CONTENT ADAPTIVE WAVELET LIFTING FOR SCALABLE LOSSLESS VIDEO CODING

Daniela Lanz, Christian Herbert, and André Kaup

Multimedia Communications and Signal Processing
Friedrich-Alexander-University Erlangen-Nürnberg (FAU)
Cauerstr. 7, 91058 Erlangen, Germany
Email: {Daniela.Lanz,Christian.Herbert,Andre.Kaup}@FAU.de

ABSTRACT

Scalable lossless video coding is an important aspect for many professional applications. Wavelet-based video coding decomposes an input sequence into a lowpass and a highpass subband by filtering along the temporal axis. The lowpass subband can be used for previewing purposes, while the highpass subband provides the residual content for lossless reconstruction of the original sequence. The recursive application of the wavelet transform to the lowpass subband of the previous stage yields coarser temporal resolutions of the input sequence. This allows for lower bit rates, but also affects the visual quality of the lowpass subband. So far, the number of total decomposition levels is determined for the entire input sequence in advance. However, if the motion in the video sequence is strong or if abrupt scene changes occur, a further decomposition leads to a low-quality lowpass subband. Therefore, we propose a content adaptive wavelet transform, which locally adapts the depth of the decomposition to the content of the input sequence. Thereby, the visual quality of the lowpass subband is increased by up to 10.28 dB compared to a uniform wavelet transform with the same number of total decomposition levels, while the required rate is reduced by 1.06% additionally.

***Index Terms*—** Lossless Coding, Scalability, Discrete Wavelet Transform, Motion Compensation

1. INTRODUCTION

Many professional tasks like surveillance systems and telemedicine applications require lossless compression due to their sensitive content. However, lossless compression naturally leads to high bit rates. Considering any wireless network with limited channel capacity, a fast transmission of the entire data is challenging. Therefore, scalable lossless video coding is desirable, which allows for transmitting a base layer (BL) with coarser quality in the first instance and afterwards one or more enhancement layers (ELs), comprising the residual video data, to reconstruct the original sequence without any loss. Basically, three different types of video scalability can be distinguished. Temporal scalability affects the frame rate, spatial scalability controls the spatial resolution, and quality scalability manipulates the fidelity of the video. Beside DCT-based coding schemes like Scalable High Efficiency Video Coding (SHVC) [1] and Sample-Based Weighted Prediction for Enhancement Layer Coding (SELC) [2], also 3-D subband coding (SBC) [3] can be applied. 3-D SBC is based on Wavelet Transforms (WT), which naturally provide scalability features without additional overhead [4]. As shown in Fig. 1, by a transformation in temporal direction, the signal is decomposed into a lowpass (LP) and a highpass (HP) subband. Both subbands offer only half the frame rate compared to the original sequence. While the LP subband is very similar to the original signal, the HP subband contains the structural information of the video sequence. Afterwards, every frame of each subband is coded by the wavelet-based coder JPEG 2000 [5], resulting in a fully scalable BL-EL-representation.

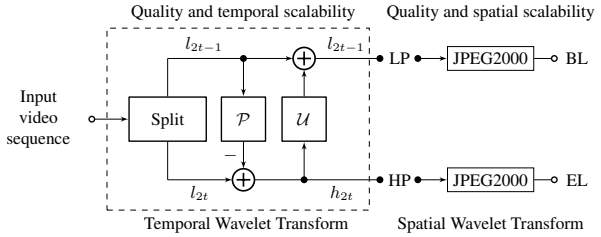

Fig. 1: Considered scenario to achieve a fully scalable representation of a video sequence. The dashed box shows the lifting structure of the wavelet transform for one decomposition level.

