealment methods of video, most of them applying directly to the reconstructed sequences (for a comparative review, see [7]). Approaches exploiting the redundancy along the temporal axis try to conceal the corrupted blocks in the current frame by selecting suitable substitute blocks from the previous frames. This approach can be reinforced by introducing data partitioning techniques [8]: data in the error prediction blocks are separated in motion vectors and DCT coefficients, which are unequally protected. This way, if the motion vector data are received without errors, the missing blocks are set to their corresponding motion-compensated blocks. However, the loss of a packet usually results in the loss of both the motion vectors and the DCT coefficients. So, many concealment methods first estimate the motion vectors associated with a missing block using the motion vectors of adjacent blocks [9, 10]. Spatial error concealment methods restore the missing blocks only based on the information decoded in the current frame. To restore the missing data, several methods can be used: minimization of a measure of variations (e.g., gradient or Laplacian) between adjacent pixels [11], each pixel in the damaged block is interpolated from the corresponding pixels in its four neighboring blocks such that the total squared border error is minimized [12], or the missing information is interpolated utilizing spatially correlated edge information from a large local neighborhood [13]. Statistical models like the Markov random fields (MRF) have also been proposed for error concealing in video [14, 15]. These methods estimate the missing pixels by exploiting spatial or spatiotemporal constraints between pixels in the original sequence. Note that such approaches can also be employed to estimate missing motion vectors [16].

The error concealment method proposed in this paper is based on a statistical model applied in the transformed wavelet domain. It is a spatiotemporal multiscale model, exhibiting the correlation between discontinuities at different resolution levels in the error prediction (temporal detail) frames.

Hierarchical dependencies between the wavelet coefficients have been largely used for still images [17], for coding in methods like EZW [18] and SPIHT [19], and for denoising [20, 21]. They rely on a quadtree model which has been thoroughly investigated, leading to a joint statistical characterization of the wavelet coefficients [22, 23]. The parent-offspring relations exhibited in the wavelet domain by still images can be extended in the temporal dimension for video sequences and thus lead to an oct-tree [24]. This one can be used to model the spatial and *temporal* dependencies between the wavelet coefficients by taking into account a vector of spatiotemporal ancestors. The extension to motion-compensated 2D + t decompositions implies taking into account additional dependencies and provides insight into the complex nature of these representations. By extending the model proposed in [22, 23] to video sequences, we propose, in this paper, a stochastic modeling of the spatiotemporal dependencies in a motion-compensated 2D + t wavelet decomposition, in which we consider the *conditional* probability law of the coefficients in a given spatiotemporal subband to be Gaussian, with variance depending on the set of the spatiotemporal neighbors. Based on this model, we provide new estimators for the proposed model, showing improved statistical performances. Then we use it to build an optimal mean square predictor for missing coefficients, which is further exploited in two applications of transmitting over packet networks: a quality enhancement technique for resolution-scalable video bit streams and an error concealment method, both applied directly to the subbands of the spatiotemporal decomposition.

The paper is organized as follows. In the next section, we present the stochastic model of spatiotemporal dependencies. In Section 3, several estimators for the model parameters are proposed and tested. In Section 4, we present the prediction method based on the stochastic model. In Sections 5, 6, and 7, we demonstrate the efficiency of our model in the quality enhancement and error concealment methods of scalable video. Section 8 concludes this paper.

2. STOCHASTIC MODELING OF THE SPATIOTEMPORAL DEPENDENCIES BETWEEN WAVELET COEFFICIENTS

The wavelet decomposition, even though ideally decorrelating the input, presents some residual hierarchical dependencies between coefficients that have been exploited in the *zerotree* structures introduced by Shapiro [18]. These parent-offspring structures, in still images, highlight the exponential decay of magnitudes of wavelet coefficients from coarse to fine scales and also their persistence, meaning that spatially correlated patterns (edges, contours, and other discontinuities) propagate through scales.

However, it was shown, for still images, that there is no significant (second-order) correlation between pairs of raw coefficients at adjacent spatial locations ("siblings"), orientations ("cousins"), or scales ("parent" and "aunts"). Instead, their *magnitudes* exhibit high statistical dependencies [22, 25]. We are interested here in exploring the statistical dependencies between the wavelet coefficients resulting from a motion-compensated spatiotemporal decomposition of a video sequence. For this 2D + t decomposition, shown in Figure 1, an extended spatiotemporal neighborhood can be considered [26]. In addition to the spatial neighbors, we take into account additional dependencies with the spatiotemporal parent, its neighbors, and the spatiotemporal "aunts" (see Figure 1).

In order to precise the model, we consider a spatiotemporal subband and let $(c_{n,m})_{1\leq n\leq N,\,1\leq m\leq M}$ be the NM coefficients in this subband. For a given coefficient $c_{n,m}$, we denote by $p_k(n,m)$ all its spatial and spatiotemporal "neighbors" (k being the index over the considered set of neighbors). Similar to the work in [22] on 2D signals, let the prediction of $a_{n,m} = |c_{n,m}|^2$ be

$$l_{n,m} = \sum_k w_k \left| p_k(n,m) \right|^2, \tag{1}$$

where $\mathbf{w} = (w_k)_k$ is the vector of weights.

The high-order statistical dependence involved by this relation can be illustrated via conditional histograms of coefficient magnitudes. In Figure 2, we present such a histogram in log-log scales, conditioned to a mean square linear prediction of squared spatiotemporal neighbors, for coefficients in spatiotemporal subbands at two different temporal resolution levels.

One can observe the increase of the variance of the model with the conditioning value which leads to a double stochastic model, in which we consider the *conditional* probability law of the coefficients in a given subband to be Gaussian, with variance depending on the set of spatiotemporal neighbors. Figure 2 suggests considering the following model:

$$\log a_{n,m} = \log\left(l_{n,m} + \alpha\right) + z_{n,m}, \tag{2}$$

where $z_{n,m}$ is an additive noise. When $l_{n,m}$ takes large values, the dependence between $\log a_{n,m}$ and $\log l_{n,m}$ is approximately linear, which is in agreement with the right part of the plot in Figure 2. In the meantime, the constant α is useful to describe the flat left part of the log histogram. From the same figure, note also the consistency of the model over the temporal scales.

