A Motion-Compensated Orthogonal Transform with Energy-Concentration Constraint

Markus Flierl and Bernd Girod
Max Planck Center for Visual Computing and Communication
Stanford University, California
{mflierl, bgirod}@stanford.edu

Abstract— **This paper discusses a transform for successive pictures of an image sequence which strictly maintains orthogonality while permitting motion compensation between pairs of pictures. The well known motion-compensated lifted Haar wavelet maintains orthogonality only approximately. In the case of zero motion fields, the motion-compensated lifted Haar wavelet is known to be orthogonal. But for complex motion fields with many multi-connected and unconnected pixels, the motion-compensated lifted Haar wavelet cannot accurately maintain its transform property and, hence, suffers a performance degradation. The presented motion-compensated orthogonal transform strictly maintains orthogonality for any motion field. Further, the transform is designed with an energy-concentration constraint. The energy of the input pictures is accumulated in the temporal low-band while the temporal high-band is zero if the input pictures are identical up to a known single-connecting motion field. This constraint makes the transform suitable for coding applications.**

I. Introduction

We address the problem of representing image sequences for coding and communication applications. Well known methods are standard hybrid video coding techniques as well as subband coding schemes. The latter are deemed to provide more flexible representations which may better adapt to heterogeneous communication scenarios. For 3-D subband coding with motion compensation, [1] proposes to distinguish between connected, covered, and uncovered pixels when incorporating motion compensation for filtering in temporal direction. Motion-compensated filtering in [2] addresses the problem of double-connected pixels and proposes an ad-hoc method to resolve the ambiguity. [3], [4], [5] choose a lifting implementation for the temporal filter and incorporate motion compensation into the lifting steps. In contrast to previous work, the lifting implementation permits a reversible filter structure, but still, it struggles with unconnected, connected, and multi-connected pixels when performing the update step. [6] proposes an optimum update step that minimizes the mean-squared reconstruction error. But note that the motion-compensated lifted Haar wavelet maintains orthogonality only approximately. For a zero motion field, the motion-compensated lifted Haar wavelet is known to be orthogonal. But for complex motion fields with many multi-connected and unconnected pixels, the motion-compensated lifted Haar wavelet cannot accurately maintain its transform property and, hence, suffers a performance degradation.

In contrast to previous work, the presented motion-compensated orthogonal transform strictly maintains orthogonality for any motion field. The transform is factored into a sequence of incremental transforms that are strictly orthogonal. The incremental transforms maintain scale counters to keep track of the scale factors that are introduced to ensure orthogonality. The decorrelation factor of each incremental transform is determined by the scale counters and is chosen such that the transform meets an energy-concentration constraint.

The paper is organized as follows: Section II introduces the motion-compensated orthogonal transform and discusses the incremental transform as well as the energy-concentration constraint. Section III proposes an adaptive spatial transform to process the resulting temporal low-band. Section IV presents the experimental results.

II. Motion-Compensated Orthogonal Transform

This section discusses how the transform is factored into incremental transforms. We outline the construction of the incremental transform and the incorporation of the energy-concentration constraint.

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two vectors representing consecutive pictures of an image sequence. The transform T maps these vectors according to

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = T \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \tag{1}$$

into two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ which represent the temporal low- and high-band, respectively. Now, we factor the transform T into a sequence of k incremental transforms T_κ such that

$$T = T_k T_{k-1} \cdots T_\kappa \cdots T_2 T_1, \tag{2}$$

where each incremental transform T_κ is orthogonal by itself, i.e., $T_\kappa^T T_\kappa = T_\kappa T_\kappa^T = I$ holds for all $\kappa = 1, 2, \cdots, k$. This guarantees that the transform T is also orthogonal. It can be imagined that the pixels of the image $\mathbf{x_2}$ are processed from top-left to bottom-right in k steps where each step κ is represented by the incremental transform T_κ.

