INVESTIGATION OF MOTION-COMPENSATED LIFTED WAVELET TRANSFORMS

Markus Flierl and Bernd Girod

Information Systems Laboratory
Stanford University, Stanford, CA 94305
mflierl@ieee.org, bgirod@stanford.edu

ABSTRACT

This paper investigates lifted wavelet transforms applied in the temporal direction of a video sequence. Due to the motion between pairs of frames, motion compensation is utilized in the lifting steps. We discuss the modified Haar and 5/3 wavelet kernel and provide experimental results for dyadic decompositions with various levels. Further, we utilize a signal model for a theoretical discussion of both kernels. We generalize and replace the dyadic decompositions by the Karhunen-Loeve Transform in order to provide theoretical performance bounds for the compression efficiency of these coding schemes.

1. INTRODUCTION

Applying a linear transform in temporal direction of a video sequence may not be very efficient if significant motion is prevalent. Motion compensation between two frames is necessary to deal with the motion in a sequence. Consequently, a combination of linear transform and motion compensation seems promising for efficient compression. For wavelet transforms, the so called *Lifting Scheme* [1] can be used to construct the kernels. A two-channel decomposition can be achieved with a sequence of prediction and update steps that form a ladder structure. The advantage is that this lifting structure is able to map integers to integers without requiring invertible lifting steps. Further, motion compensation can be incorporated into the prediction and update steps as proposed in [2]. The fact that the lifting structure is invertible without requiring invertible lifting steps makes this approach feasible. We cannot count on invertible lifting steps as, in general, motion compensation is not invertible. If it is invertible, this motion-compensated wavelet transform based on lifting permits a linear transform along the motion trajectories in a video sequence.

2. CODING SCHEME

The investigated coding schemes process the video sequence in groups of K pictures (GOP). First, we decompose each GOP in temporal direction. The dyadic decomposition utilizes a motion-compensated wavelet which will be discussed later in more detail. The temporal transform provides K output pictures that are intra-frame encoded. In order to allow a comparison to a classic hybrid coder, we utilize for the intra-frame coder a 8×8 DCT with run-length coding. If we employ a Haar wavelet and set the motion vectors to zero, the dyadic decomposition will be an orthonormal transform. Therefore, we select the same quantizer step-size for all K intra-frame encoder. The motion information that is required for the motion-compensated wavelet transform is estimated in each decomposition level depending on the results of the lower level. Further, we employ half-pel accurate motion compensation with bi-linear interpolation.

2.1. Motion-Compensated Lifted Haar Wavelet

First, we discuss the lifting scheme with motion compensation for the Haar wavelet [2].

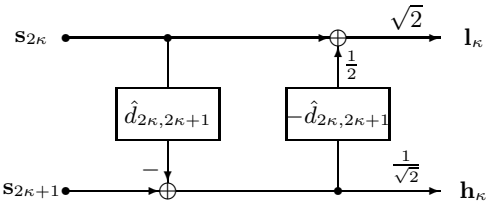

Fig. 1. Haar transform with motion-compensated lifting steps. The update step uses the negative motion vector of the prediction step.

Fig. 1 depicts a Haar transform with motion-compensated lifting steps. The even frames of the video sequence $\mathbf{s}_{2\kappa}$ are displaced by the estimated value $\hat{d}_{2\kappa,2\kappa+1}$ to predict its odd frames $\mathbf{s}_{2\kappa+1}$. The prediction step is followed by an update step with the displacement $-\hat{d}_{2\kappa,2\kappa+1}$. We use a block-size of 16×16 and half-pel accurate motion compensation in the prediction step and select the motion vectors such that they minimize the squared error in the high-band $\mathbf{h}_\kappa$. In general, the block-motion field is not invertible but we still utilize the negative motion vectors for the update step. Additional scaling factors in low- and high-band are necessary to normalize the transform.

2.2. Motion-Compensated Lifted 5/3 Wavelet

The Haar wavelet is a short filter and provides limited coding gain. We expect better coding efficiency with longer wavelet kernels. In the following, we discuss the lifted 5/3 wavelet kernel with motion compensation [2].