This model amounts to

$$\left| c_{n,m} \right| = \left(l_{n,m} + \alpha\right)^{1/2} e^{z_{n,m}/2} \tag{3}$$

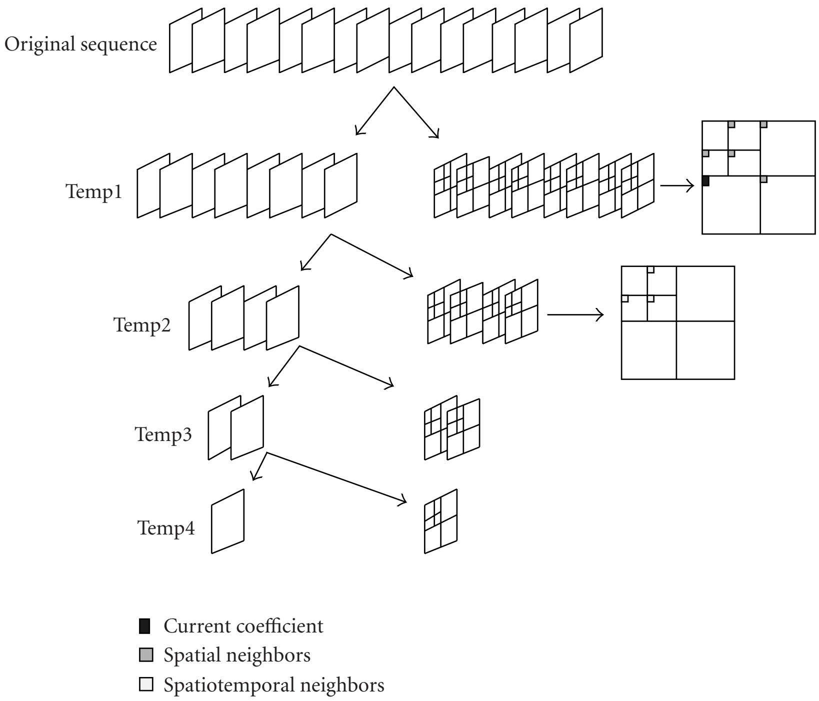


Figure 1: Spatiotemporal neighbors of a wavelet coefficient in a video sequence (the original group of frames (GOF) is decomposed over four temporal levels). Dependencies are highlighted with their spatial (parent, cousins, aunts) and spatiotemporal neighbors (temporal parent and temporal aunts). Temp*i* stands for the *i*th temporal decomposition level.

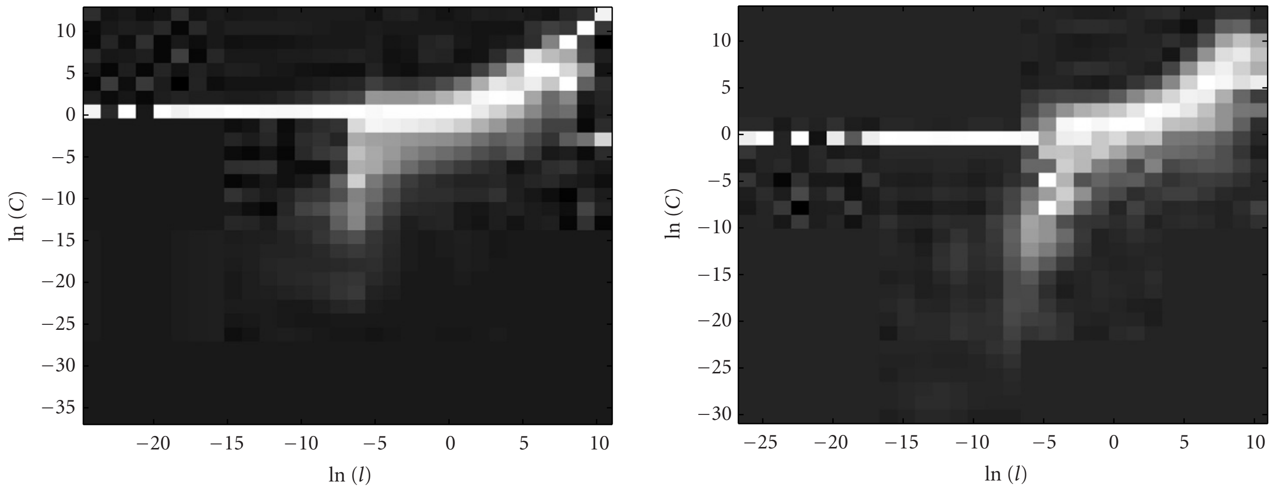


Figure 2: Log-log histogram of squared wavelet coefficients, conditioned to a linear prediction of squared spatiotemporal neighbors. Left: first temporal level. Right: second temporal level.

and by reintroducing the sign, we have

$$c_{n,m} = \left(l_{n,m} + \alpha\right)^{1/2} e^{z_{n,m}/2} s_{n,m}, \tag{4}$$

where $s_{n,m} \in \{-1, 1\}$. We suppose the noise to be normal, that is, $\beta_{n,m} = e^{z_{n,m}/2} s_{n,m} \sim \mathcal{N}(0, 1)$. This leads to a Gaussian conditional distribution for the spatiotemporal coefficients of the form

$$g\left(c_{n,m} \mid \sigma_{n,m}^2\right) = \frac{1}{\sqrt{2\pi}\sigma_{n,m}} e^{-c_{n,m}^2/2\sigma_{n,m}^2}, \tag{5}$$

where

$$\sigma_{n,m}^2 = \sum_k w_k \left| p_k(n, m) \right|^2 + \alpha \tag{6}$$

and $\mathbf{p}(n, m) = \left(p_k(n, m)\right)_k$ is the vector of neighbors.

3. MODEL ESTIMATION

In order to estimate the parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \mathbf{w} \\ \alpha \end{pmatrix} \tag{7}$$

of the model, we use the wavelet coefficients $(c_{n,m})_{(1\leq n\leq N, 1\leq m\leq M)}$ (where N, M represent the image size) to build several criteria and compare their estimation performances. Ideally, a criterion $J_{N,M}(\boldsymbol{\theta})$ should satisfy some nice properties, such as the following.

(1) A parameter estimator should be such that $\widehat{\boldsymbol{\theta}}_{N,M} = \arg\min_{\boldsymbol{\theta}} J_{N,M}(\boldsymbol{\theta})$.
(2) $J_{N,M}(\boldsymbol{\theta}) \to J(\boldsymbol{\theta})$ when $N, M \to \infty$, the convergence being almost sure (or, at least, in probability).
(3) $J(\boldsymbol{\theta}) \geq J(\boldsymbol{\theta_0})$, with $\boldsymbol{\theta_0}$ being the vector of the true parameters;

These conditions define what is called a "contrast" in statistics [27]. However, they may be difficult to satisfy in practice, and one can therefore require slightly weaker constraints to be satisfied. In the sequel, we will check whether the following two alternative constraints are satisfied by the proposed criteria:

$$\mathrm{E}\{J_{N,M}(\boldsymbol{\theta})\} \geq \mathrm{E}\{J_{N,M}(\boldsymbol{\theta_0})\}, \tag{8}$$

$$J_{N,M}(\boldsymbol{\theta_0}) \longrightarrow J(\boldsymbol{\theta_0}) \quad \text{in probability, when } N, M \longrightarrow \infty. \tag{9}$$

In the above equation, $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. We now introduce the criteria and discuss their properties with respect to the above constraints.

(1) *Least squares (LS)*. The criterion proposed in [22] is a least mean squares one, which can be written as

$$J_{N,M}(\boldsymbol{\theta}) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left(c_{n,m}^2 - \sigma_{n,m}^2(\boldsymbol{\theta})\right)^2. \tag{10}$$

For the probability law of the coefficients given by (5) and (6), it can be easily shown that this criterion satisfies relation (8) (with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta_0}$), but condition (9) holds only subject to some additional ergodicity conditions on $c_{n,m}^2 - \sigma_{n,m}^2(\boldsymbol{\theta_0})$.

