INTER-RESOLUTION TRANSFORM FOR SPATIALLY SCALABLE VIDEO CODING

Markus Flierl and Pierre Vandergheynst

Signal Processing Institute
Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland
{markus.flierl,pierre.vandergheynst}@epfl.ch

ABSTRACT

Spatial scalability of video signals can be achieved with critically sampled spatial wavelet schemes but also with an overcomplete spatial representation. Critically sampled schemes struggle with the problem that critically sampled high-bands are shift-variant. Therefore, efficient motion compensation is challenging. On the other hand, overcomplete representations can be shift-invariant, thus permitting efficient motion compensation in the spatial subbands, but they have to be designed carefully to achieve high compression efficiency. This paper discusses an orthonormal transform for decomposing two different spatial scales of the same image. The transform is such that it minimizes the impact of the quantization noise on the reconstructed video signal at the decoder. Further, we investigate the decorrelation property of the transform. Finally, we compare to the compression efficiency of a Laplacian pyramid, a conventional scheme for an overcomplete representation of images, and observe coding gains up to 1 dB.

1. INTRODUCTION

Rate-distortion efficient coding of image sequences can be accomplished with motion-compensated temporal transforms [1, 2, 3, 4, 5]. Employing the temporal transform directly to the images of the sequence may be too limiting for targeted scalability properties of video representations. In particular, desirable video coding schemes should provide efficient spatial scalability of the video signal. If a motion-compensated temporal transform is utilized, it is favorable to employ this transform to the spatial subbands of the input images [6]. Such an architecture achieves good spatial scalability but is burdened with a degradation in rate-distortion performance. This burden is rooted in the fact that spatial decompositions utilize either critically sampled representations or overcomplete representations of the spatial subbands. Critically sampled representations lack the property of shift-invariance which seems to be crucial for efficient motion compensation. On the other hand, overcomplete representations can be shift-invariant, but rate-distortion efficient encoding is challenging.

This paper discusses a coding scheme with spatial scalability properties that can be interpreted as an extension of the spatial scalability concept as it is known from, e.g., the video coding standard ITU-T Rec. H.263 [7]: The pictures of the spatial base layer are spatially up-sampled in order to obtain pictures with the same spatial resolution as the pictures of the next spatial enhancement layer. These up-sampled pictures are used to predict the pictures of the next spatial enhancement layer. But this spatial prediction is just one step in our inter-resolution transform which requires also a spatial update step. The spatial update step will provide the desired orthogonality that spatial prediction is not capable of.

Our coding scheme decomposes spatially the input pictures of various sizes into spatial subbands. This spatial decomposition is based on a transform that is orthonormal for the spatial frequencies of the lower resolution. Depending on the chosen filter, the decorrelation properties of the inter-resolution transform can be close to optimal permitting an efficient rate-distortion performance. The orthogonality of the transform assures an efficient embedded representation of the image sequence at various resolutions. The spatial low-band can be critically sampled to reduce the encoding/decoding complexity of the spatial base layer. On the other hand, the spatial high-bands keep their shift-invariance property and permit efficient motion compensation.

2. INTER-RESOLUTION TRANSFORM

Assume that we have the k-th pictures of an image sequence in CIF resolution $\mathbf{s}_k^{(1)}$. For the lower resolution, we target a sequence of QCIF images. The pictures $\tilde{\mathbf{s}}_k^{(0)}$ in QCIF resolution are obtained by sub-sampling the corresponding ones in CIF resolution by 2. Before sub-sampling, we employ the

