WAVELET-DOMAIN DISTRIBUTED VIDEO CODING BASED ON CONTINUOUS-VALUED SYNDROMES

Lorenzo Cappellari

Dept. of Information Engineering, University of Padova
via Gradenigo 6/B, 35131 Padova, Italy
email: lorenzo.cappellari@dei.unipd.it

ABSTRACT

In current video codecs, the temporal correlation between successive frames of a video sequence is exploited by means of predictive coding. Successful utilization of this paradigm in a wireless and mobile scenario is impaired due to both the computationally-demanding encoder-side motion-compensation procedure and the severe degradation of the decoded video quality when losses of encoder-decoder synchronization occur. The application of distributed source coding principles has been recently proposed as solution to both problems. In this paper, a wavelet-domain video coder is investigated that uses continuous-valued syndromes to efficiently code the coefficients at the encoder. Motion compensation, to form the side-information that allows for reconstruction, takes instead place at the decoder. The proposed video coder, whose performance is similar to other distributed video coders found in literature, does not require any feedback channel and produces a spatial-quality scalable stream.

1. INTRODUCTION

The growing popularity of video streaming applications has brought the need for computationally light, robust and scalable video compression. To solve the lightness and the robustness issues, many research groups have investigated the feasibility of application of distributed source coding principles [1] to video coding. As opposed to traditional coding, in distributed video coding the frames of the video sequence are seen as correlated sources than cannot communicate each other, and are coded independently using a *light* encoder. Many of the current proposals perform this kind of compression into the Discrete Cosine Transform (DCT) domain [2, 3], and try to reuse some of the tools employed by current video codecs such as, for example, H.264/AVC. Other solutions, e.g. [4], compress the video sequence into the Discrete Wavelet Transform (DWT) domain, achieving at the same time spatial scalability.

In this paper we propose a wavelet-domain distributed video coding solution based on continuous-valued syndromes [5]. The rest of the paper is organized as follows. In Section 2 the principles of distributed source coding using continuous-valued syndromes are reviewed. The experimental correlation analysis of real video data in the wavelet domain is then presented in Section 3. This analysis procedure is then taken into account in the design of the proposed video coder, that is described in Section 4. Experimental results that show the effectiveness of the proposal are presented in Section 5. Conclusions and further remarks are finally drawn in Section 6.

2. DISTRIBUTED SOURCE CODING USING CONTINUOUS-VALUED SYNDROMES.

The theoretical possibility to encode the output of a random source $\mathbf{X}$ in presence of a correlated side-information $\mathbf{Y}$ at the decoder as efficiently as if $\mathbf{Y}$ was known as well at the encoder was theoretically investigated in the seventies [1]. If $\mathbf{X}$ and $\mathbf{Y}$ are stationary, memoryless, and jointly Gaussian, there is such a possibility. Among the practical schemes approaching this result, the system described in [6] (distributed source coding using syndromes, DISCUS) uses syndrome-based coding to reach this objective.

Essentially, words $X \in \mathbb{R}^N$ formed by N realizations from the source are pre-quantized into a discrete set $\mathscr{C} \subset \mathbb{R}^N$, which is called the *source code*, that has a certain induced group structure. By means of a certain *channel code* $\mathscr{C}_0$, this group is in turn partitioned into a *finite* number of isomorphic *cosets* $\mathscr{C}_i \subset \mathscr{C}$, indexed by some bit-strings i. Upon encoding, the value of S such that the quantized value of X (denoted by $W \in \mathscr{C}$) belongs to $\mathscr{C}_S$, which in the channel coding lingo may be called (or identified by) the *syndrome*, is sent at the decoder. The decoder recovers an estimate of W by quantizing the side information Y into the coset $\mathscr{C}_S$. If the minimum Euclidean distance between the points of the channel code $\mathscr{C}_0$ (and between the points of any coset $\mathscr{C}_i$, $i > 0$) is greater then twice the Euclidean distance between Y and W, then this closest point search reveals the right value of W, from which X can be estimated.