Fig. 2 depicts the 5/3 transform with motion-compensated lifting steps. Similar to the Haar transform, the update steps use the negative motion vectors of the corresponding prediction steps. But for this transform, the odd frames are predicted by a linear combination of two displaced neighboring even frames. Again, we use a block-size of 16×16 and half-pel accurate motion compensation in the prediction steps and choose the motion vectors $\hat{d}_{2\kappa,2\kappa+1}$ and $\hat{d}_{2\kappa+2,2\kappa+1}$ such that they minimize the squared error in the high-band $\mathbf{h}_{2\kappa+1}$. The corresponding update steps use the negative motion vectors.

M. Flierl is on leave from the Telecommunications Laboratory, University of Erlangen-Nuremberg, Germany

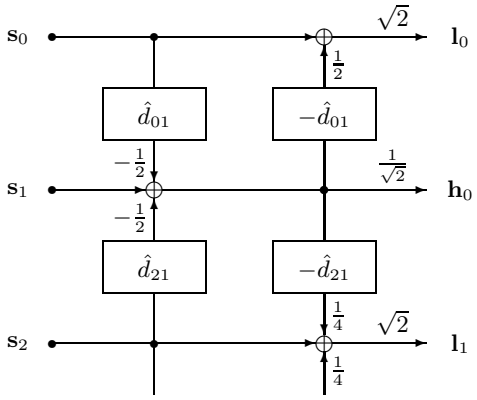

Fig. 2. 5/3 transform with motion-compensated lifting steps. The update steps use the negative motion vectors of the corresponding prediction steps.

2.3. Experimental Results

For the experiments, we subdivide the test sequences *Mother & Daughter* and *Mobile & Calendar*, each with 288 frames, into groups of K pictures. We decompose the GOPs independent of each other and in the case of the 5/3 wavelet, we use a cyclic extension as it is slightly beneficial for some sequences.

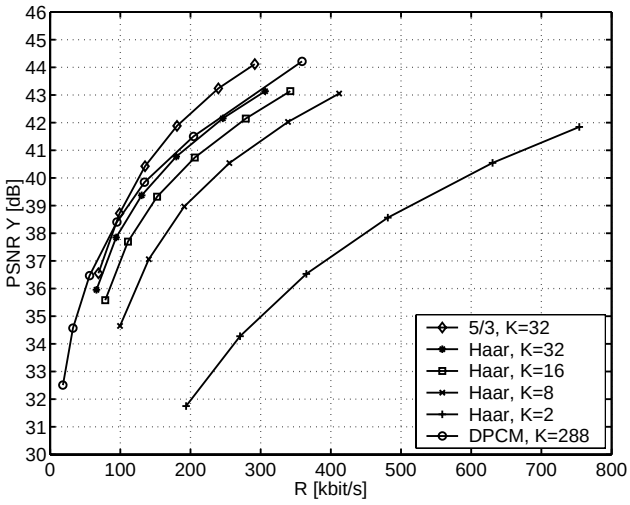

Fig. 3. Luminance PSNR vs. total bit-rate for the QCIF sequence *Mother & Daughter* at 30 fps. A dyadic decomposition is used to encode groups of $K = 2$, 8, 16, and 32 pictures with the Haar kernel, and $K = 32$ with the 5/3 kernel. Results for a basic hybrid video codec with 287 inter-frames are given for reference.

Figs. 3 and 4 show the luminance PSNR over the total bit-rate for the sequences *Mother & Daughter* and *Mobile & Calendar* encoded with groups of $K = 2$, 8, 16, and 32 pictures with the Haar kernel. The bit-rate savings diminish very quickly as the GOP size approaches 32 pictures. Please note that for *Mobile & Calendar* at lower bit-rates the wavelet codec outperforms a basic hybrid video codec (intra- and inter-frames, 16×16 block-motion compensation, half-pel accuracy, previous reference picture, and 8×8 DCT) with a very large GOP size. Note also that the 5/3 decomposition with a GOP size of 32 outperforms not only the corresponding Haar decomposition but also the basic hybrid codec with K=288.