(2) *Maximum likelihood (ML)*. We propose the use of an approximate ML estimator:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{n=1}^{N}\prod_{m=1}^{M} g\left(c_{n,m} \mid \sigma_{n,m}^2(\boldsymbol{\theta})\right). \tag{11}$$

This amounts to minimizing the following criterion:

$$J_{N,M}(\boldsymbol{\theta}) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\{\frac{c_{n,m}^2}{\sigma_{n,m}^2(\boldsymbol{\theta})} + \log\sigma_{n,m}^2(\boldsymbol{\theta})\right\}. \tag{12}$$

Again, it is easy to verify that this criterion satisfies relation (8) (with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta_0}$) for the conditional law of interest, but condition (9) requires ergodicity conditions on $\log\sigma_{n,m}^2(\boldsymbol{\theta_0})$.

(3) Looking for a criterion satisfying (9), we introduce *a more efficient criterion (EC)*, defined by

$$J_{N,M}(\boldsymbol{\theta}) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\frac{\gamma\left|c_{n,m}\right|^{\beta}}{\sigma_{n,m}^{\beta}(\boldsymbol{\theta})} - 1\right)^2, \tag{13}$$

where γ and β are two positive real parameters.

For a very large number of coefficients ($N, M \to \infty$), according to the law of large numbers, the criterion $J_{N,M}(\boldsymbol{\theta_0})$ converges in probability to the following expression:

$$J(\boldsymbol{\theta_0}) = \gamma^2\,\mathrm{E}\left\{\frac{\left|c_{n,m}\right|^{2\beta}}{\sigma_{n,m}^{2\beta}(\boldsymbol{\theta_0})}\right\} - 2\gamma\,\mathrm{E}\left\{\frac{\left|c_{n,m}\right|^{\beta}}{\sigma_{n,m}^{\beta}(\boldsymbol{\theta_0})}\right\} + 1. \tag{14}$$

Besides, we have $\mathrm{E}\{|c_{n,m}|^{\beta}|\mathbf{p}(n,m)\} = C_{\beta}^{c}\sigma_{n,m}^{\beta}(\boldsymbol{\theta_0})$, where

$$C_{\beta}^{c} = 2\int_{0}^{\infty} u^{\beta} g(u|1)\,du. \tag{15}$$

Expression (13) thus leads to

$$\begin{aligned}&\mathrm{E}\left\{J_{N,M}(\boldsymbol{\theta}) \mid (\mathbf{p}(n,m))_{1\leq n\leq N,\,1\leq m\leq M}\right\} \\ &\quad = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[\gamma^2 C_{2\beta}^{c}\frac{\sigma_{n,m}^{2\beta}(\boldsymbol{\theta_0})}{\sigma_{n,m}^{2\beta}(\boldsymbol{\theta})} - 2\gamma C_{\beta}^{c}\frac{\sigma_{n,m}^{\beta}(\boldsymbol{\theta_0})}{\sigma_{n,m}^{\beta}(\boldsymbol{\theta})} + 1\right].\end{aligned} \tag{16}$$

The parameter γ should be chosen so as to guarantee that $\mathrm{E}\{J_{N,M}(\boldsymbol{\theta_0})\} \leq \mathrm{E}\{J_{N,M}(\boldsymbol{\theta})\}$ for all $\boldsymbol{\theta}$, with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta_0}$. This condition is satisfied if

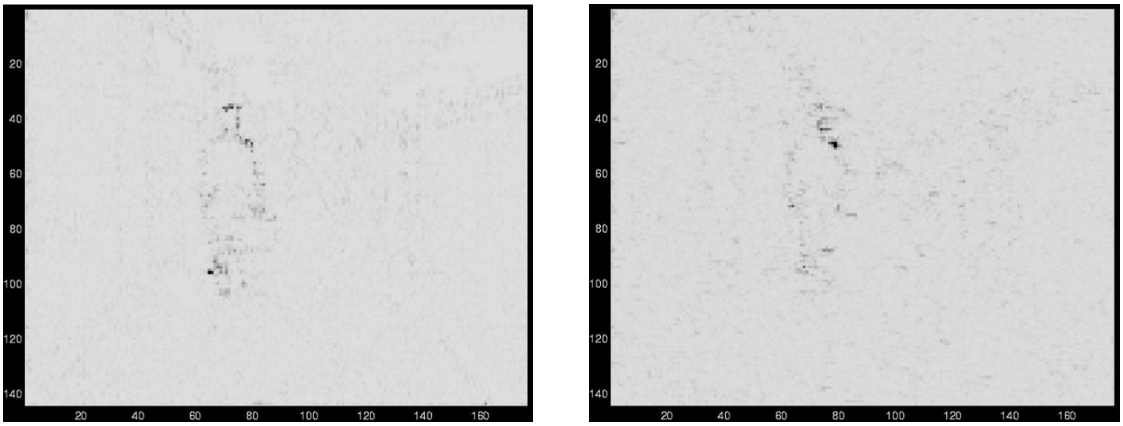
$$\boldsymbol{\theta} \longmapsto \gamma^2 C_{2\beta}^{c}\frac{\sigma_{n,m}^{2\beta}(\boldsymbol{\theta_0})}{\sigma_{n,m}^{2\beta}(\boldsymbol{\theta})} - 2\gamma C_{\beta}^{c}\frac{\sigma_{n,m}^{\beta}(\boldsymbol{\theta_0})}{\sigma_{n,m}^{\beta}(\boldsymbol{\theta})} + 1 \tag{17}$$

is minimum for $\boldsymbol{\theta} = \boldsymbol{\theta_0}$. After some simple calculations, it can be shown that by choosing $\gamma = C_{\beta}^{c}/C_{2\beta}^{c}$, the above property is satisfied.

We can notice that due to the Gaussian assumption in the particular case $\beta = 2$, we get $C_2^c = 1$, $C_4^c = 3$. In this case, the criterion in (16) is equivalent to a *modified least squares (MLS)* criterion, leading to the following minimization:

$$\sum_{n,m}\left(\frac{c_{n,m}^4}{3\sigma_{n,m}(\boldsymbol{\theta})^4} - 2\frac{c_{n,m}^2}{\sigma_{n,m}(\boldsymbol{\theta})^2}\right). \tag{18}$$

One of the advantages of the third criterion (EC) over the former two (LS, ML) is that no additional ergodicity conditions are required for (9) to be satisfied. In the next section, we provide evidence through Monte Carlo simulations for the improved mean square estimation error achieved by the new criterion.



(a)

(b)

Figure 3: (a) Vertical detail subband at the highest spatial resolution of the first temporal decomposition level for "hall monitor" sequence. (b) Simulated subband, using the conditional law given in (5).

Table 1: Parameter estimation: the first column indicates the various spatiotemporal neighbors whose weights are estimated (see Figure 1). The second column indicates the value of the true parameters; the next four the MSE of the estimation by the four proposed methods over 50 realizations.