However, it is as well possible to partition the Euclidean space $\mathbb{R}^N$ itself into an *infinite* number of isomorphic cosets $\mathscr{C}_\lambda \subset \mathbb{R}^N$, and consider the value of S such that $X \in \mathscr{C}_S$ as a *continuous-valued* syndrome [5]. As suggested in [5], the quantization error relative to the quantization of the vector X using trellis coded quantization (TCQ) [7] can be for example identified as S. Upon receiving an approximation $\tilde{S}$ of S, the decoder simply quantizes Y into $\mathscr{C}_{\tilde{S}}$. If the minimum Euclidean distance between the points of $\mathscr{C}_\lambda$, $\forall\lambda$, is greater then twice the Euclidean distance between Y and $X + (\tilde{S} - S)$, then this closest point search exactly reveals the value of $X + (\tilde{S} - S)$, i.e. gives a close estimate of X.

The advantages of the continuous-domain solution come essentially from the fact that the encoding is made by two almost independent steps. In the first step, a continuous-valued syndrome is in fact formed depending only on the *current* expected correlation between the variable to be coded X and the side information Y. Then, the second step, that codes the syndrome, is only driven by the *current* desired transmission rate, and can easily adapt to changing transmission channel conditions. The scheme is in addition very suitable for quality scalable transmission, e.g. using embedded quantization for syndrome coding.

3. CORRELATION ANALYSIS

In the case that $X = \{x_i\}$ and $Y = \{y_i\}$ are jointly stationary random processes and the source X differs from the side information Y by an *innovation signal* $N = \{n_i\}$, i.e. $x_i = y_i + n_i$ with n_i and y_i uncorrelated, the best continuous-syndrome formation block should employ cosets $\mathscr{C}_\lambda$ with normalized cell-volume

$$V_{\mathscr{C}} = K \frac{\sigma_N^2}{G_{\mathscr{C}}} \quad , \tag{1}$$

where σ_N^2 is the innovation variance, $G_{\mathscr{C}}$ is the normalized second moment of the cells [8], and K depends on both the statistics of N and the cell-shape [5]. In particular, K is roughly 2 for independent Gaussian innovations and $\mathbb{Z}/4\mathbb{Z}$ TCQ-based syndrome-formation[1].

Assuming that a video decoder has correctly reconstructed some adjacent frames of a video sequence, similarly to what is done in [2, 9] a motion compensated prediction for the current frame can be formed that serves as side information Y for the corresponding encoder, which can use continuous-valued syndromes to *intra*-code the current frame X. However, there are mainly three differences between the real and the ideal case, namely:

1. it can still be assumed that $x_i = y_i + n_i$, with n_i and y_i roughly uncorrelated [10], but the statistics (in particular the variance) of the innovation signal is not stationary;
2. the statistics of N is not Gaussian;
3. adjacent values of N, n_i and n_{i+1}, are not independent nor uncorrelated [10].

To cope with these problems, it is possible to, respectively,

1. consider the normalized source samples $x'_i = x_i/\sigma_{n_i}$ (and normalized side samples $y'_i = y_i/\sigma_{n_i}$), where $\sigma^2_{n_i}$ is an estimate of the current innovation signal variance, such that $x'_i = y'_i + n'_i$ with $\sigma^2_{n'_i} = \sigma^2_{N'} = 1, \forall i$;
2. experimentally change the value of K to adapt to the non-Gaussian statistics of N';
3. consider as source and side samples their transformed-domain equivalents.

In this paper, with the objective to obtain a spatial scalable video sequence representation, it is proposed to code the image samples in the wavelet domain. The remaining of this Section then discusses how the innovation signal variance can be estimated at the encoder without actually knowing the side-information.

As suggested in [3], there certainly exist a relation between the innovation signal variances and the block-based MSE estimates between the current and the previous frame (both available at the encoder). The greater this estimate is, the farther the side information is expected to be from the source. Using a setup similar to the one in [9], in Fig. 1 the side block-MSE estimate (i.e. the quality of the motion compensated prediction used by the decoder as side information) is plotted versus the block-MSE estimate between the current and the previous frame (computed by the encoder with no needs for motion estimation or compensation), and it is evident that the experimental data supports this hypothesis.

[1] At low transmission rates, there is actually a dependence of K on the quality of the reconstructed syndrome; the value 2 refers to a rate of $1 \div 2$ bit/sample.