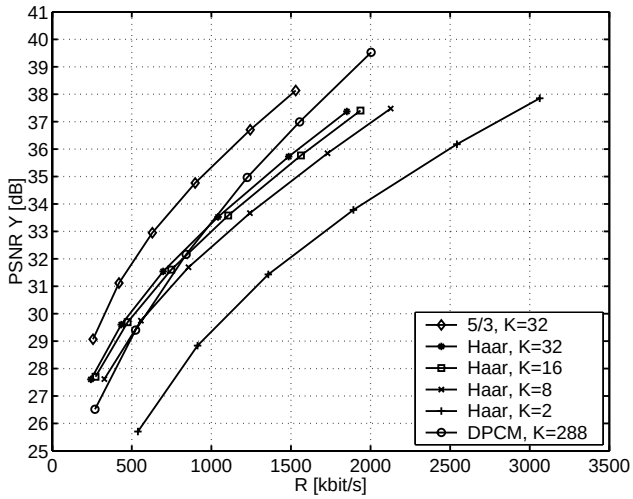

Fig. 4. Luminance PSNR vs. total bit-rate for the QCIF sequence *Mobile & Calendar* at 30 fps. A dyadic decomposition is used to encode groups of $K = 2$, 8, 16, and 32 pictures with the Haar kernel, and $K = 32$ with the 5/3 kernel. Results for a basic hybrid video codec with 287 inter-frames are given for reference.

3. THEORETICAL SIGNAL MODEL

Let $\mathbf{s}_k = \{\mathbf{s}_k[l], l \in \Pi\}$ be scalar random fields over a two-dimensional orthogonal grid Π with horizontal and vertical spacing of 1. The vector $l = (x, y)^T$ denotes a particular location in the lattice Π. We interpret $\mathbf{s}_k$ as the k-th of K pictures to be encoded. Further, the signal $\mathbf{s}_k[l]$ is band-limited and we obtain a displaced version of it as follows: we shift the ideal reconstruction of the band-limited signal by the continuous-valued displacement vector d and re-sample it on the original grid. With this signal model, any shift-invariant displacement operation is invertible.

3.1. Motion-Compensated Lifted Haar Wavelet

Now, given this signal model, we revisit the motion-compensated lifted Haar wavelet in Fig. 1 and remove the displacement operators in the lifting steps such that we can isolate a lifted Haar wavelet without displacement operators.

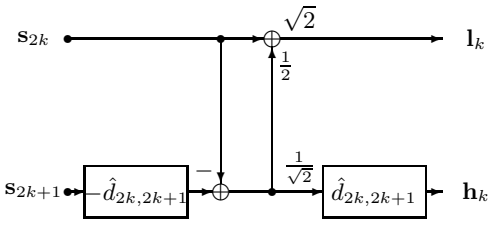

Fig. 5. Equivalent Haar wavelet without shifts in the lifting steps.

Fig. 5 shows the equivalent Haar wavelet where the displacement operators are pre- and post-processing operators with respect to the original Haar transform. This scheme is equivalent to Fig. 1, if the displacement operators are invertible.

We continue and perform the dyadic decomposition of a GOP with the equivalent Haar wavelet. For that, the displacements of the equivalent Haar blocks have to be added. We assume that the estimated displacements between pairs of frames are additive such that, e.g., $\hat{d}_{02} + \hat{d}_{23} = \hat{d}_{03}$. As the true displacements are also

additive, e.g. $d_{02} + d_{23} = d_{03}$, and differ from the estimated displacement by the displacement error, i.e. $d_{ij} = \hat{d}_{ij} + \Delta_{ij}$, we conclude that the displacement errors are also additive, e.g. $\Delta_{02} + \Delta_{23} = \Delta_{03}$.

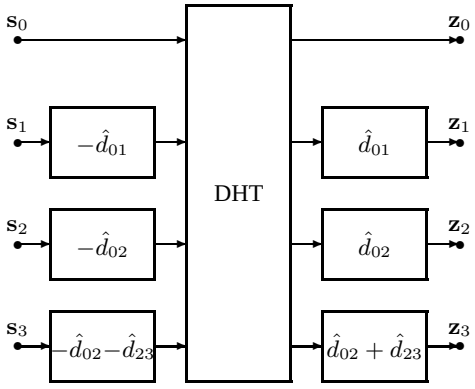

Fig. 6. Dyadic Haar Transform (DHT) without shifts in the lifting steps for $K = 4$ pictures.

Fig. 6 depicts a dyadic decomposition for $K = 4$ pictures based on the equivalent Haar wavelet in Fig. 5. The dyadic Haar transform without displacements in the lifting steps is labeled by DHT. The displacements $\hat{d}_{0k}$ are pre- and post-processing operators with respect to the original dyadic Haar decomposition DHT.