Spatiotemporal neighbors	Model parameters	LS	ML	EC ($\beta = 2$)	EC ($\beta = 1$)
wUp	0.1008	0.0670	0.0405	0.0315	0.0369
wLeft	0.2940	0.1158	0.0272	0.0266	0.0318
wcous1	0.0389	0.0331	0.0132	0.0249	0.0141
wcous2	0.2121	0.1487	0.0487	0.0326	0.0496
wpar	0.0073	0.0350	0.0070	0.0101	0.0071
waunt1	0.0358	0.0230	0.0070	0.0106	0.0080
waunt2	0.0685	0.0273	0.0048	0.0067	0.0045
wpartm	0.0341	0.0147	0.0042	0.0045	0.0051
wLeftpartm	0.0012	0.0079	0.0022	0.0032	0.0033
wUppartm	0.0054	0.0065	0.0048	0.0056	0.0049
waunt1tm	0.0013	0.0034	0.0024	0.0058	0.0021
waunt2tm	0.0002	0.0092	0.0011	0.0009	0.0008
α	0.3663	0.5121	0.0975	0.0465	0.0774

3.1. Illustration examples

In order to illustrate the previous theoretical results, we consider a lifting-based motion-compensated temporal Haar decomposition [3] of a video sequence, applied on groups of 16 frames, with 4 temporal and 4 spatial resolution levels. The motion estimation/compensation in the Haar temporal decomposition uses a full search block matching algorithm with half-pel motion accuracy and the spatial multiresolution analysis (MRA) is based on the biorthogonal 9/7 filters. The spatiotemporal neighborhood consists of 12 coefficients of the current one: its Up and Left neighbors, its spatial parent, aunts, and cousins, and its spatiotemporal parent together with its Up and Left neighbors and spatiotemporal aunts. In order to check the validity of our model, the parameters estimated by least mean squares on a given subband have been used to generate a Gaussian random field having the same conditional probability density as our model. The real subband (which is, in this case, the vertical detail subband at the highest spatial resolution of the first temporal decomposition level for "hall monitor" sequence) and a typical simulated one (with the parameters estimated by MLS criterion) are shown in Figure 3. Based on the synthetic data, the different estimators presented in Section 3 have been compared and the parameter values estimated over 50 realizations are presented in Table 1. A critical point in the estimation is that in order to keep the variance of the model positive, we need to constrain the weights to be positive. As we can notice from this table, the EC with $\beta = 2$ proves to be the most robust and of the best performance compared to the LS and ML criteria especially for the neighbors which are more significant.

In the second part of this paper, we introduce a prediction method based on our stochastic model before presenting two applications of it: the quality improvement of scalable video and error concealment when packet losses occur during video transmission.

4. PREDICTION STRATEGY

In a packet network without QoS (quality of service), even considering a strong channel protection for the most important parts of the bit stream, some of the packets will be lost during the transmission due to network congestion or bursts of error. In this case, an error concealment method should be applied by the decoder in order to improve the quality of the reconstructed sequence.

The stochastic model presented in the previous sections can be applied to the prediction of the subbands that are not received by the decoder. Indeed, a spatiotemporal MRA as described in Section 2 naturally provides a hierarchical subband structure, allowing to transmit information by decreasing order of importance. The decoder receives, therefore, the coarser spatiotemporal resolution levels first and then, with the help of the spatiotemporal neighbors, can predict the finest resolution ones.

The conditional law of the coefficients exhibited in (5) is used to build an optimal mean square error (MSE) estimator of the magnitude of each coefficient, given its spatiotemporal ancestors. This leads to the following predictor:

$$\begin{aligned} |\hat{c}_{n,m}| &= \mathrm{E}\left\{|c_{n,m}| \,|\, \mathbf{p}(n,m)\right\} \\ &= \int_{-\infty}^{\infty} |c_{n,m}|\, g\left(c_{n,m} \,|\, \sigma^2_{n,m}\right) dc_{n,m}. \end{aligned} \tag{19}$$

After some simple calculations, we get the optimal estimator expression:

$$|\hat{c}_{n,m}| = \sqrt{\frac{2}{\pi}}\sigma_{n,m}, \tag{20}$$

with $\sigma_{n,m}$ given in (6) and the model parameters estimated using the criterion in (14).

The choice of the spatiotemporal neighbors used by the predictor, in the context of a scalable bit stream, has been made in such a way as to avoid error propagation. Supposing the coarser spatial level of each frame is received (e.g., it can be better protected against channel errors), we restrict the choice of the coefficients $p_k(n,m)$ in our model to the spatial parent, spatial aunts, and the spatiotemporal parent, its neighbors, and the spatiotemporal aunts of the current coefficient. As the bit stream is resolution scalable, all these spatiotemporal ancestors belong to the spatiotemporal subbands that have already been received by the decoder and can therefore be used in a causal prediction.

Note that our statistical model and therefore the proposed prediction do not take into account the sign of the coefficients. As the sign of the coefficients remains an important piece of information, data partitioning can be used to separate it from the magnitude of the coefficients, in order to better protect it in the video bit stream. Efficient algorithms for encoding the sign of wavelet coefficients are already available (see, e.g., [28]). In the sequel, we will consider therefore that the sign has been correctly decoded.

5. MODEL-BASED QUALITY ENHANCEMENT OF SCALABLE VIDEO

In the first application, we consider scalable video transmission over heterogeneous networks and we are interested in improving the spatial scalability properties. In this case, the adaptation of the bit stream to the available bandwidth can lead to discarding the finest spatial detail subbands during the transmission. However, if the decoder has display size and CPU capacity to decode in full resolution, the lack of the finest frequency details would result in a low-quality, oversmoothed, reconstructed sequence. We propose to use the stochastic model developed in Section 2 to improve the rendering of the spatiotemporal details in the reconstructed sequence. Thus, the decoder will receive the coarser spatial resolution levels at each temporal level and predict with the help of the spatiotemporal neighbors the finest resolution ones. We propose to use, for the prediction, the optimal MSE estimator of the magnitude of each coefficient, given its spatiotemporal ancestors presented in Section 4.

Note that this strategy can also be seen as a quality scalability, since bit rate reduction is achieved by not transmitting the finest frequency details.

In order to apply this method, as we can recall from the Table 1, it is more convenient to use the EC criterion with $\beta = 1$. Its performance in the considered neighborhood is better than that of the same criterion with $\beta = 2$.

For simulations, we have considered the spatiotemporal neighborhood consisting of the 8 coefficients mentioned above. We send the three low-resolution spatial levels of each temporal detail frame and predict the highest resolution detail subbands using our model. We compare this procedure with the reconstruction of the full resolution using the finest spatial detail subbands set to zero, which would be the reconstruction strategy of a simpler decoder.

In Figure 4, we present the MSE of the spatial reconstruction of each temporal detail frame at different temporal resolution levels. One can observe the significant decrease in reconstruction error by using the proposed prediction strategy. Another observation is related to the MSE value in itself, which is highest at the last temporal resolution level. This is related to the higher energy of the low-resolution temporal detail subbands.