In this work, we focus on the optimization of the temporal scalability, which is controlled by the temporal WT highlighted by the dashed box in Fig. 1. The recursive application of the WT to the LP subband of the previous stage halves the frame rate for every decomposition level. This is advantageous for similar frames of the video sequence. In contrast, if huge changes occur among subsequent frames, the visual quality suffers significantly from multiple decomposition levels. This is why motion compensation (MC) should be included into the WT. However, MC always leads to a higher entire rate, mainly caused by the motion information, which has to be transmitted as additional overhead [6]. Further, there exists no practical approach for perfect MC. Hence, the error propagation will increase for a higher number of decomposition levels, leading to an inferior visual quality of the LP subband. Therefore, the temporal scaling should be adapted to the video sequence. This can be reached by our proposed content adaptive wavelet lifting (CA-WL), which provides fine temporal resolution for high dynamic parts of a video sequence, while parts with few changes among subsequent frames are resolved coarser. After a brief overview of 3-D SBC, the proposed CA-WL is described in detail. Simulation results are given in the next section, followed by a short conclusion and outlook at the end of the paper.

2. 3-D SUBBAND CODING

An efficient implementation of the discrete WT was proposed by Sweldens [7]. The so-called lifting structure consists of three steps: split, predict, and update. The block diagram of the lifting structure for a decomposition in temporal direction is depicted in the dashed

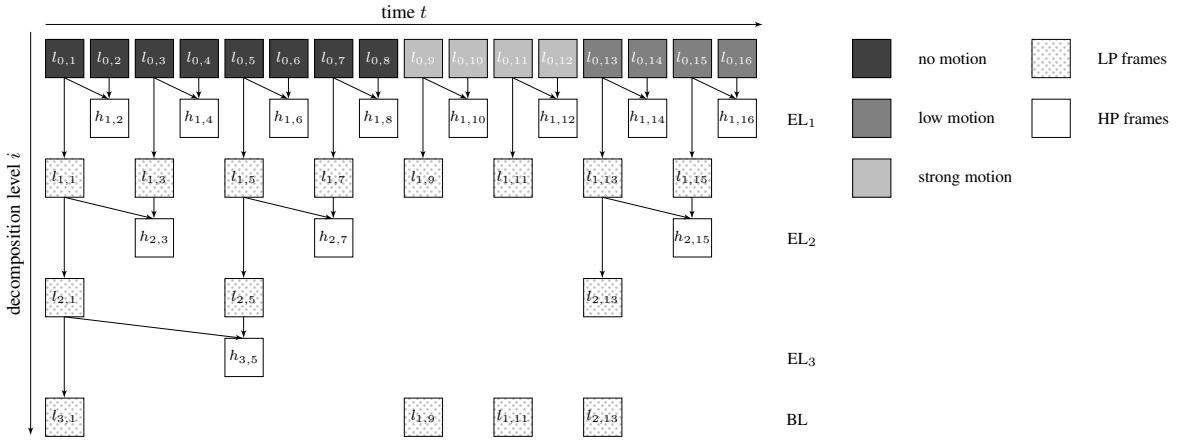

Fig. 2: Basic decomposition scheme of the CA-WL for $i_{\max}$=3 decomposition levels resulting in one BL and three ELs. Depending on the underlying motion in the original sequence, the local depth of the decomposition differs.

box in Fig. 1. In the first step, splitting is performed by decomposing the input signal into even- and odd-indexed frames l_{2t} and l_{2t-1}. In the second step, the even frames are predicted from the odd frames by a prediction operator $\mathcal{P}$. Subtracting the predicted values $\mathcal{P}(l_{2t-1})$ from the even frames, results in the HP coefficients h_{2t}. In the third step, the HP coefficients are filtered by an update operator $\mathcal{U}$ and are added back to the odd frames, resulting in the LP coefficients. To get coarser temporal resolutions, the lifting scheme can be iterated on the LP subband by $i_{\max} = \log_2(T)$ decomposition levels, where T equals the total number of original frames.