3.2. Motion-Compensated Lifted 5/3 Wavelet

We also apply the invertible displacement operator to the motion-compensated lifted 5/3 wavelet in Fig. 2 and obtain the equivalent 5/3 wavelet in Fig. 7. Due to the structure of the 5/3 wavelet, we have displacements between the frames 2κ & $2\kappa+1$, $2\kappa+2$ & $2\kappa+1$, and 2κ & $2\kappa+2$ (in the next decomposition level). Again, we assume that the estimated displacements are additive such that, e.g., $\hat{d}_{01} - \hat{d}_{21} = \hat{d}_{02}$. With this assumption, the displacement operators between the levels cancel out and several decomposition levels are possible without displacements between the levels.

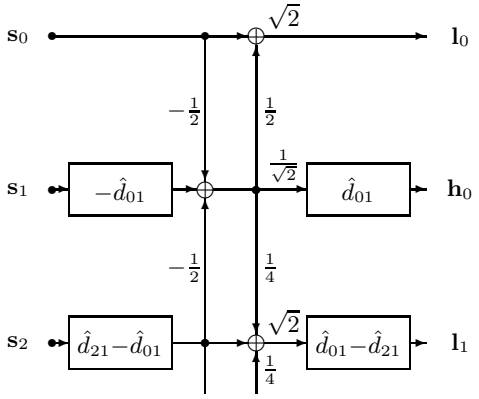

Fig. 7. Equivalent 5/3 wavelet without shifts in the lifting steps.

The equivalent dyadic 5/3 transform has the same pre- and post-processing displacement operators as the equivalent dyadic Haar transform in Fig. 6 but the DHT is replaced by the original dyadic 5/3 decomposition.

3.3. Generalized Signal Model

Now, we assume that the pictures $\mathbf{s}_k$ are shifted versions of a "clean" video signal $\mathbf{v}$ with the true displacements d_{0k} and distorted by independent additive white Gaussian noise $\mathbf{n}_k$. Combining this signal model with the equivalent dyadic decomposition, we can eliminate the absolute displacements and restrict ourselves to the displacement error Δ_{0k} in the k-th picture. In the following, we do not consider particular displacement errors Δ_{0k}. We rather specify statistical properties and consider them as random variables $\mathbf{\Delta}_k$, statistically independent from the "clean" signal $\mathbf{v}$ and the noise $\mathbf{n}_k$. The noise signals $\mathbf{n}_\mu$ and $\mathbf{n}_\nu$ are also mutually statistically independent.

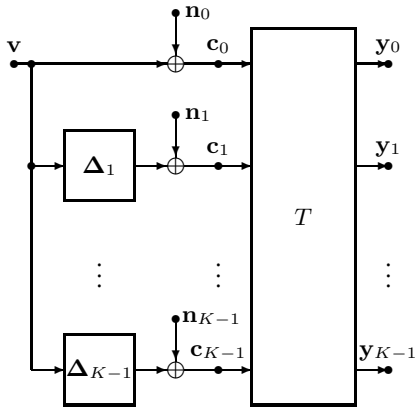

Fig. 8. Motion compensation for a group of K pictures.

Fig. 8 depicts the generalized model with the displacement-free and linear transform T for a group of K pictures. The motion-compensated pictures $\mathbf{c}_1, \ldots, \mathbf{c}_{K-1}$ are aligned with respect to the first picture $\mathbf{c}_0$. According to Fig. 6, the signals $\mathbf{z}_k$ are independently intra-frame encoded. As the absolute displacements have no influence on the performance of the intra-frame encoder, we omit them and consider only the direct output signals $\mathbf{y}_k$ of T.

Now, assume that the random fields $\mathbf{v}$ and $\mathbf{c}_k$ are jointly wide-sense stationary with the real-valued scalar two-dimensional power spectral densities $\Phi_{\mathbf{vv}}(\omega)$ and $\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega)$. The power spectral densities $\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega)$ are elements in the power spectral density matrix of the motion-compensated pictures $\Phi_{\mathbf{cc}}$. The power spectral density matrix of the decorrelated signal $\Phi_{\mathbf{yy}}$ is given by $\Phi_{\mathbf{cc}}$ and the transform T,

$$\Phi_{\mathbf{yy}}(\omega) = T(\omega)\Phi_{\mathbf{cc}}(\omega)T^H(\omega), \tag{1}$$

where T^H denotes the Hermitian of T and $\omega = (\omega_x, \omega_y)^T$ the vector-valued frequency.