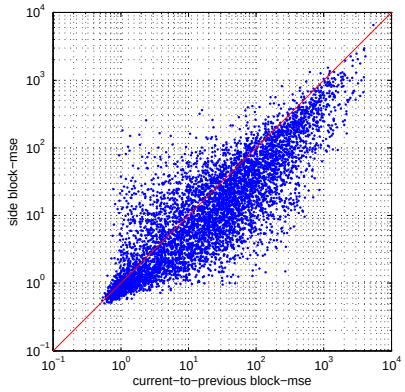


Figure 1: 16×16-block-MSE correlation between the side-to-original and the current-to-previous image quality, for the sequence carphone (QCIF).

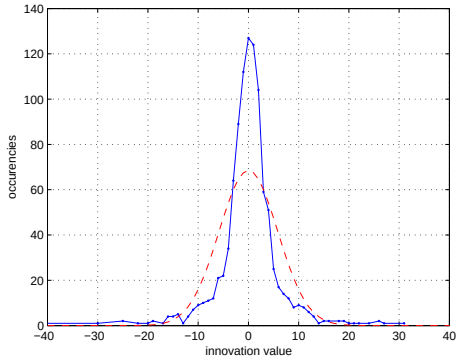


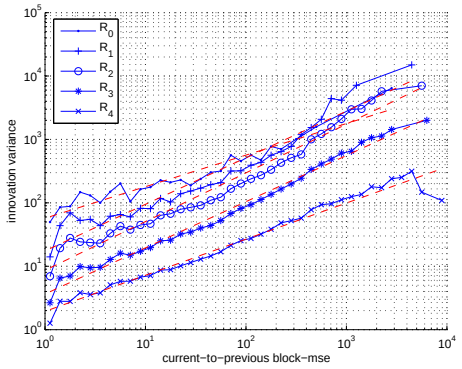
Figure 2: Experimental distribution of the innovation signal of the band HL_2, when the current-to-previous block-MSE equals about 56; the dashed curve is the Gaussian distribution with same mean and variance.

To go further inside this analysis, each DWT coefficient can be labelled with the current-to-previous block-MSE estimate corresponding to the image block to which the coefficient belongs, considering the interleaved wavelet-coefficient representation proposed by the standard JPEG2000 [11]. Then, the statistics of all the coefficients having the same label can be gathered (an example is provided in Fig. 2). In particular, the corresponding experimental innovation signal variance for all coefficients having the same label (and belonging to the same resolution R_i) is shown versus the label itself in Fig. 3.

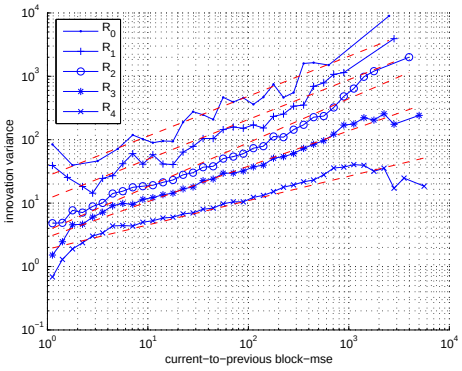
The almost linear relation (but different at each resolution) between the magnitude of these two quantities, suggests that it is reasonable to estimate the innovation variances $\sigma^2_{n_i}$ from the corresponding current-to-previous MSE-block estimate. As a remark, note that different video sequences have a very similar variance-label relationship, and hence that this estimate will be as good for any other natural source video sequence.

4. CODER DESCRIPTION

With the results of the previous Section in mind, the proposed syndrome-based distributed video coder, shown in Fig. 4, was easily devised. The main role is played by the encoder-

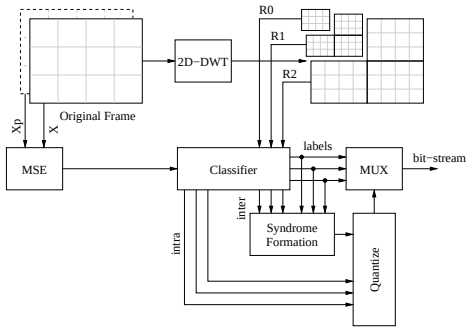


(a) carphone (QCIF)

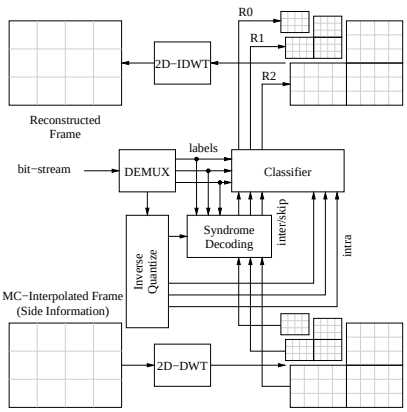


(b) foreman (CIF)

Figure 3: Innovation signal variances for different block-MSE classes, at the various resolutions.