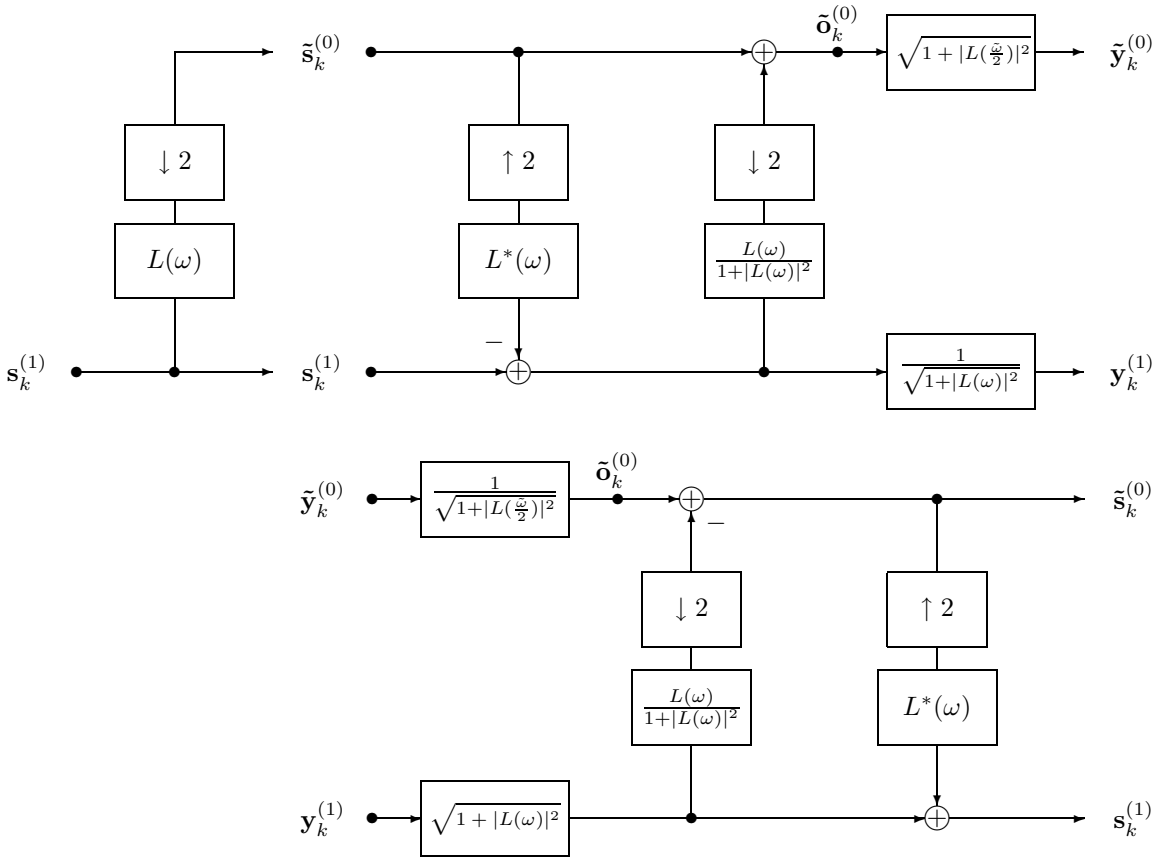

Fig. 1. **Top:** Inter-resolution Haar transform for the high-resolution image $\mathbf{s}_k^{(1)}$ and the low-resolution image $\tilde{\mathbf{s}}_k^{(0)}$ which is generated with the low-pass filter $L(\omega)$. The prediction step uses up-sampling and filtering whereas the update step utilizes filtering and down-sampling. **Bottom:** Inverse inter-resolution Haar transform for the high-band image $\mathbf{y}_k^{(1)}$ and the low-band image $\tilde{\mathbf{y}}_k^{(0)}$. It is an orthonormal transform for the spatial frequencies of the lower resolution with a low-resolution representation of the low-band image.

low-pass filter $L(\omega)$. Given these pictures in QCIF and CIF resolution, we employ the inter-resolution transform as depicted at the top of **Fig. 1** to obtain the spatial low-band $\tilde{\mathbf{y}}_k^{(0)}$ in QCIF resolution and the spatial high-band $\mathbf{y}_k^{(1)}$ in CIF resolution. A lifting implementation is used for the transform where the prediction step uses up-sampling and filtering, and the update step filtering and down-sampling. Finally, linear filters are used to normalize the transform. Depending on the filter $L(\omega)$, we achieve only approximately perfect decorrelation with the transform. We design an orthonormal transform to minimize the impact of the quantization noise on the reconstructed images in CIF resolution. At the decoder, we utilize the inverse inter-resolution transform as depicted at the bottom of **Fig. 1**. Note that the images $\tilde{\mathbf{o}}_k^{(0)}$ in QCIF resolution (**Fig. 1**) are chosen to represent the spatial base layer. With that, the spatial high-band can be dropped without degrading the spatial base layer. This is a desirable feature for spatially scalable video coding.