We adopt the expressions for the cross spectral densities $\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}$ from [3]

$$\Phi_{\mathbf{c}_\mu\mathbf{c}_\nu}(\omega) = E\left\{e^{-j\omega^T(\mathbf{\Delta}_\mu - \mathbf{\Delta}_\nu)}\right\}\Phi_{\mathbf{vv}}(\omega) + \Phi_{\mathbf{n}_\mu\mathbf{n}_\nu}(\omega) \tag{2}$$

and assume a power spectrum $\Phi_{\mathbf{vv}}$ that corresponds to an exponentially decaying isotropic autocorrelation function with a correlation coefficient $\rho_{\mathbf{v}} = 0.93$.

For the k-th displacement error $\mathbf{\Delta}_k$, a 2-D normal distribution with variance $\sigma^2_{\mathbf{\Delta}}$ and zero mean is assumed where the x- and y-components are statistically independent. The expected value in (2) depends on the variance of the displacement error with respect to the reference picture $\mathbf{c}_0$ (*absolute displacement accuracy*) and the variance of the difference displacement error between pairs of non-reference pictures (*relative displacement accuracy*). We assume that each picture in a GOP can be the reference picture $\mathbf{c}_0$. That is, there is no preference among the pictures in a GOP and the variances of the absolute displacement error are the same for all

$K-1$ motion-compensated pictures. Based on the dyadic decomposition with motion-compensated lifted wavelets and the assumption that there is no preference among the pictures in a GOP, we assume that absolute and relative displacement accuracy are identical. The differences of absolute displacement errors are equal to relative displacement errors as we assume additive estimated displacements. This results in correlated displacement errors [4]. We abbreviate the expected value in (2) with $P(\omega, \sigma_{\mathbf{\Delta}}^2)$ which is the characteristic function of the Gaussian displacement error.

With that, we obtain for the power spectral density matrix of the motion-compensated pictures

$$\frac{\Phi_{\mathbf{cc}}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} = \begin{pmatrix} 1+\alpha & P & \cdots & P \\ P & 1+\alpha & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & 1+\alpha \end{pmatrix}. \quad (3)$$

$\alpha = \alpha(\omega)$ is the normalized spectral density of the noise $\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)$ with respect to the spectral density of the "clean" video signal.

$$\alpha(\omega) = \frac{\Phi_{\mathbf{n}_k\mathbf{n}_k}(\omega)}{\Phi_{\mathbf{vv}}(\omega)} \quad \text{for} \quad k = 0, 1, \ldots, K-1 \quad (4)$$

T represents the dyadic Haar transform or the dyadic 5/3 transform. In terms of decorrelation and coding gain, the 5/3 wavelet performs better than the Haar wavelet as shown in Figs. 3 and 4. In the following, we are interested in theoretical performance bounds and choose the Karhunen-Loeve Transform (KLT). The normalized eigenvalues of the power spectral density matrix $\Phi_{\mathbf{cc}}$ are $\lambda_1(\omega) = 1 + \alpha(\omega) + (K-1)P(\omega)$ and $\lambda_{2,3,\ldots,K}(\omega) = 1 + \alpha(\omega) - P(\omega)$. The power spectral density matrix of the transformed signals $\Phi_{\mathbf{yy}}$ is diagonal. The first eigenvector just adds all components and scales with $1/\sqrt{K}$. For the remaining eigenvectors, any orthonormal basis can be used that is orthogonal to the first eigenvector. That is, the KLT for our signal model is not dependent on ω. Note, that for this simple signal model, the Haar transform is also a KLT.