(a) encoder



(b) decoder

Figure 4: Structure of the proposed video coder.

side classifier block, which, using the current-to-previous block-MSE estimates of the current frame, partitions the DWT coefficients of any resolution into many classes according to the estimated innovation variance $\sigma^2_{n_i}$. In particular:

- if $\sigma^2_{n_i}$ is under a certain threshold σ^2_m (that depends on a desired target quality), the coefficient is classified as *skip*, and no information at all is sent (other than the label);
- if $\sigma^2_{n_i}$ is over a certain threshold σ^2_M (equal to the mean expected variance of the source coefficients belonging to the same resolution), the coefficient itself, classified as *intra*, is coded;
- otherwise, depending on $\sigma^2_{n_i}$, the coefficient is classified into a certain number of *inter* classes, with the aim to use σ_{n_i} itself to normalize the coefficient before continuous-syndrome formation, as discussed in the previous Section.

As a remark, σ^2_M is set as described above because in that case it is more convenient to code the coefficient itself since its variance is smaller than the corresponding innovation signal variance. Fig. 5, that plots the estimated variances of the DWT coefficients at the various resolutions versus the coefficient label, shows that, as expected, there is essentially no dependence on the label. Again, it should be noted that different sequences have very similar mean coefficient variances. Moreover, since the labels have to be sent at the decoder and hence represent overhead information, only 4 *inter* coefficient classes are actually used in the experiments, each one using a different average value of σ_{n_i} for normalization.

The *intra* coefficients and the syndromes are then uniformly quantized according to a desired target rate (or quality). In particular, embedded scalar quantization has been used that permits to choose the actual rate during the decoding operation. Considering that the signal reconstructed by the syndrome decoder is the *normalized* signal x'_i, care is taken in quantizing the corresponding syndrome with a quantization step inversely proportional[2] to the *scaling factor* σ_{n_i}.

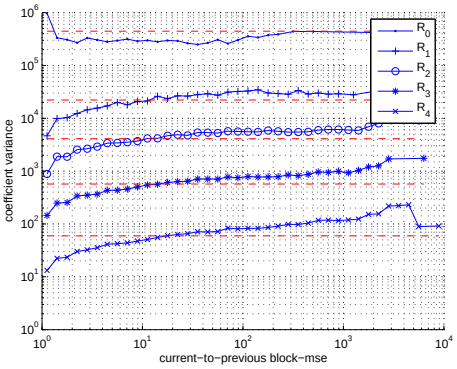
The decoder, upon receiving the labels and the quantized DWT coefficients relative to the desired resolutions (and relying on some motion compensated side information created, for example, as explained in the next Section), reconstructs the DWT coefficients for the current frame and inverts the transformation into the pixel-domain. In particular:

- the *skip* coefficients are directly taken from the side information;
- the *intra* coefficients are directly taken from the bitstream;
- the *inter* coefficients result from the syndrome-decoding procedure, and are then normalized-back through multiplication by σ_{n_i}.

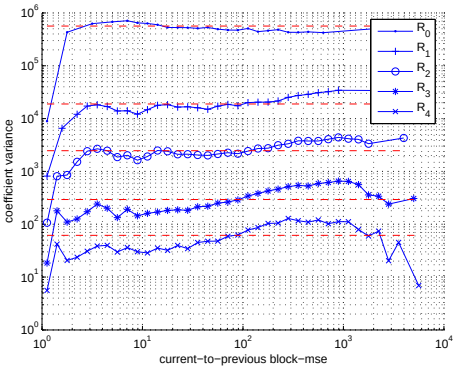
5. EXPERIMENTAL RESULTS

The experimental setup is similar to the one proposed in [9]. In particular, the even indexed frames are coded as *intra* (i.e. like if all the coefficients where classified as *intra*), while the remaining ones are coded with the algorithm

[2] All the wavelet coefficients are energy normalized in a way such that the same quantization step should be used across the different resolutions.