This inter-resolution transform is an extension of "upward prediction" as it is used in ITU-T Rec. H.263 to achieve spatial scalability. Upward prediction does not provide an orthogonal decomposition and the advantages of the inter-resolution transform are as follows: First, additional quantization noise due to SNR scalability has the least impact on the reconstructed video as the transform is orthonormal. Second, the features of spatial scalability can be carefully chosen at the encoder such that the decoder is able to reconstruct efficiently the desired spatial subresolutions. Third, accurate motion compensation is possible in all bands as we use a shift-invariant representation.

Note that our multiresolution representation for images is related to the Laplacian pyramid [8]. The basic idea of the Laplacian pyramid is the following: First, a coarse approximation of the original image is derived by low-pass filtering and down-sampling. Based on this coarse version, the original is predicted by up-sampling and filtering, and the

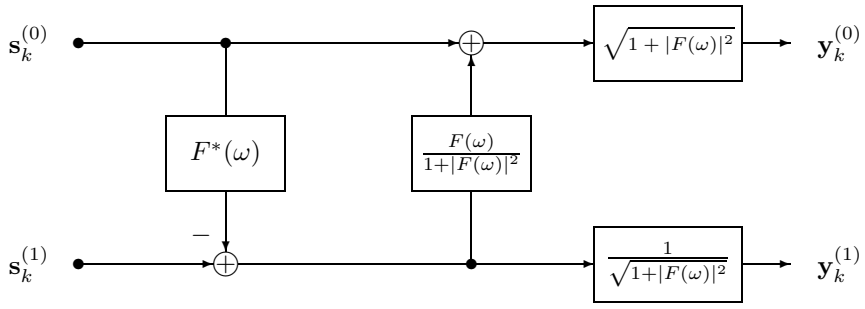

Fig. 2. Equivalent inter-resolution Haar transform. Prediction and update step uses just filtering.

difference is calculated as the prediction error. For the reconstruction, the signal is obtained by simply adding back the difference to the prediction from the coarse signal. In [9, 10], the Laplacian pyramid is studied using the frame theory and it is demonstrated that the usual reconstruction is suboptimal. The proposed filter bank for the reconstruction of the Laplacian pyramid uses a projection that leads to an improvement over the usual method in the presence of noise. As the Laplacian pyramid is an overcomplete representation, it is well known that the dual frame operator should be used for the reconstruction. But it is rarely used in practice. The authors demonstrate that gains around 1 dB are actually possible over the usual reconstruction.

3. EQUIVALENT INTER-RESOLUTION TRANSFORM

To study the inter-resolution transform in more detail, we discuss the equivalent inter-resolution transform in **Fig. 2**. We select the same resolution for both input pictures but describe the low-resolution images as band-limited signals in the base-band $\mathcal{B} = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\frac{\pi}{2}, \frac{\pi}{2}]$. We neglect spectral replica due to sampling and use for the transform the low-pass filter

$$F(\omega) = L(\omega)\mathbf{1}_{\mathcal{B}}(\omega), \tag{1}$$

where $\mathbf{1}_{\mathcal{B}}(\omega)$ is 1 for frequencies in the base-band $\mathcal{B}$ and 0 elsewhere. In particular, the base-band limited picture $\mathbf{s}_k^{(0)}$ is obtained by filtering the high-resolution picture $\mathbf{s}_k^{(1)}$ with $F(\omega)$. In **Fig. 2**, the prediction step interpolates the base-band limited picture $\mathbf{s}_k^{(0)}$ and utilizes the interpolation filter $F^*(\omega)$ which is the complex conjugate of $F(\omega)$. The prediction is subtracted from the high-resolution picture $\mathbf{s}_k^{(1)}$. The update step employs the sampling filter $F(\omega)/[1+|F(\omega)|^2]$. To obtain a normalized transform, the low- and high-band images are filtered. The normalized high-band image is denoted by $\mathbf{y}_k^{(1)}$, the normalized low-band image by $\mathbf{y}_k^{(0)}$. The equivalent transform in matrix notation reads

$$T(\omega) = \frac{1}{\sqrt{1+|F(\omega)|^2}} \begin{pmatrix} 1 & F(\omega) \\ -F^*(\omega) & 1 \end{pmatrix}. \tag{2}$$

The transform is orthonormal for any given $F(\omega)$ and all frequencies $\omega = (\omega_x, \omega_y)$.