Since the lifting structure offers a flexible framework, it can be modified in multiple ways. By introducing rounding operators as introduced in [8], integer to integer transforms can be achieved, which yield perfect reconstruction. This makes the lifting structure of the WT highly attractive for many professional applications by offering a scalable representation and lossless reconstruction at the same time. However, due to temporal displacements in the sequence, blurriness and ghosting artifacts will appear in the LP subband. These can be alleviated by incorporating MC methods directly into the lifting structure without losing the property of perfect reconstruction. This is called motion compensated temporal filtering (MCTF) [9] and can be achieved by realizing the prediction operator $\mathcal{P}$ by the warping operator $\mathcal{W}_{2t-1\to 2t}$. Instead of the original odd frames, a compensated version is subtracted from the even frames. In case of the Haar wavelet filters, the prediction step is given by

$$h_{2t} = l_{2t} - \lfloor \mathcal{W}_{2t-1\to 2t}(l_{2t-1}) \rfloor. \tag{1}$$

However, to achieve an equivalent wavelet transform, the compensation has to be inverted in the update step [10]. By reversing the index of $\mathcal{W}$, the LP coefficients of the Haar transform can be calculated by

$$l_{2t-1} = l_{2t-1} + \left\lfloor \frac{1}{2} \mathcal{W}_{2t\to 2t-1}(h_{2t}) \right\rfloor. \tag{2}$$

$\mathcal{W}$ can be realized by different approaches of MC. In this work, we will employ a block-based approach.

3. CONTENT ADAPTIVE WAVELET LIFTING

Considering video sequences from surveillance systems or medical data sets, which comprise a temporal acquisition of images with contrast medium, there will be parts, where almost no motion occurs over time. In this case, an adaptive temporal scaling is advantageous, which performs iteratively a further decomposition, if subsequent frames are similar enough. If there are no changes over several frames, they shall be represented by only one LP frame. For significant changes among subsequent frames, for example when the contrast medium gets visible, these changes shall be represented in the LP subband with finer temporal resolution.

Fig. 2 shows the basic approach of our proposed content adaptive wavelet transform. Index i indicates the number of the current decomposition level. For i=0, no decomposition has been done so far, which corresponds to the original video sequence. In the first row, a schematic video sequence is given, which consists of three sections, each with a different amount of moving content. The corresponding amount of motion is described by the legend on the right side of Fig 2. While in this example the first decomposition is performed for the entire sequence, the second decomposition is performed only on the frames with no or low motion. The third decomposition is exclusively done on frames with no motion. The resulting BL and ELs are marked at the right side. Since the maximum decomposition level $i_{\max}$ is equal to 3, three ELs are generated. By combining the ELs with the BL at the decoder side, the original sequence can be reconstructed step by step without any loss.

3.1. Calculation of the Stopping Criterion

Haar WTs can be represented with tree structures [4]. For 3-D SBC, the tree structure is given by decomposing two subsequent frames into LP and HP frames. To realize the CA-WL, the costs of the single nodes in every tree have to be considered. If the combined costs of the child nodes exceed the costs of the parent node, this means for an arbitrary signal s, if

$$\mathcal{C}(s_{i,[2t-1,2t]}) \leq (\mathcal{C}(s_{i+1,2t-1}) \cup \mathcal{C}(s_{i+1,2t})) \tag{3}$$

holds, then the child nodes shall be pruned from the tree. Here, $\mathcal{C}(\cdot)$ describes a cost functional, which represents the coding costs, such as entropy [11] or rate-distortion [12]. In this work, every decomposition level is performed for the entire input sequence in advance, before a retrospective evaluation of the resulting costs is done. Further, we decided to use a rate-distortion-based approach for calculating the coding costs. Therefore, we formulate the Lagrangian cost

Table 1: For the evaluation, sequences from surveillance systems, medical applications, and the HEVC test data set are employed. All sequences are used in 4:0:0 color sub-sampling format.