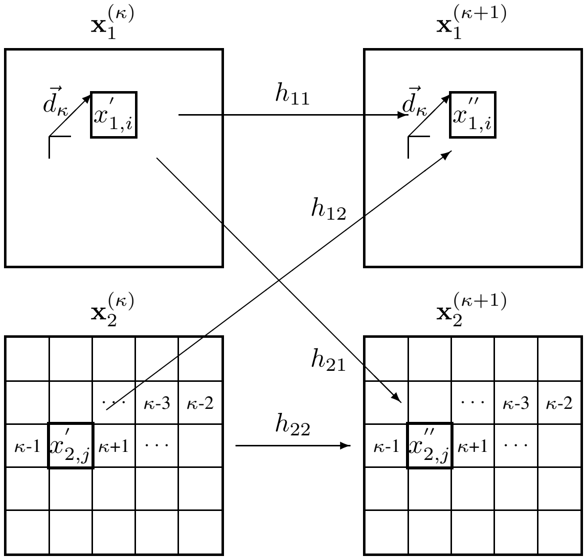
This work has been supported by the Max Planck Center for Visual Computing and Communication.

A. The Incremental Transform

Let $\mathbf{x}_1^{(\kappa)}$ and $\mathbf{x}_2^{(\kappa)}$ be two vectors representing consecutive pictures of an image sequence if $\kappa = 1$, or two output vectors of the incremental transform $T_{\kappa-1}$ if $\kappa > 1$. The incremental transform T_κ maps these vectors according to

$$\begin{pmatrix} \mathbf{x}_1^{(\kappa+1)} \\ \mathbf{x}_2^{(\kappa+1)} \end{pmatrix} = T_\kappa \begin{pmatrix} \mathbf{x}_1^{(\kappa)} \\ \mathbf{x}_2^{(\kappa)} \end{pmatrix} \tag{3}$$

into two vectors $\mathbf{x}_1^{(\kappa+1)}$ and $\mathbf{x}_2^{(\kappa+1)}$ which will be further transformed into the temporal low- and high-band, respectively.



$\mathbf{x}_1^{(\kappa)}$
$\mathbf{x}_1^{(\kappa+1)}$
$\vec{d}_\kappa$
$x'_{1,i}$
h_{11}
$\vec{d}_\kappa$
$x''_{1,i}$
h_{12}
$\mathbf{x}_2^{(\kappa)}$
$\mathbf{x}_2^{(\kappa+1)}$
h_{21}
$\cdots$ κ-3 κ-2
κ-1 $x'_{2,j}$ κ+1 $\cdots$
h_{22}
$\cdots$ κ-3 κ-2
κ-1 $x''_{2,j}$ κ+1 $\cdots$

Fig. 1. The incremental transform T_κ for two frames $\mathbf{x}_1^{(\kappa)}$ and $\mathbf{x}_2^{(\kappa)}$ which strictly maintains orthogonality for any motion field between the two frames.

Fig. 1 depicts the process accomplished by the incremental transform T_κ with its input and output images as defined above. The incremental transform removes the energy of the j-th pixel $x'_{2,j}$ in the image $\mathbf{x}_2^{(\kappa)}$ with the help of the i-th pixel $x'_{1,i}$ in the image $\mathbf{x}_1^{(\kappa)}$ which is linked by the motion vector $\vec{d}_\kappa$ (or of the j-th block with the help of the i-th block if all the pixels of the block have the same motion vector $\vec{d}_\kappa$). The energy-removed pixel value $x''_{2,j}$ is obtained by a linear combination of the pixel values $x'_{1,i}$ and $x'_{2,j}$ with scalar weights h_{21} and h_{22}. The energy-concentrated pixel value $x''_{1,i}$ is also obtained by a linear combination of the pixel values $x'_{1,i}$ and $x'_{2,j}$ but with scalar weights h_{11} and h_{12}. All other pixels are simply kept untouched.