Finally, we discuss the decorrelation property of the equivalent transform. This property characterizes the coding efficiency of the spatially scalable coding scheme. Given the power spectral density $\Phi_{\mathbf{ss}}^{(1)}$ of the high-resolution images $\mathbf{s}_k^{(1)}$, we obtain for the power spectral density matrix $\Phi_{\mathbf{yy}}(\omega)$ of the spatial subbands $\mathbf{y}_k^{(\cdot)}$

$$\Phi_{\mathbf{yy}}(\omega) = \begin{pmatrix} \frac{4|F(\omega)|^2}{1+|F(\omega)|^2} & \frac{2F(\omega)[1-|F(\omega)|^2]}{1+|F(\omega)|^2} \\ \frac{2F^*(\omega)[1-|F(\omega)|^2]}{1+|F(\omega)|^2} & \frac{[1-|F(\omega)|^2]^2}{1+|F(\omega)|^2} \end{pmatrix} \Phi_{\mathbf{ss}}^{(1)}. \tag{3}$$

Note that we achieve perfect decorrelation only if the off-diagonal elements of the power spectral matrix are zero, i.e. $F(\omega)[1 - |F(\omega)|^2] = 0$. Good approximations can be achieved with an arbitrarily accurate low-pass filter.

4. EXPERIMENTAL RESULTS

We obtain experimental results with a video coding scheme based on motion-compensated lifted wavelets [4, 5]. We employ the inter-resolution transform in **Fig. 1** to the images of the video in CIF resolution and its down-sampled version in QCIF resolution. Each resulting spatial subband is temporally decomposed in groups of 32 pictures (GOP). The dyadic decomposition utilizes either the motion-compensated Haar wavelet or the motion-compensated 5/3 wavelet. For the temporal lifting steps, half-pel accurate motion compensation with 16×16 blocks is used. The temporal update step employs the negative motion vector of the corresponding temporal prediction step. The motion vectors are chosen such that they minimize the square error in the corresponding temporal high-band. The temporal transform provides

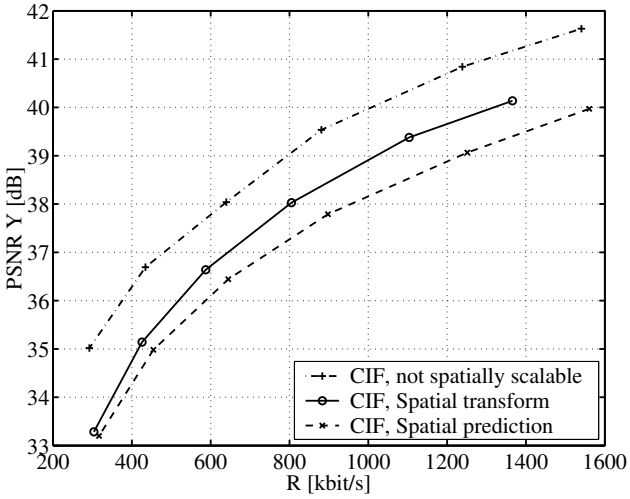

Fig. 3. Rate-distortion performance of the complete embedded representation for the sequence *Container Ship* at 30 fps in CIF resolution. The motion-compensated Haar kernel for groups of 32 pictures is used to compare the inter-resolution transform with inter-resolution prediction. The performance of the motion-compensated temporal Haar transform applied to the original CIF sequence is given as a reference.

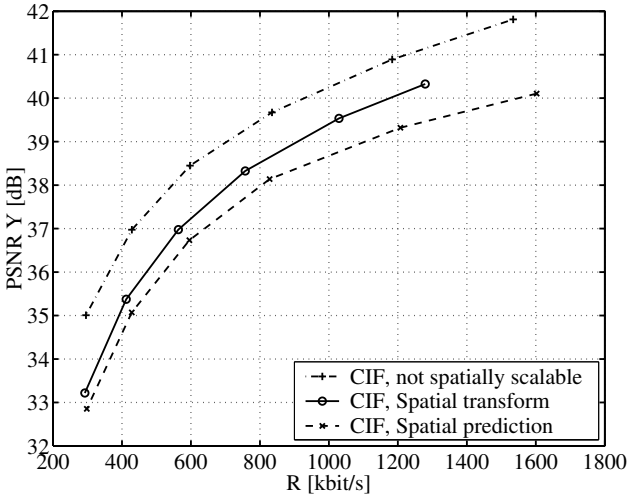

Fig. 4. Rate-distortion performance of the complete embedded representation for the sequence *Salesman* at 30 fps in CIF resolution. The motion-compensated Haar kernel for groups of 32 pictures is used to compare the inter-resolution transform with inter-resolution prediction. The performance of the motion-compensated temporal Haar transform applied to the original CIF sequence is given as a reference.