		Spatial resolution	Number of frames
Surv	*AirportNight1*	688×352	500
	AirportNight2	688×432	500
	AirportNight3	688×372	500
	AirportDay1	688×432	500
Med	*MedFrontal*	512×512	29
	MedSagittal	512×512	29
HEVC	*ClassC*	832×480	300
	ClassD	416×240	300

functional for a signal s

$$C(s) = D(s) + \lambda R(s). \tag{4}$$

To determine the distortion $D(s)$ of the resulting LP frame, we calculate the MSE of the corresponding wavelet coefficients compared to the original signal according to [13]. In this work, not only the similarity of the LP frame to the odd-indexed frame, but also the similarity to the even indexed frame is considered, which is a very important aspect for many professional applications. The rate $R(s)$ is composed of the required rate for lossless coding of the LP and HP frames and, in case of MC, the file size of the motion vectors. Then, the current decomposition can be evaluated locally by comparing the R-D costs of the children to the costs of the parent node for a given value λ. If the R-D costs of the child nodes exceed the costs of the parent node, thus if the following inequality

$$\begin{aligned} &D(l_{i,[2t-1,2t]}) + \lambda R(l_{i,[2t-1,2t]}) \leq \\ &(D(l_{i+1,2t-1}) + D(h_{i+1,2t})) + \lambda(R(l_{i+1,2t-1}) + R(h_{i+1,2t})) \end{aligned} \tag{5}$$

holds, then a further decomposition is prevented. The Lagrange multiplier λ can be any positive value. By choosing large values for λ, the rate is decreased, while for small values the distortion is decreased.

3.2. Handling of the Overhead

For lossless reconstruction, the decomposition depth for every part of the input sequence has to be transmitted additionally. Therefore, a vector $\boldsymbol{v}$ is generated, whose length corresponds to T. This vector is initialized with zeros and gets an increment of 1 at every temporal position of a LP frame after one decomposition level. The position to the corresponding HP frame is set to zero. Consequently, the non-zero entries correspond to the number of applied decomposition levels i for every temporal position of a LP frame, while the distance d to the corresponding HP frame is given by $d{=}2^{i-1}$. For the schematic video sequence in Fig. 2, vector $\boldsymbol{v}$ is generated as follows:

$$\begin{aligned} &\text{Initialize } \boldsymbol{v}: && (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \\ &\boldsymbol{v} \text{ after level } i{=}1: && (1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0) \\ &\boldsymbol{v} \text{ after level } i{=}2: && (2,0,0,0,2,0,0,0,1,0,1,0,2,0,0,0) \\ &\boldsymbol{v} \text{ after level } i{=}3: && (3,0,0,0,0,0,0,0,1,0,1,0,2,0,0,0) \end{aligned}$$

The entire vector $\boldsymbol{v}$ is encoded using Context Adaptive Binary Arithmetic Coding (CABAC) [14] and is transmitted to the decoder side. Then, for lossless reconstruction of the previous stage, the decoder can easily determine the decomposition level and the temporal positions of the LP and HP frames regarding the original video sequence.

Table 2: Absolute $\text{PSNR}_{\text{LP}_t}$ [dB] and relative rate [%] differences of our proposed CA-WL compared to the U-WL with (bottom) and without (top) block-based MC. Positive numbers denote a better visual quality and a higher rate of our proposed CA-WL and vice versa.