The scalar weights $h_{\mu\nu}$ are arranged into the matrix

$$H = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \tag{4}$$

which is required to be orthogonal. For a 2×2 matrix, one scalar *decorrelation factor* a_n is sufficient to capture all possible orthogonal transforms. We use the form

$$H = \frac{1}{\sqrt{1+a_n^2}} \begin{pmatrix} 1 & a_n \\ -a_n & 1 \end{pmatrix}, \tag{5}$$

where a_n is a positive real value to remove the energy in the image $\mathbf{x}_2$ and to concentrate the energy in the image $\mathbf{x}_1$. The decorrelation factor a_n will be determined in the next subsection which discusses the energy-concentration constraint.

To summarize, the incremental transform T_κ touches only pixels that have the same motion vector. Of these, T_κ performs only a linear combination with pixels pairs that are connected by this motion vector. All other pixels are simply kept untouched. This is reflected with the following matrix notation

$$T_\kappa = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \cdots & 0 & h_{11} & 0 & \cdots & 0 & h_{12} & 0 & \cdots \\ \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots \\ \cdots & 0 & h_{21} & 0 & \cdots & 0 & h_{22} & 0 & \cdots \\ \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{6}$$

where the diagonal elements equal to 1 represent the untouched pixels and where the elements $h_{\mu\nu}$ represent the pixels subject to linear operations.

Note that for accomplishing the transform T, each pixel in $\mathbf{x}_2$ is touched only once whereas the pixels in $\mathbf{x}_1$ may be touched multiple times or never. Further, the order in which the incremental transforms T_κ are applied does not affect the orthogonality of T. But the order may affect the energy concentration of the transform T.

B. The Constraint of Energy Concentration

The decorrelation factor a_n for each pixel touched by the incremental transform has to be chosen such that energy is removed from the image $\mathbf{x}_2$. We discuss a method that reduces the energy in the high-band to zero for any motion vector field if the input pictures are identical and of constant intensity.

Consider the pixel pair $x_{1,i}$ and $x_{2,j}$ to be processed by the incremental transform T_κ. To determine the decorrelation factor a_n for the pixel $x_{2,j}$, we assume that the pixel $x_{2,j}$ is connected to the pixel $x_{1,i}$ such that $x_{2,j} = x_{1,i}$. Consequently, the resulting "to be high-band" pixel $x''_{2,j}$ shall be zero. Note that the pixel $x_{1,i}$ may have been processed previously by T_τ, where $\tau < \kappa$. Therefore, let v_n be the *scale factor* for the pixel $x_{1,i}$ such that $x'_{1,i} = v_n x_{1,i}$. The pixel $x_{2,j}$ is used only once during the transform process T and no scale factor needs to be considered, i.e., $x'_{2,j} = x_{2,j}$. Let v_m be the scale factor for the pixel $x_{1,i}$ after it has been processed by T_κ. Now, the pixels $x'_{1,i}$ and $x'_{2,j}$ are processed by T_κ as follows:

$$\begin{pmatrix} v_m x_{1,i} \\ 0 \end{pmatrix} = \frac{1}{\sqrt{1+a_n^2}} \begin{pmatrix} 1 & a_n \\ -a_n & 1 \end{pmatrix} \begin{pmatrix} v_n x_{1,i} \\ x_{1,i} \end{pmatrix} \tag{7}$$

The condition of energy concentration is satisfied if

$$a_n = \frac{1}{v_n} \quad \text{and} \tag{8}$$

$$v_m = \sqrt{v_n^2 + 1}. \tag{9}$$

Now, let n be the *scale counter*. n simply counts how often the pixel $x_{1,i}$ is used as reference for motion compensation. In the beginning, i.e., before the transform is applied, the scale counter for each pixel $x_{1,i}$ is $n = 0$ and the scale value is $v_0 = 1$. With a scale counter n for each pixel $x_{1,i}$, (9) and (8) simplify to $v_n = \sqrt{n+1}$ and $a_n = \frac{1}{\sqrt{n+1}}$, respectively.