32 subbands that are intra-frame encoded with a 8×8 DCT and run-length coding. These 32 intra-frame coder use the same quantizer step-size.

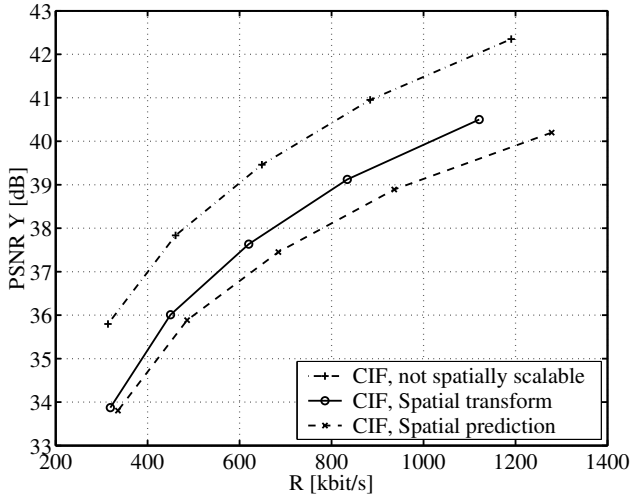

Fig. 5. Rate-distortion performance of the complete embedded representation for the sequence *Container Ship* at 30 fps in CIF resolution. The motion-compensated 5/3 kernel for groups of 32 pictures is used to compare the inter-resolution transform with inter-resolution prediction. The performance of the motion-compensated temporal 5/3 transform applied to the original CIF sequence is given as a reference.

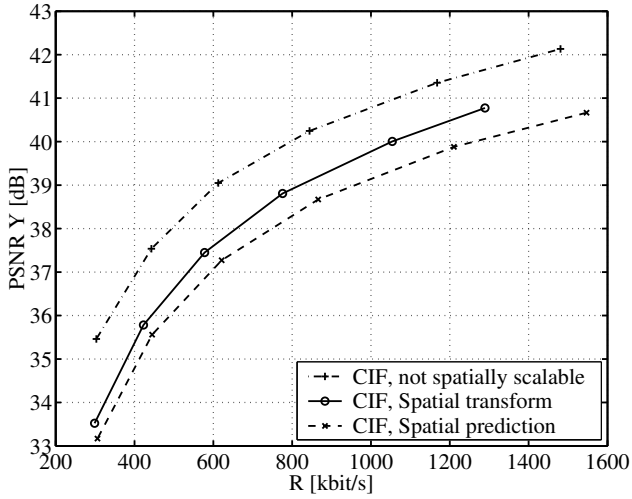

Fig. 6. Rate-distortion performance of the complete embedded representation for the sequence *Salesman* at 30 fps in CIF resolution. The motion-compensated 5/3 kernel for groups of 32 pictures is used to compare the inter-resolution transform with inter-resolution prediction. The performance of the motion-compensated temporal 5/3 transform applied to the original CIF sequence is given as a reference.

As outlined in Section 3, the equivalent inter-resolution transform is orthonormal. Therefore, we select the same quantizer step-size for the intra-frame encoder in the spatial