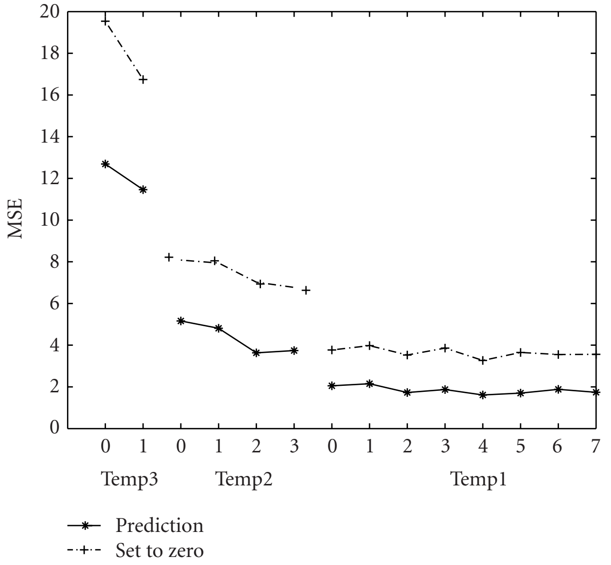


Figure 4: MSE of the spatial reconstruction of the detail frames at each temporal resolution level in a GOF.

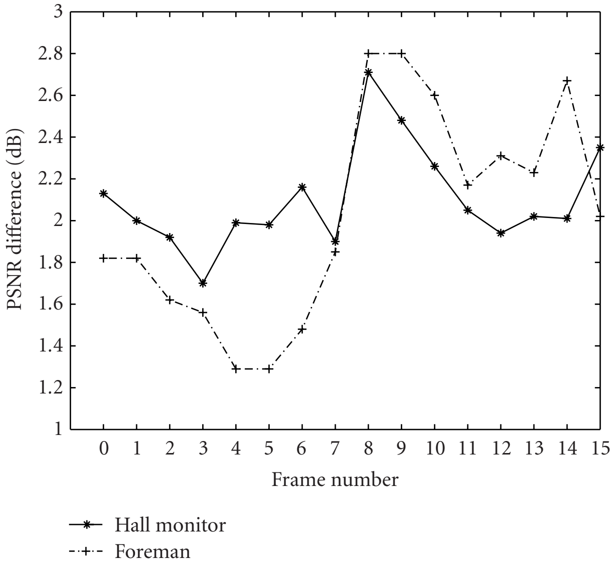
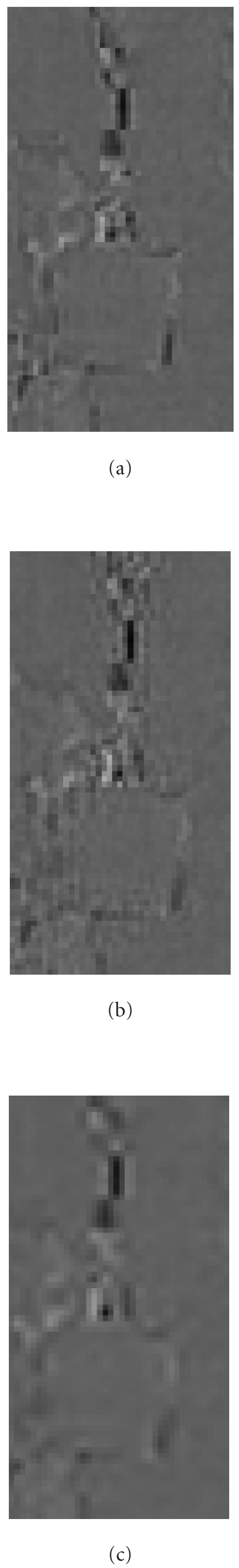


Figure 5: PSNR improvement for a GOF of 16 frames of the "foreman" and "hall monitor" CIF sequences, when we predict the *finest frequency* subbands at different temporal resolution levels.

In Figure 5, we present the PSNR improvement of the reconstructed sequence obtained by predicting the finest frequency subbands at all the temporal resolution levels with our model, instead of setting them to zero. As we can see, for two different sequences, the PSNR improvement varies between 1.3 dB and 2.7 dB.

In Figure 6, we present the reconstructed temporal detail frames of the first temporal resolution level of the "hall monitor" sequence. (a) Is the real reconstructed temporal detail



(a)

(b)

(c)

Figure 6: Zoom in a temporal detail frame at the first temporal resolution level. (a) Original frame. (b) Reconstructed detail frame when we predict its *finest resolution* subbands. (c) Reconstructed frame when we set them to zero.

frame, (b) is the reconstructed frame when we predict the finest subbands, and (c) is the reconstructed frame when we set its finest subbands to zero. As we can see, the third image proved to be more blurred than the real one and the frame reconstructed with the help of our model has sharper edges and outlines.

6. ERROR CONCEALMENT IN THE SPATIOTEMPORAL WAVELET DOMAIN

The application we consider in this section is the transmission of scalable video bit stream over IP networks, prone to packet losses. The packetization strategy will highly influence the error concealment methods that we need to apply. Indeed, depending on the application and on the level of protection desired (and the overhead allowed for error protection), several strategies of packetization can be envisaged for the spatiotemporal coefficients, such as:

(1) one spatial subband per packet;
(2) all subbands with the same spatial resolution and orientation in one packet;
(3) all subbands at the same spatial level in each temporal detail frame in one packet.

We further analyze the influence of losing a packet at different spatiotemporal levels in each one of these settings and the ability of our prediction model to provide error concealment.

(1) First, we analyze the concealment ability of our model when the packetization method consists of taking one subband per packet. In this case, if a spatiotemporal subband is lost, we predict it with the help of the neighbors of the coarser spatial and temporal resolution levels that we assume have been received by the decoder without losses. In Figure 7, we present the MSE of the reconstruction of a detail frame when we lose a subband at different spatial and temporal resolution levels. The MSE of the reconstructed frames using the prediction based on the statistical model is better than the one obtained by setting to zero the coefficients corresponding to the lost subband. Note also that, as expected, the loss of a subband at the last temporal resolution level influences the MSE of the reconstructed frame more than at any other temporal level.

An interesting point that comes out from these results is that the spectral behavior of a temporal detail frame is different from that of still images. One can see, from Figure 7, that the energy of the subbands at different spatial resolution levels does not decay across the scales, as observed for still images, but the medium and high frequency levels have more power than the lowest frequency one. This is due to the fact that the frames we are studying represent temporal prediction errors, therefore containing spatial patterns very similar to edges, whose energy is concentrated at rather high spatial frequencies.

Another useful point is to see how the prediction of a subband at different spatial resolution levels influences the reconstruction of a frame in the original sequence. Thus, in Table 2, we present the MSE of the reconstruction of a frame in the original sequence when we lose a subband of a temporal detail frame at a given temporal resolution level (numbered 1, 2, 3) and at each spatial resolution level (denoted by Tables 1, 2, 3).