The above mentioned condition of energy concentration is applicable only to pictures of the first level of the temporal decomposition. At the first level, each pixel $x_{2,j}$ of the picture $\mathbf{x}_2$ is used only once during the transform process. Therefore, there is no need to maintain a scale counter for the pixels in the picture $\mathbf{x}_2$. But at the second level of temporal decomposition, both pictures $\mathbf{x}_1$ and $\mathbf{x}_2$ are temporal low-bands resulting from transforms at the first level. Therefore, we need to consider scale factors v_{n_1} and v_{n_2} for the pixels $x_{1,i}$ and $x_{2,j}$ at higher levels of the temporal decomposition.

Let v_{n_1} and v_{n_2} be the scale factors for the pixels $x_{1,i}$ and $x_{2,j}$ that have been processed previously by T_τ, where $\tau < \kappa$. Let v_m be the scale factor for the pixel $x_{1,i}$ after it has been processed by T_κ. There is no need for an updated scale counter which corresponds to the pixels $x_{2,j}$ as they are touched only once during the transform process. To determine the decorrelation factors a_n of a transform T_κ at higher levels of the dyadic decomposition, we assume identical pictures of constant intensity as the input of the transforms at the first level. Again, we design a transform at a higher level of the dyadic decomposition such that the resulting "to be high-band" $\mathbf{x}_2^{(\kappa+1)}$ is zero. Now, the pixels $x_{1,i}$ at higher levels of the dyadic decomposition are processed by T_κ as follows:

$$\begin{pmatrix} v_m x_{1,i} \\ 0 \end{pmatrix} = \frac{1}{\sqrt{1+a_n^2}} \begin{pmatrix} 1 & a_n \\ -a_n & 1 \end{pmatrix} \begin{pmatrix} v_{n_1} x_{1,i} \\ v_{n_2} x_{1,i} \end{pmatrix} \tag{10}$$

The condition of energy concentration is satisfied if

$$a_n = \frac{v_{n_2}}{v_{n_1}} \quad \text{and} \tag{11}$$

$$v_m = \sqrt{v_{n_1}^2 + v_{n_2}^2}. \tag{12}$$

According to the definition of the scale counters at the first level of dyadic decomposition, the resulting scale factors are $v_{n_1} = \sqrt{n_1+1}$ and $v_{n_2} = \sqrt{n_2+1}$. For the second level, the condition of energy concentration in (12) requires the scale factors to satisfy $v_m = \sqrt{n_1+1+n_2+1}$. This result allows us to extend the definition of the scale counter to be applicable to any level of a dyadic decomposition if the scale counter m of the next higher level of dyadic decomposition satisfies the following *scale counter update rule*:

$$m = n_1 + n_2 + 1 \tag{13}$$

Consequently, the scale factor at the second level is $v_m = \sqrt{m+1}$ with the scale counter update rule. Moreover, the relation between the scale counter and the scale factor is valid for any level of a dyadic decomposition.

$$v_n = \sqrt{n+1} \tag{14}$$

This allows us to state the condition of energy concentration in terms of the scale counter. The decorrelation factor is

$$a_n = \frac{\sqrt{n_2+1}}{\sqrt{n_1+1}} \tag{15}$$

with the scale counters n_1 and n_2 which are maintained according to (13).

III. Adaptive Spatial Transform

The spatial transform that further decomposes the temporal low-band has to consider the scale factors that have been used during the temporal decomposition. In the following, we outline an adaptive spatial transform that is suitable.

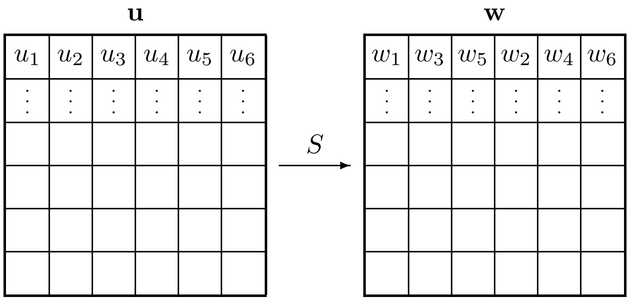

u
w
u_1 u_2 u_3 u_4 u_5 u_6
S
w_1 w_3 w_5 w_2 w_4 w_6

Fig. 2. The adaptive horizontal transform S for a temporal low-band $\mathbf{u}$.