The rate difference [3] is used to measure the improved compression efficiency for each picture k.

$$\Delta R_k = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} \log_2 \left(\frac{\Phi_{\mathbf{y}_k\mathbf{y}_k}(\omega)}{\Phi_{\mathbf{c}_k\mathbf{c}_k}(\omega)} \right) d\omega \quad (5)$$

It represents the maximum bit-rate reduction (in bit per sample) possible by optimum encoding of the transformed signal $\mathbf{y}_k$, compared to optimum intra-frame encoding of the signal $\mathbf{c}_k$ for Gaussian wide-sense stationary signals for the same mean squared reconstruction error. The overall rate difference ΔR is the average over all pictures and is used to evaluate the efficiency of motion-compensated transform coding. Assuming the KLT, we obtain for the overall rate difference

$$\Delta R = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{K-1}{2K} \log_2 \left(1 - \frac{P(\omega, \sigma_{\mathbf{\Delta}}^2)}{1+\alpha(\omega)} \right) + \frac{1}{2K} \log_2 \left(1 + (K-1) \frac{P(\omega, \sigma_{\mathbf{\Delta}}^2)}{1+\alpha(\omega)} \right) d\omega. \quad (6)$$

The case of a very large number of motion-compensated pictures is of special interest for the comparison to video coding with motion-compensated prediction.

$$\Delta R_{K\to\infty} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} \log_2 \left(1 - \frac{P(\omega, \sigma_{\mathbf{\Delta}}^2)}{1+\alpha(\omega)} \right) d\omega \quad (7)$$

According to [3], the performance of motion-compensated prediction with optimum Wiener filter achieves a rate difference of

$$\Delta R_{\text{MCP}} = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \frac{1}{2} \log_2 \left(1 - \frac{P^2(\omega, \sigma_{\mathbf{\Delta}}^2)}{[1+\alpha(\omega)]^2} \right) d\omega. \quad (8)$$

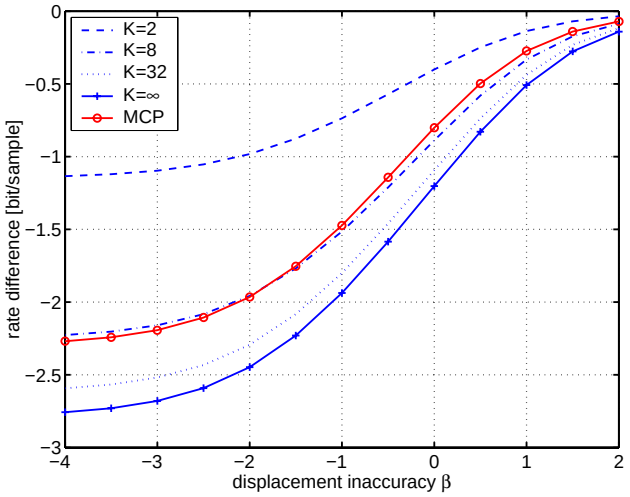

Fig. 9. Rate difference for motion-compensated transform coding with groups of K pictures over the displacement inaccuracy β. The performance of motion-compensated prediction with Wiener filter is labeled by MCP. The residual noise level is -30 dB.

Fig. 9 depicts the rate difference according to (6) and (8) over the displacement inaccuracy $\beta = \log_2(\sqrt{12}\sigma_{\mathbf{\Delta}})$ for a residual noise level of -30 dB. We observe that the rate difference starts to saturate for $K = 32$ and that motion-compensated transform coding outperforms motion-compensated prediction by at most 0.5 bits per sample. These observations are consistent with the experimental results in the previous section.

4. CONCLUSION

We investigate experimentally and theoretically motion-compensated lifted wavelet transforms. We implement dyadic Haar and 5/3 wavelets and observe the superiority of the later. Based on an ideal signal model and the additivity of estimated displacements, we develop equivalent transforms without displacement operators in the lifting steps. Further, we determine performance bounds with the Karhunen-Loeve Transform and observe that we outperform video coding with motion-compensated prediction by at most 0.5 bits per sample.

5. REFERENCES

[1] W. Sweldens, "The lifting scheme: A construction of second generation wavelets," *SIAM Journal on Mathematical Analysis*, vol. 29, no. 2, pp. 511–546, 1998.

[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, pp. 1029–1032.

[3] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *IEEE Transactions on Image Processing*, vol. 9, no. 2, pp. 173–183, Feb. 2000.

[4] M. Flierl and B. Girod, "Video coding with motion compensation for groups of pictures," in *Proceedings of the IEEE International Conference on Image Processing*, Rochester, NY, Sept. 2002.