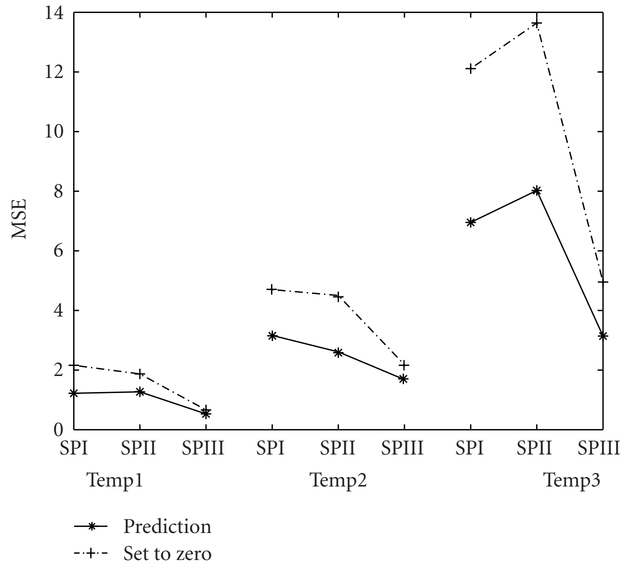


FIGURE 7: MSE of the spatial reconstruction of the first temporal detail frame on losing the horizontal subband at different temporal resolution levels and for three spatial resolution levels. SP*i* stands for the *i*th spatial resolution level and Temp*i* for the *i*th temporal decomposition level.

In this case, the reconstruction quality using our optimal predictor is proved to be superior to the reconstruction performed with the details corresponding to the lost subbands set to zero. We also notice that, as expected, the loss of a subband at the third temporal resolution level is more damaging for the reconstruction than at another temporal level.

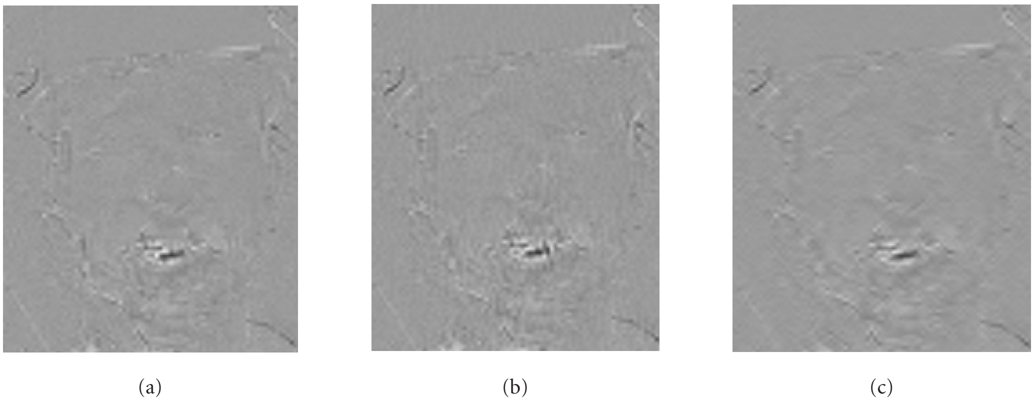
In Figure 8, we show a detail of a reconstructed frame at the first temporal resolution level, assuming that a subband at the second spatial resolution level was lost.

(2) Next, we consider the packetization technique in which all the subbands of the same spatial resolution and orientation level at the same temporal resolution level belong to a packet. In Figure 9, we present the PSNR improvement of the reconstructed sequence assuming that we lose a packet at each temporal level. We notice here that our model leads to a higher improvement of the PSNR (up to 2.5 dB) when the lost packet is at the first temporal resolution level, where the prediction errors do not propagate through the temporal synthesis procedure.

(3) The third method of packetization considered consists of taking the subbands of the same spatial resolution level in each temporal detail frame in one packet. In Table 3, we present the MSE of a reconstructed frame of the original sequence in case we lose a spatial resolution level (first or second) of a temporal detail frame at different temporal resolution levels. We observe that as we move to coarser temporal

Table 2: MSE of the reconstruction of the first frame of the original sequence, when prediction of a subband at different spatial resolution levels and at different temporal resolution levels is used, compared with setting to zero the lost coefficients.

	Temp1			Temp2			Temp3		
	SPI	SPII	SPIII	SPI	SPII	SPIII	SPI	SPII	SPIII
Set to zero	0.93	0.81	0.28	0.86	0.89	0.42	1.14	1.32	0.49
Prediction	0.53	0.56	0.22	0.63	0.51	0.33	0.68	0.77	0.31



(a) (b) (c)

Figure 8: First temporal detail frame at the first temporal resolution level. (a) Original frame. (b) Reconstructed detail frame when we predict the lost horizontal subband of the *second spatial resolution level*. (c) Reconstructed detail frame when we set it to zero.

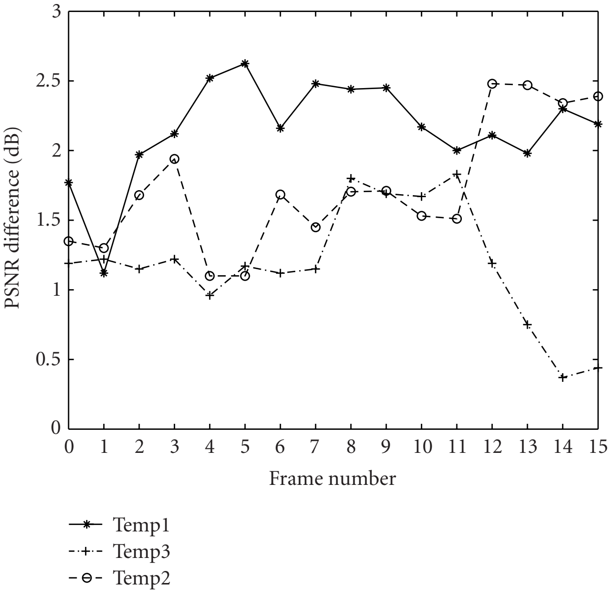


Figure 9: PSNR improvement (prediction versus setting to zero) of a reconstructed GOF of the original sequence "foreman" in CIF format, 30 fps, when we lose the horizontal subbands of the *second spatial resolution level* at each temporal resolution level.

resolution levels, the loss of the coarser spatial resolution level becomes more significant. This could be expected, as the coefficients of a coarser temporal and spatial resolution level are bigger than those of a finer one and so even a small error at the prediction becomes important in the reconstruction of the original frames.

Figure 10 compares the reconstruction of a temporal detail frame with the proposed method with the one that consists of setting to zero the coefficients corresponding to the lost packet. We observe the oversmoothing resulting from the latter method and the good visual rendering of the high frequency details obtained with the proposed method.