Fig. 2 depicts the temporal low-band $\mathbf{u}$ and its horizontal decomposition $\mathbf{w}$. Let u_{2r+1} and u_{2r+2} be the odd and even horizontal samples of the the temporal low-band $\mathbf{u}$. The adaptive spatial transform S maps these pixels according to

$$\begin{pmatrix} w_{2r+1} \\ w_{2r+2} \end{pmatrix} = S \begin{pmatrix} u_{2r+1} \\ u_{2r+2} \end{pmatrix} \tag{16}$$

into spatial low- and high-band coefficients w_{2r+1} and w_{2r+2}, respectively. The transform matrix S is the same as the matrix H in (5). The decorrelation factor a_n is also determined by (15), where n_2 denotes the scale counter of the even pixels and n_1 that of the odd pixels in the picture $\mathbf{u}$. The scale counter update rule for the spatial decomposition is also given by (13), where m depicts the updated scale counter for the odd pixels in the picture $\mathbf{w}$.

After updating the scale counter, the scale counter for the even pixels in the picture $\mathbf{w}$ are set to zero. Consequently, a standard Haar transform ($a_n = 1$) is applied to obtain the horizontal and diagonal subbands. But an adaptive vertical transform is used to obtain the low-pass and the vertical subband. The adaptive vertical transform is analogous to the adaptive horizontal transform.

IV. Experimental Results and Discussion

Experimental results assessing the energy compaction are obtained for the QCIF sequences *Foreman*, *Bus*, and *Mother & Daughter*. Our coding scheme with the motion-compensated orthogonal transform is compared to schemes which use a motion-compensated lifted Haar wavelet with and without update step. In addition, the performance of closed loop coding with hierarchical P pictures is reported.

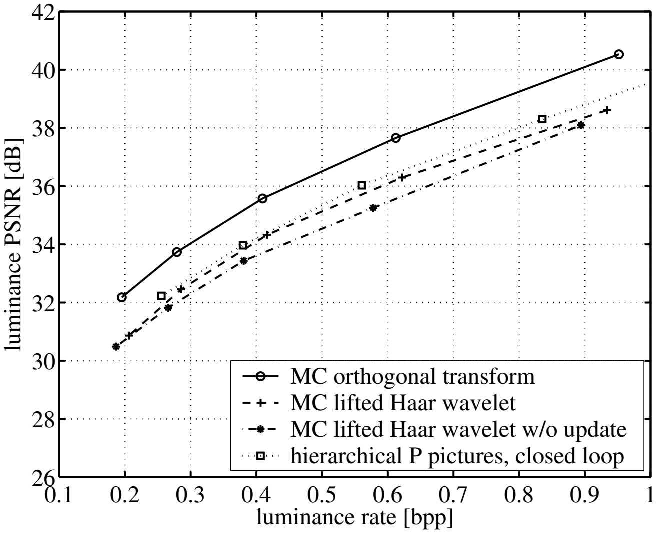


Fig. 3. PSNR over the rate for the luminance signal of the QCIF sequence *Foreman* at 30 fps with 288 frames.