		λ	Surv	Med	ClassC	ClassD	Total average
No MC	$\Delta\ \text{PSNR}_{\text{LP}_t}$	1	4.12	5.28	12.83	18.07	**8.88**
		3	1.64	1.91	6.32	11.4	**5.3**
		5	0.97	1.16	3.73	8.89	**3.67**
		7	0.65	1.16	4.09	8.27	**3.5**
	Δ File size	1	5.99	0.09	11.64	9.07	**6.56**
		3	0.8	-0.96	2.53	5.77	**2.18**
		5	0.23	-1.29	0.84	4.04	**1.08**
		7	0.16	-1.29	0.25	3.07	**0.67**
Block-based MC	$\Delta\ \text{PSNR}_{\text{LP}_t}$	1	9.3	15.56	6.99	14.15	**10.98**
		3	8.17	13.89	9.6	11.26	**10.28**
		5	7.42	13.89	9.21	9.55	**9.47**
		7	7.27	13.89	8.95	8.42	**9.02**
	Δ File size	1	0.16	-5.58	1.98	6.9	**1.34**
		3	-0.52	-5.64	-1.7	1.35	**−1.06**
		5	-0.69	-5.64	-1.96	0.65	**−1.38**
		7	-0.8	-5.64	-2.18	0.29	**−1.57**

4. EXPERIMENTAL RESULTS

Our simulation setup comprises surveillance videos, medical sequences with contrast medium, and natural sequences from the HEVC test data set [15]. The dimensions are summarized in Table 1. The bit depth for all sequences constitutes 8 bits per sample. All surveillance sequences are characterized by a static background and some moving objects in the foreground. The medical sequences origin from Digital Subtraction Angiography (DSA), showing the inflow of a contrast medium into a human cranium in frontal and sagittal perspectives.

In the following, we will compare our proposed CA-WL to a uniform wavelet lifting (U-WL) with the same number of total decomposition levels. The single frames of each subband are encoded by JPEG 2000, using the OpenJPEG [16] implementation with four spatial wavelet decomposition steps in xy-direction. Further, we evaluate the CA-WL and the U-WL with and without a block-based MC, respectively. For block-based MC, the block size is set to 8, while the search range starts with a size of 8 and is doubled for every decomposition level until a maximum size of 64. The increasing search range is important, since the input frames of higher decomposition levels have a larger temporal distance, which has to be covered. To keep the computational effort realistic, we limit the increment of the search range by 64 pixels. The resulting motion vectors are encoded using the QccPack library [17]. Then, the entire file size is composed of the rate resulting from each subband, the required motion vectors and the coding costs for transmitting vector $\boldsymbol{v}$. The visual quality of the resulting LP subband is measured by the same metric as already used in Section 3.1, but in terms of $\text{PSNR}_{\text{LP}_t}$ [13].

Table 2 gives the differences regarding $\text{PSNR}_{\text{LP}_t}$ in [dB] and the entire file size in [%] of our proposed method compared to the U-WL for all data sets with and without the application of MC and for different values of λ. As can be seen in the right column, our method always results in a better visual quality compared to the U-WL. By increasing λ, the file size is reduced, while the $\text{PSNR}_{\text{LP}_t}$ gains are also decreased. However, for $\lambda{=}7$, the $\text{PSNR}_{\text{LP}_t}$ gains are still positive. By including MC into both methods, we are able to get a lower rate than for the U-WL, resulting in positive $\text{PSNR}_{\text{LP}_t}$ gains at the same time. For $\lambda{=}3$, the file size can be reduced by up

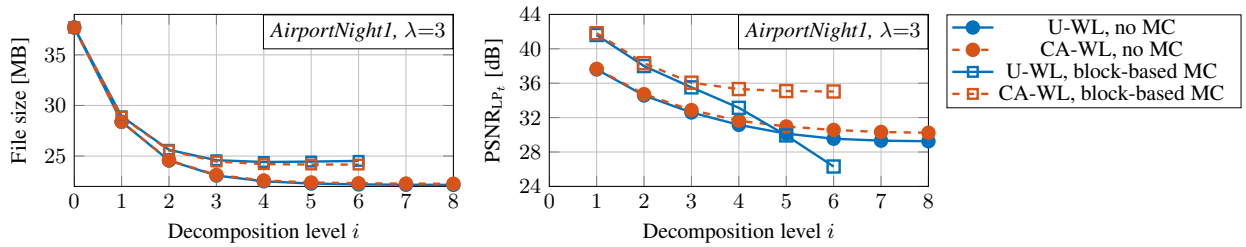

Fig. 3: Absolute rate and $\text{PSNR}_{\text{LP}_t}$ results from the *AirportNight1* sequence with and without MC, using λ=3. The results are displayed over all reached decomposition levels i. The proposed method is characterized by the dashed lines.