7. ERROR CONCEALMENT OF SCALABLE BITSTREAMS

In the previous simulation results, we have assumed that, except for the lost packet, all the other subbands have been correctly received by the decoder. Here, we consider an even worse scenario: bandwidth reduction during the transmission requires to cut from the bit stream the finest detail subbands, and, in addition, some packets are lost from the remaining bit stream. The main difference from the previous situation is that we need to predict not only the lost packet, but also the finest spatial resolution level. Some of the subbands in this level will be predicted based on spatiotemporal neighbors that also result from a prediction. As this procedure inherently introduces a higher error, we show by simulation results that the reconstruction of the full resolution video sequence has better quality than what we can obtaine by a "naïve" decoder (which, as in the previous section, would set to zero all the unknown coefficients).

We next examine the error concealment ability of our model in the same three packetization strategies as in Section 6.

Table 3: MSE of a reconstructed frame of the original sequence when we lose the first or second spatial resolution level of a temporal detail frame at each temporal resolution level.

	Temp1		Temp2		Temp3	
	SPI	SPII	SPI	SPII	SPI	SPII
Set to zero	1.59	1.12	1.52	1.27	2.52	2.58
Prediction	0.85	0.83	1.02	1.15	1.74	1.97

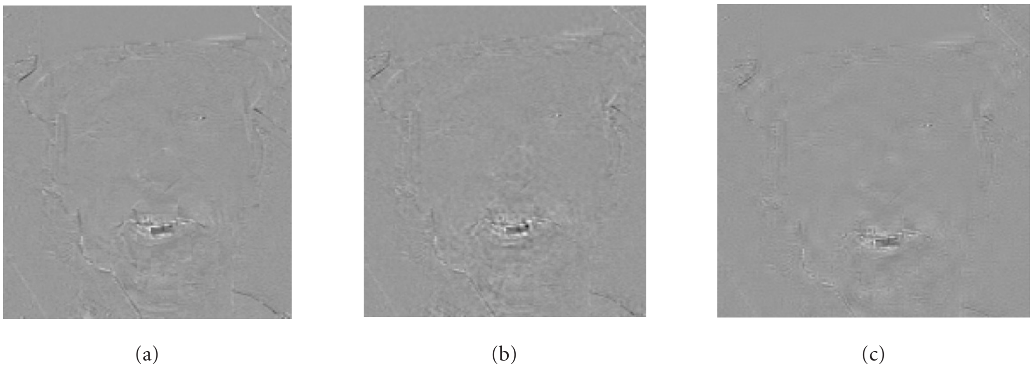


Figure 10: Temporal detail frame at the first temporal resolution level. (a) Original frame. (b) Reconstructed frame obtained by predicting the (lost) *second spatial resolution level.* (c) Reconstructed frame when we set to zero the details corresponding to the lost packet.

(1) For the first packetization strategy (one subband per packet), the MSE of a frame in the original sequence when we lose a subband at the second spatial resolution level at different temporal resolution levels is computed. The difference in MSE when using our prediction method compared with the “naive” decoder is about 1 at the first temporal resolution level and about 1.5 for the second and the third temporal resolution level. This variation can be explained by the fact that the loss of a subband at the second and third spatial resolution levels influences more the reconstruction of the original frame as this loss affects the spatial neighbors used in the reconstruction of the finest spatial subbands at the same temporal resolution level and also the spatiotemporal neighbors of the finest spatial subbands at the next finer temporal resolution level.

(2) In the second case (a packet includes all the subbands of the same orientation and spatial resolution level, at the same temporal resolution level), Figure 11 illustrates the PSNR improvement of the original sequence in case we predict the finest resolution subbands after having predicted a lost packet at a coarser resolution level, compared to the case where all these lost subbands are set to zero.

The higher improvement of the PSNR at the finest temporal resolution level is due to the fact that in this case, the loss of the packet influences only the reconstruction of the temporal detail frame at this temporal resolution level. On the contrary, a loss at any other temporal level influences also the prediction of the subbands at finer temporal resolution.

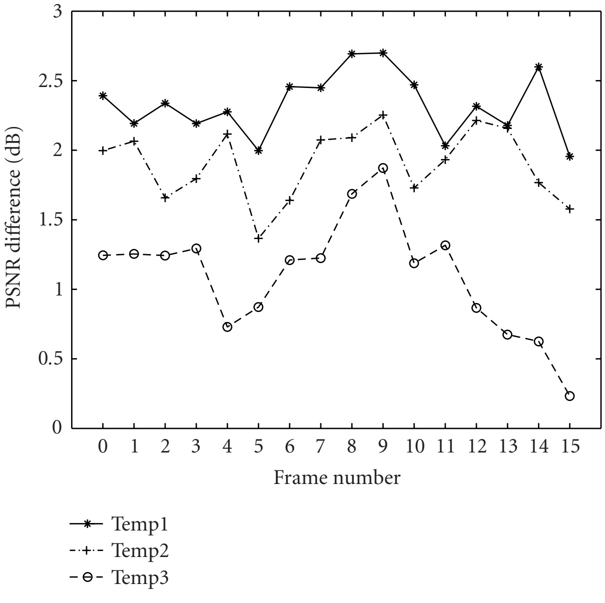


Figure 11: Improvement of the PSNR of the reconstructed GOF of the original sequence “foreman” when we lose the horizontal subbands of the second spatial resolution level at each temporal resolution level and we predict them and the finest spatial resolution subbands.

(3) At the end, we examine the third packetization technique (a packet includes all the subbands at a given spatial resolution level for each temporal detail frame). Table 4 illustrates the MSE when, in the reconstruction of the original sequence, we predict the lost second spatial resolution level as well as the finest ones compared to the case where both of these spatial resolution levels are considered to be lost and set to zero. We remark that even in the case where we lose the whole second spatial resolution level, our model is able to

Table 4: MSE of the reconstructed first frame of a GOF of original sequence “foreman” when we lose the second spatial resolution level of the first temporal detail frame at each temporal resolution level.

	Temp1	Temp2	Temp3
Set to zero	3.50	4.34	6.63
Prediction	3.20	3.61	5.95

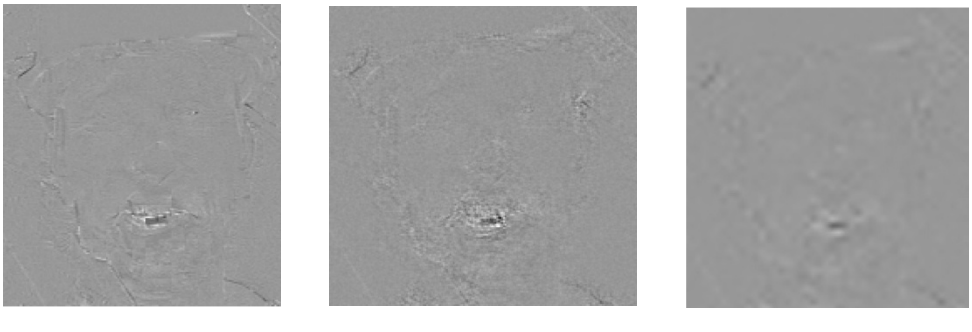


Figure 12: First temporal detail frame at the first temporal resolution level. (a) Original subband. (b) Reconstructed frame when we predict the second and then the first spatial resolution level. (c) Reconstructed frame when we set lost subbands to zero.

successfully predict it from the received subbands and, based on this, to predict also the finer spatial resolution level.