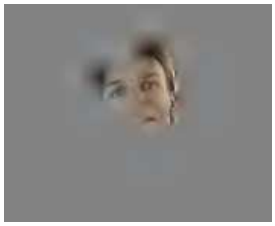
(a) carphone (QCIF)



(b) foreman (CIF)

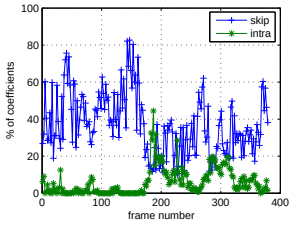
Figure 5: DWT-domain coefficient-variance for different block-MSE classes, at the various resolutions.

(a) *skip* coefficients only



(b) *intra* coefficients only

(c) all *inter* coefficients



(d) coefficient distribution

Figure 6: Reconstruction of the 25-th frame of the sequence carphone (QCIF) using subsets of coefficients with different labels (a,b,c); percentage of *skip* and *intra* coefficients used in each frame (d).

presented in the previous Section. Side information is obtained by block-based overlapped motion compensated interpolation of the previous and of the subsequent frames, using blocks of 16×16 pixels. Overlapping serves to avoid blocking artifacts, which would cause a performance degradation due to false high energy in the high frequency wavelet bands. The same CDF 9/7 wavelet kernel used in the standard JPEG2000 [11] has been then used for spatial transformation; in particular, four wavelet analysis (and synthesis) levels have been applied. Syndrome formation and decoding is based on 8-states $\mathbb{Z}/4\mathbb{Z}$ TCQ.

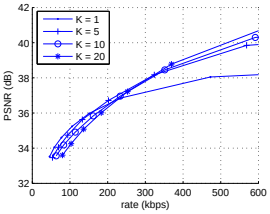
In the preliminary results, the mean transmission rate for a given quantization step is derived from the estimated entropy of the quantized coefficients, in bit/sample. This is in turn experimentally measured in each wavelet band assuming that each coefficient is coded into a context given by the magnitude of the previously coded coefficient, according to a certain coefficient scanning order. In particular, row-scanning is assumed in the LH_i bands, column-scanning is assumed in the HL_i bands, and zig-zag scanning is assumed in the LL_0 band (the smaller resolution) and in the remaining HH_i bands.

As first experiment, it is interesting to check if the classification into the various classes is reasonable according to the actual video data content. The positive answer comes from examination of Fig. 6. As can be noted, *skip* coefficients are correctly used to describe the more steady part of the background of the scene, *intra* coefficients are correctly used in areas with high motion, i.e. in the foreground, while the remaining coefficients are set as *inter* coefficients. Fig. 6(d) shows then that the coder is able to choose on the fly more *intra* coefficients where there is more motion in the sequence (i.e. around the 200-th frame).

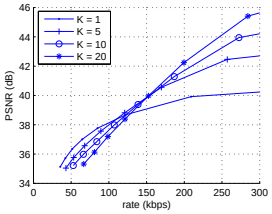
As stated before, varying the parameter K allows for adaptation of the coder to the non-Gaussian data statistics. As shown in Fig. 2, the coefficient distribution is usually more long-tailed than a Gaussian (and sometimes it is in fact assumed as Laplacian), and probably $K > 2$ should be then used to obtain the best performance. This is due to the fact that the continuous-syndrome decoding procedure produces an error each time the distance between Y and $X + (\tilde{S} - S)$ is actually greater than twice the minimum distance between the points of $\mathscr{C}_\lambda$ [5], and $K \cong 2$ balances the number of errors and the fidelity of reconstruction (in case of no errors) for best performance only for Gaussian data. In case of more long-tailed distributions, it should be probably used a greater K to get a similar optimum error rate, accepting a lower fidelity of reconstruction in case of no errors.

Preliminary experiments have been carried on with the goal of finding the more suitable value of K on the sequence carphone, compressed at QCIF resolution. As shown in Fig. 7(c), small values of K produce a great amount of errors, but, at low bit-rates, permit to obtain a better rate-distortion performance, in reconstructing both the *full-size* QCIF resolution (Fig. 7(a)) and the *reduced-size* sub-QCIF resolution (Fig. 7(b)). In particular, it turned out that the best value of K is a function of the desired target bit-rate.