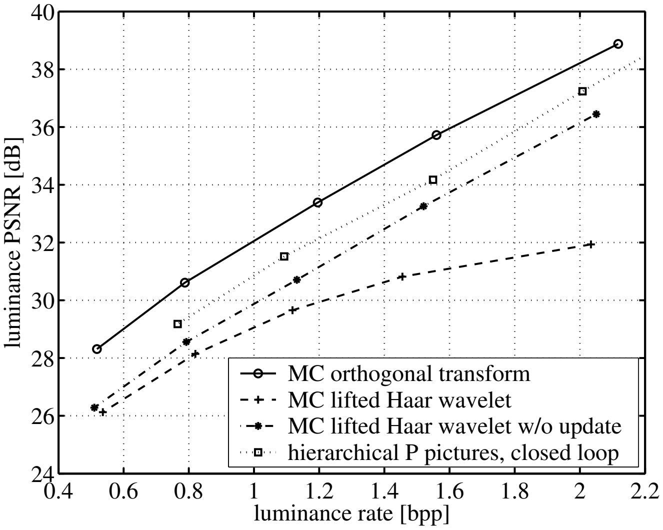


Fig. 4. PSNR over the rate for the luminance signal of the QCIF sequence *Bus* at 15 fps with 64 frames.

For the coding process with the orthogonal transform, a scale counter n is maintained for every pixel of each picture. The scale counters are an immediate results of the utilized motion vectors and are only required for the processing at encoder and decoder. The scale counters do not have to be encoded as they can be recovered from the motion vectors.

All schemes operate with a GOP size of 16 frames. The block size for motion compensation is limited to 8×8. For simplicity, the resulting temporal subbands are coded with JPEG 2000. The temporal high-bands are coded directly, whereas the temporal low-band is re-scaled with (14) before encoding. For optimal rate allocation, Lagrangian costs are determined where the distortion term considers the scale factors applied to the temporal low-band.

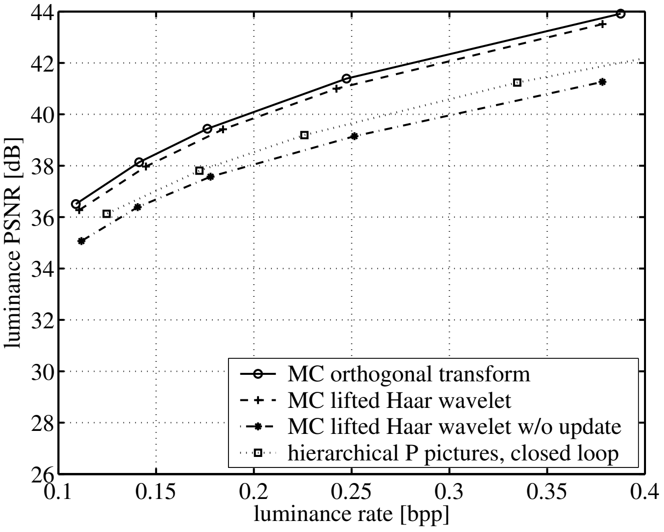


Fig. 5. PSNR over the rate for the luminance signal of the QCIF sequence *Mother & Daughter* at 30 fps with 64 frames.

Figs. 3, 4, and 5 depict the rate distortion performances for the luminance signals of the test sequences. Results for the orthogonal transform, the lifted Haar wavelet with and without update step, as well as for closed loop coding with hierarchical P pictures are given. For all test sequences, the orthogonal transform outperforms the reference schemes. For *Bus*, the significant motion in the sequence degrades the performance of the lifted Haar wavelet as the update step introduces substantial noise. For *Mother & Daughter*, the orthogonal transform outperforms the lifted Haar wavelet with update step by a small margin as the weak motion in the sequence does not substantially harm the lifted Haar wavelet.

V. Conclusions

This paper presents a motion-compensated orthogonal transform which strictly maintains orthogonality for any motion field. It outperforms the motion-compensated lifted Haar wavelet which is not able to maintain orthogonality in all cases. The orthogonality principle improves energy compaction and provides a highly robust video representation.

Acknowledgement

The authors thank Aditya Mavlankar for helpful discussions.

References

[1] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, Sept. 1994.

[2] S.-J. Choi and J. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[3] 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*, vol. 3, Salt Lake City, UT, May 2001, pp. 1793–1796.

[4] 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*, vol. 2, Thessaloniki, Greece, Oct. 2001, pp. 1029–1032.

[5] 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.

[6] B. Girod and S. Han, "Optimum update for motion-compensated lifting," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 150–153, Feb. 2005.