Figure 12 shows the reconstructed images when we lose the second spatial resolution level of a temporal detail frame at the first temporal resolution level. Compared to Figure 10, the frame obtained using the prediction method keeps almost the same amount of details, while the image obtained by setting to zero all the lost subbands suffered an even worse degradation.

8. CONCLUSION

In this paper, we have first presented a statistical model for the spatiotemporal coefficients of a motion-compensated wavelet decomposition of a video sequence. We have deduced an optimal MSE predictor for the lost coefficients and used these theoretical results in two applications to scalable video transmission over packet networks. In the first application, we have shown significant quality improvement achieved by this technique in spatiotemporal resolution enhancement. In the second one, we have proved the error concealment properties conferred by our stochastic model on a scalable video bit stream, under different packet loss conditions and with different packetization strategies. Our future work concerns the study of sign prediction methods of the wavelet coefficients in 2D + t decompositions of video sequences.

ACKNOWLEDGMENT

Part of this work has been presented at the NSIP workshop, June 2003 [26].

REFERENCES

[1] S.-J. Choi and J. W. Woods, “Motion-compensated 3-D subband coding of video,” *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155–167, 1999.

[2] S.-T. Hsiang and J. W. Woods, “Invertible three-dimensional analysis/synthesis system for video coding with half-pixel-accurate motion compensation,” in *Proc. SPIE Conference on Visual Communications and Image Processing (VCIP ’99)*, vol. 3653, pp. 537–546, San Jose, Calif, USA, January 1999.

[3] B. Pesquet-Popescu and V. Bottreau, “Three-dimensional lifting schemes for motion compensated video compression,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP ’01)*, vol. 3, pp. 1793–1796, Salt Lake City, Utah, USA, May 2001.

[4] D. Turaga and M. van der Schaar, “Unconstrained temporal scalability with multiple reference and bi-directional motion compensated temporal filtering,” doc. m8388, MPEG meeting, Fairfax, Va, USA, November 2002.

[5] J. R. Ohm, “Complexity and delay analysis of MCTF interframe wavelet structures,” doc. m8520, MPEG meeting, Klagenfurt, Austria, July 2002.

[6] J. W. Woods, P. Chen, and S.-T. Hsiang, “Exploration experimental results and software,” doc. m8524, MPEG meeting, Shanghai, China, October 2002.

[7] S. Shirani, F. Kossentini, and R. Ward, “Error concealment methods, a comparative study,” in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering (CCECE ’99)*, vol. 2, pp. 835–840, Edmonton, Alta, Canada, May 1999.

[8] R. Talluri, “Error-resilient video coding in the ISO MPEG-4 standard,” *IEEE Communications Magazine*, vol. 36, no. 6, pp. 112–119, 1998.

[9] M. Ghanbari and V. Seferidis, “Cell-loss concealment in ATM video codecs,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 3, no. 3, pp. 238–247, 1993.

[10] W. M. Lam, A. R. Reibman, and B. Liu, “Recovery of lost or erroneously received motion vectors,” in *Proc.*

IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '93), vol. 5, pp. 417–420, Minneapolis, Minn, USA, April 1993.

[11] Y. Wang, Q.-F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *IEEE Trans. Communications*, vol. 41, no. 10, pp. 1544–1551, 1993.

[12] S. S. Hemami and T. H.-Y. Meng, "Transform coded image reconstruction exploiting interblock correlation," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 1023–1027, 1995.

[13] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Trans. Image Processing*, vol. 4, no. 4, pp. 470–477, 1995.

[14] P. Salama, N. B. Shroff, and E. J. Delp, "Error concealment in encoded video streams," in *Signal Recovery Techniques for Image and Video Compression and Transmission*, N. P. Galatsanos and A. K. Katsaggelos, Eds., pp. 199–234, Kluwer Academic, Boston, Mass, USA, 1998.

[15] S. Shirani, F. Kossentini, and R. Ward, "A concealment method for video communications in an error-prone environment," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1122–1128, 2000.

[16] Y. Zhang and K.-K. Ma, "Error concealment for video transmission with dual multiscale Markov random field modeling," *IEEE Trans. Image Processing*, vol. 12, no. 2, pp. 236–242, 2003.

[17] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Processing*, vol. 10, no. 11, pp. 1647–1658, 2001.

[18] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, 1993.

[19] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 6, no. 3, pp. 243–250, 1996.

[20] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Processing*, vol. 50, no. 11, pp. 2744–2756, 2002.

[21] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Interference in Wavelet Based Models*, vol. 141 of *Lecture Notes in Statistics*, pp. 291–308, Springer-Verlag, New York, NY, USA, 1999.

[22] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Wavelet Applications in Signal and Image Processing VII*, vol. 3813, pp. 188–195, Denver, Colo, USA, July 1999.

[23] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Processing*, vol. 8, no. 12, pp. 1688–1701, 1999.

[24] B.-J. Kim, Z. Xiong, and W. A. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 8, pp. 1374–1387, 2000.

[25] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 1056–1068, 2001.

[26] G. Feideropoulou, B. Pesquet-Popescu, J. C. Belfiore, and G. Rodriguez, "Non-linear modelling of wavelet coefficients for a video sequence," in *Proc. IEEE Workshop on Nonlinear Signal and Image Processing (NSIP '03)*, Grado, Italy, June 2003.

[27] C. Gourieroux and A. Monfort, *Statistics and Econometric Models*, Cambridge University Press, New York, NY, USA, 1995.

[28] A. Deever and S. Hemami, "Efficient sign coding and estimation of zero-quantized coefficients in embedded wavelet image codecs," *IEEE Trans. Image Processing*, vol. 12, no. 4, pp. 420–430, 2003.

Georgia Feideropoulou received the M.S. degree in electronic and computer engineering from the Technical University of Crete, Greece, in 2000, and the DEA (Diplome d'Etudes Approfondies) degree in telecommunication systems from the École Nationale Supérieure des Télécommunications (ENST) in Paris in 2001. She is currently pursuing the Ph.D. degree in video processing at the ENST in Paris. Her research interests include video compression and joint source-channel coding.

Béatrice Pesquet-Popescu received the M.S. degree in telecommunications from the "Politehnica" Institute in Bucharest in 1995 and the Ph.D. degree from the École Normale Supérieure de Cachan in 1998. In 1998, she was a Research and Teaching Assistant at Université Paris XI, and in 1999, she joined Philips Research France, where she worked for two years as a Research Scientist in scalable video coding. Since October 2000, she is an Associate Professor in multimedia at the École Nationale Supérieure des Télécommunications (ENST). EURASIP gave her a Best Student Paper Award in the IEEE Signal Processing Workshop on Higher-Order Statistics in 1997, and in 1998, she received a Young Investigator Award granted by the French Physical Society. She holds 19 patents in the area of wavelet-based video coding. Her current research interests are in scalable video coding, multimedia applications, and statistical image analysis.