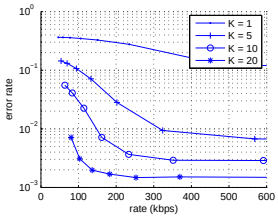
With this consideration in mind, the sequence foreman, at QCIF resolution, was compressed using the optimum values of K found for the sequence carphone. The average performance on the first 100 frames, compared to other results found in the literature, is shown in Fig. 8. In particular, two solutions from [2] (both based on turbo-codes) and another one from our previous paper [9] (similarly based on continuous-valued syndromes) are taken into consideration.



(a) quality of the reconstructed sequence (at QCIF resolution)



(b) quality of the reconstructed sequence (at sub-QCIF resolution)



(c) error rate (in reconstructing the QCIF resolution)

Figure 7: Results for the sequence carphone (compressed at QCIF resolution).

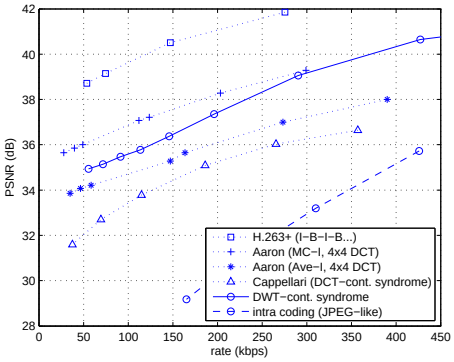


Figure 8: Performance of different video codecs, averaged on the first 100 frames of the sequence foreman (QCIF) at 25 frames per second. (Data relative to *intra* frames are not taken into account.)

All of them apply distributed source coding ideas into the DCT domain, and hence do not allow for spatial scalability.

It can be noted that the solution proposed in this paper, which moreover is quality-scalable since the corresponding curve is obtained with embedded quantization, outperforms both the solution proposed in [9] and one of the solutions (AVE-I) from [2], while being very close to the other solution (MC-I) proposed in [2]. However, it should be emphasized that both solutions proposed in [2] use a feedback channel between decoder and encoder, that helps in the estimation of the actual side-source correlation, but that it is very difficult to control in practical scenarios.

6. CONCLUSION

A scalable video coder based on distributed coefficient coding in the wavelet domain has been proposed. Deep analysis of coefficient correlation allows for the design of a quite effective video coding solution that does not need any feedback channel between the decoder and the encoder nor any motion estimation at the encoder. Correlation estimation does not represent a complex task since it is based on a simple block-based MSE estimation. Due to the intrinsic robustness of the distributed coding paradigm and to the quality-spatial scalability, the proposed coder is very suitable for applications in wireless and mobile environments, where errors may occur and terminals may have very different processing/visualization means. Preliminary results showed that the performance of the coder is close to the one of similar solutions found in the literature.

REFERENCES

[1] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[2] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Proc. Visual Communications and Image Processing (VCIP-2004)*, San Jose, CA, USA, Jan. 2004.

[3] R. Puri and K. Ramchandran, "PRISM: a "reversed" multimedia coding paradigm," in *Proc. of IEEE Intl. Conf. on Image Processing*, 14-17 September 2003, vol. 1, pp. 617–620.

[4] R. Bernardini, R. Rinaldo, P. Zontone, D. Alfonso, and A. Vitali, "Wavelet domain distributed coding for video," in *Proc. of IEEE Intl. Conf. on Image Processing*, 8-11 October 2006, pp. 245–248.

[5] L. Cappellari and G. A. Mian, "A practical algorithm for distributed source coding based on continuous-valued syndromes," in *Proc. of European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, 4-8 Sept. 2006.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[7] M. W. Marcellin and T. R. Fisher, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 82–93, Jan. 1990.

[8] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, NY, USA, 1988.

[9] L. Cappellari and G. A. Mian, "An algorithm for intra-frame video coding based on continuous-valued syndromes," in *Conf. Rec. of 40th IEEE Asilomar Conf. on Signals, Syst. and Comput.*, Pacific Grove, CA, U.S.A., 29 Oct.-1 Nov. 2006, pp. 1090–1094.

[10] C.-F. Chen and K. K. Pang, "The optimal transform of motion-compensated frame difference images in a hybrid coder," *IEEE Trans. Circuits Syst. II*, vol. 40, no. 6, pp. 393–397, June 1993.

[11] D. S. Taubman and M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standard and Practice*, Kluwer Academic Publishers, Boston/Dordrecht/-London, 2002.