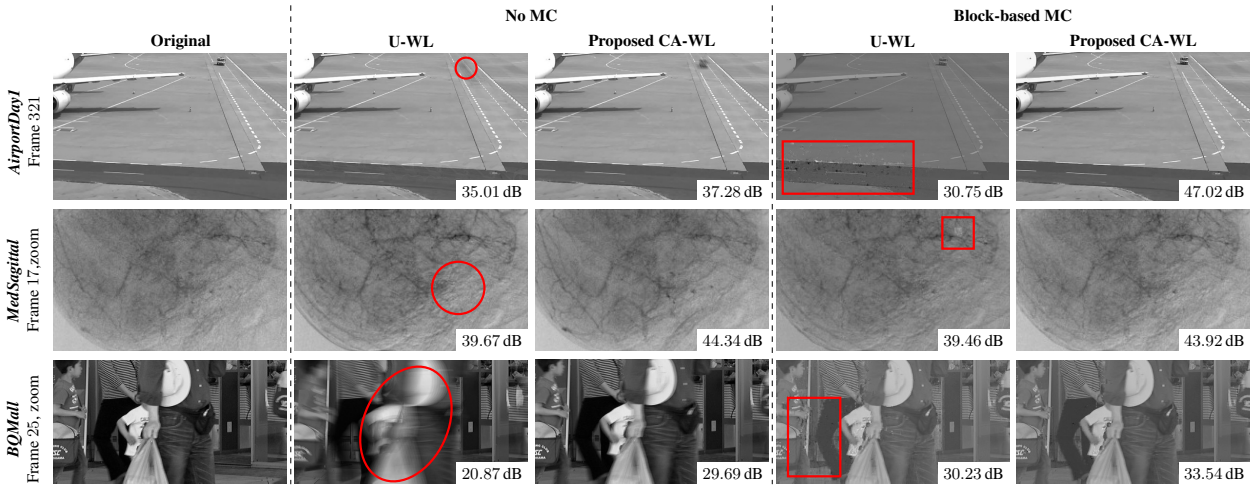

Fig. 4: Comparison of the visual quality of one frame from each test data set compared to the corresponding LP frames of a U-WL and our CA-WL with and without block-based MC, for $\lambda = 3$. The rectangles depict blocking artifacts, the circles indicate missing objects, and the ellipses show blurring artifacts.

to 1.06% in total average, while the visual quality is increased by 10.28 dB, as the right column of Table 2 shows.

To demonstrate the performance of our proposed CA-WL in more detail, Fig. 3 presents the absolute rate and $\text{PSNR}_{\text{LP}_t}$ results from the *AirportNight1* sequence with and without MC, using λ=3. As the left plot offers, the entire file size for incorporating MC is always higher than omitting MC. However, the visual quality is significantly higher by including MC, which is very important for many professional applications. But for higher decomposition levels i, the error propagation due to imperfect MC is increasing. The right plot shows the strong decreasing $\text{PSNR}_{\text{LP}_t}$ results, which can be prevented by our proposed CA-WL. Simultaneously, the overall file size can be decreased, as the left plot of Fig. 3 shows.

In Fig. 4, some visual results for each data set are presented. From left to right, the reference frame and the corresponding LP frames from the U-WL and the proposed CA-WL are shown for a value of λ=3, without MC and with block-based MC, respectively. Disturbing artifacts and loss of content, resulting from the U-WL, are highlighted in every frame. The rectangles depict blocking artifacts, the circles indicate locations of objects, which are canceled out completely, and the ellipses show blurring artifacts. As the right column shows, the CA-WL is capable to compensate this lack of data fidelity efficiently and gives a reliable impression of the actual content of the sequence. This is also proven by the PSNR values given at the bottom right corner of each frame.