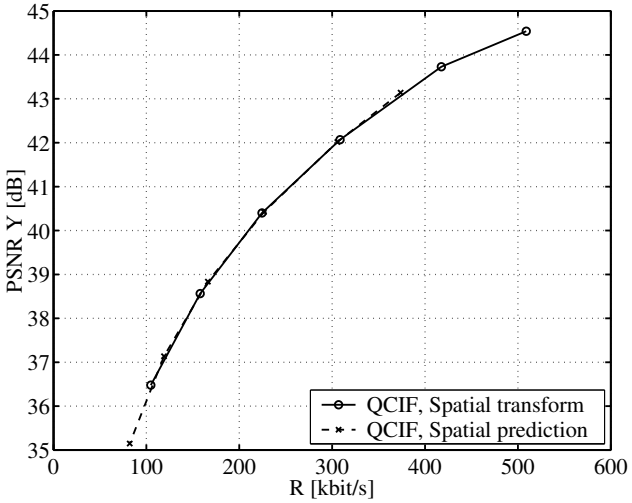

Fig. 7. Rate-distortion performance of the sub-stream that represents the sequence *Container Ship* at 30 fps in QCIF resolution. The motion-compensated Haar kernel is used to encode groups of 32 pictures. For the lower resolution, the efficiency of the inter-resolution transform is the same as that of inter-resolution prediction.

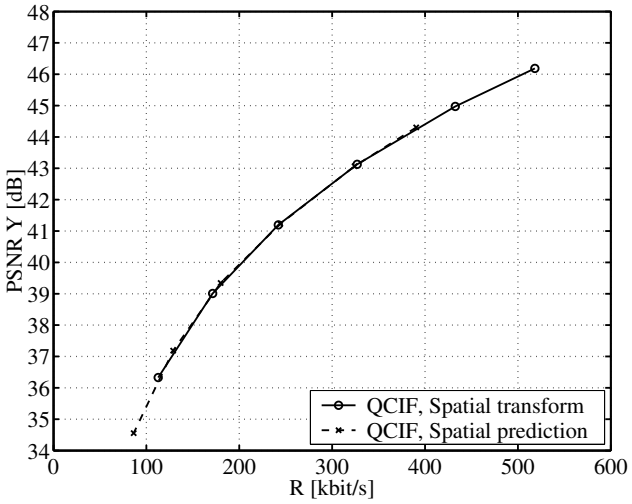

Fig. 8. Rate-distortion performance of the sub-stream that represents the sequence *Salesman* at 30 fps in QCIF resolution. The motion-compensated Haar kernel is used to encode groups of 32 pictures. For the lower resolution, the efficiency of the inter-resolution transform is the same as that of inter-resolution prediction.

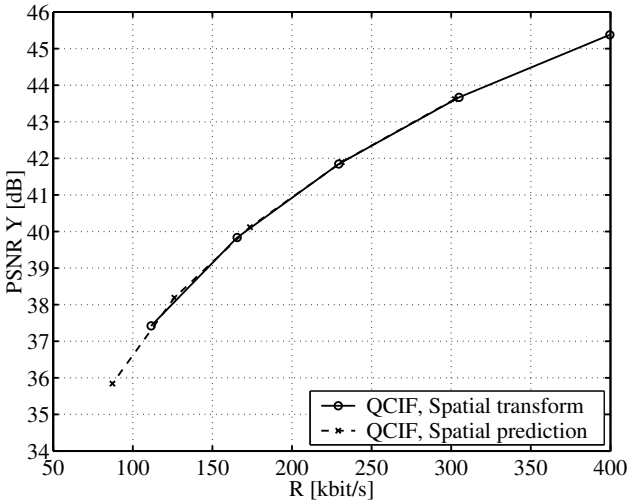

Fig. 9. Rate-distortion performance of the sub-stream that represents the sequence *Container Ship* at 30 fps in QCIF resolution. The motion-compensated 5/3 kernel is used to encode groups of 32 pictures. For the lower resolution, the efficiency of the inter-resolution transform is the same as that of inter-resolution prediction.

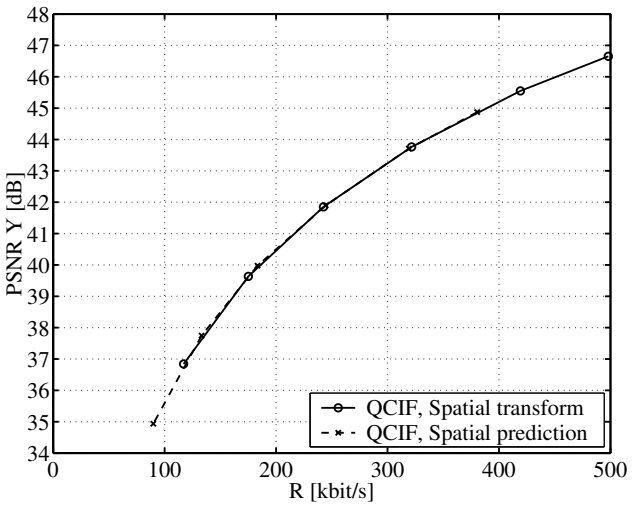

Fig. 10. Rate-distortion performance of the sub-stream that represents the sequence *Salesman* at 30 fps in QCIF resolution. The motion-compensated 5/3 kernel is used to encode groups of 32 pictures. For the lower resolution, the efficiency of the inter-resolution transform is the same as that of inter-resolution prediction.

low-band as well as for those in the spatial high-band. Further, the motion information in the spatial low-band is used to predict the motion information in the spatial high-band. That is, the enhancement layer carries the refined motion information with respect to the motion information in the spatial low-bands. As we have an overcomplete representation of the spatial high-bands, the motion information of the spatial low-bands can be exploited efficiently.

According to **Fig. 1**, the spatial low-band is filtered by

$\sqrt{1+|L(\frac{\tilde{\omega}}{2})|^2}$. If the low-pass filter $L(\omega)$ is ideal, filtering is simplified to scaling by $\sqrt{2}$. As the implemented low-pass filter is approximately ideal, we simplify the scheme by scaling the low-band with $\sqrt{2}$. For the spatial high-band, base-band frequencies are scaled by $1/\sqrt{2}$, whereas others pass unscaled. This can be achieved either by linear filtering, or by selecting appropriately the quantizer step-sizes for the frequency coefficients of the intra-frame encoder.

Figs. 3 and **4** depict the rate-distortion performance of the complete embedded representation for the sequences *Container Ship* and *Salesman* in CIF resolution, respectively. The motion-compensated Haar wavelet is used to encode 288 frames of each sequence. As a reference, the motion-compensated temporal Haar transform is also applied to the original CIF sequences. **Figs. 5** and **6** show the corresponding results for the motion-compensated 5/3 wavelet. It can be observed that the inter-resolution transform outperforms the reference scheme with inter-resolution prediction (i.e. the transform without spatial update step and normalization) by up to 1 dB. Further, the gap to the performance of the non-scalable representation is somewhat larger for the motion-compensated 5/3 wavelet. This may be due to inefficient encoding of the motion information. Please note also that the used intra-frame codec is not SNR scalable. The rate-distortion points are obtained by choosing appropriately the quantizer step-size for each encoding.

Finally, **Figs. 7** and **8** show the rate-distortion performance of the spatial base layer signal $\tilde{\mathbf{o}}_k^{(0)}$ in QCIF resolution for the sequences *Container Ship* and *Salesman*, respectively. The results for the motion-compensated Haar wavelet are shown. **Figs. 9** and **10** depict the corresponding results for the motion-compensated 5/3 wavelet. No degradation for the spatial transform is observed if the low-resolution sequence is selected appropriately. This is independent of the utilized motion-compensated wavelet kernel.

5. CONCLUSION

This paper discusses an orthonormal inter-resolution transform for spatially scalable video coding. We use an overcomplete spatial representation for two different spatial scales of the same image and perform motion-compensated temporal filtering on the shift-invariant spatial subbands. The transform minimizes the impact of the quantization noise on the reconstructed video signal at the decoder. The scheme is related to the Laplacian pyramid with prediction of the higher resolution, but the discussed transform uses additionally a spatial update step. This update step orthogonalizes the decomposition and enables our coding scheme to outperform the compression efficiency of the Laplacian pyramid.

6. REFERENCES

[1] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, vol. 3, pp. 1793–1796.

[2] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, Oct. 2001, vol. 2, pp. 1029–1032.

[3] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[4] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 561–575, Aug. 2004.

[5] M. Flierl, P. Vandergheynst, and B. Girod, "Video coding with lifted wavelet transforms and complementary motion-compensated signals," in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, San Jose, CA, Jan. 2004, vol. 5308, pp. 497–508.

[6] D. Taubman, "Successive refinement of video: Fundamental issues, past efforts and new directions," in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, Lugano, Switzerland, July 2003, pp. 649–663.

[7] ITU-T, *Recommendation H.263++ (Video Coding for Low Bitrate Communication)*, 2000.

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

[9] M.N. Do and M. Vetterli, "Frame reconstruction of the Laplacian pyramid," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001, pp. 3641–3644.

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