5. CONCLUSION

Scalable lossless video coding and a high visual quality of the corresponding BL is very important for many professional applications. Wavelet-based video coding provides full scalability without additional overhead. The temporal resolution can be controlled by the recursive application of the WT in temporal direction to the LP subband of the previous stage. This leads to lower bit rates, but if the content of the underlying video sequence comprises strong motion, the visual quality of the BL is degraded significantly. We proposed a method which locally adapts the temporal scaling by evaluating a Lagrangian cost functional in every transformation step and prevents further decomposition, if the costs of the current level are higher than the costs of the previous level. This way, we can increase the visual quality of the BL by 10.28 dB compared to the U-WL with block-based MC, while the required rate is reduced by 1.06% at the same time. Further work aims at the development of an algorithm to determine the optimum value of λ in a rate-distortion sense, based on the characteristics of the underlying sequence.

6. ACKNOWLEDGMENT

We gratefully acknowledge that this work has been supported by the Deutsche Forschungsgemeinschaft (DFG) under contract number KA 926/4-3.

7. REFERENCES

[1] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, Jan 2016.

[2] A. Heindel, E. Wige, and A. Kaup, "Low-complexity enhancement layer compression for scalable lossless video coding based on HEVC," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1749–1760, Aug 2017.

[3] G. Karlsson and M. Vetterli, "Three dimensional sub-band coding of video," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, New York City, NY, USA, Apr 1988, vol. 2, pp. 1100–1103.

[4] J. Garbas, B. Pesquet-Popescu, and A. Kaup, "Methods and tools for wavelet-based scalable multiview video coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 2, pp. 113–126, Feb 2011.

[5] ITU-T and ISO/IEC, "JPEG 2000 Image Coding System: Core Coding System," in *ITU-T Rec. T.800 and ISO/IEC 15444-1:2004*, Sep 2004.

[6] W. Schnurrer, N. Pallast, T. Richter, and A. Kaup, "Temporal scalability of dynamic volume data using mesh compensated wavelet lifting," *IEEE Trans. on Image Processing*, vol. 27, no. 1, pp. 419–431, Jan 2018.

[7] W. Sweldens, "Lifting scheme: a new philosophy in biorthogonal wavelet constructions," in *Proc. SPIE Int. Symp. on Optical Science, Engineering, and Instrumentation*, San Diego, CA, USA, Sep 1995, vol. 2569, pp. 68–79.

[8] A.R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, "Lossless image compression using integer to integer wavelet transforms," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Oct 1997, vol. 1, pp. 596–599.

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

[10] N. Bozinovic, J. Konrad, W. Zhao, and C. Vazquez, "On the importance of motion invertibility in MCTF/DWT video coding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, Mar 2005, pp. 49–52.

[11] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 713–718, March 1992.

[12] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. on Image Processing*, vol. 2, no. 2, pp. 160–175, Apr 1993.

[13] D. Lanz, J. Seiler, K. Jaskolka, and A. Kaup, "Compression of dynamic medical CT data using motion compensated wavelet lifting with denoised update," in *Proc. IEEE Picture Coding Symposium (PCS)*, San Francisco, CA, USA, June 2018, pp. 1–5.

[14] Ian H. Witten, Radford M. Neal, and John G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, June 1987.

[15] F. Bossen, "Common test conditions and software reference configurations," *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, Jan 2013.

[16] A. Descampe, F. Devaux, H. Drolon, D. Janssens, and Y. Verschueren, "OpenJPEG 2.0.0," Nov 2012.

[17] J.E. Fowler, "Qccpack: An open-source software library for quantization, compression, and coding," in *Proc. SPIE Applications of Digital Image Processing XXIII*, San Diego, CA, USA, Aug 2000, vol. 4115, pp. 294–301.