Improvements of the ZeroTree-based Video Coding by using Scalable Three-dimension Spatio-temporal Orientation Tree Data Structure

ZHANG ZONG-PING[1,2], LIU GUI-ZHONG[3], LIU KUN[1], and PENG JI-HU[1,2]
[1]EDA Key Lab. Research Institute of Tsinghua University in Shenzhen, Shenzhen 518057 P.R.China
[2]Department of Electronic Engineering, Tsinghua University, Beijing 100084 P.R.China
[3]Department of Information & Communication Engineering, Xi'an Jiaotong University, Xi'an, 710049, P. R. China

Abstract: Based on the analysis to the information distribution and the parent-child correlation of three-dimensional (3D) wavelet coefficients, a scalable 3D spatio-temporal orientation tree (3D-SOT) is proposed to improve the performance of embedded zerotree wavelet video codec. Simulations on the 3D-SPIHT coding system show that the proposed scalable 3D-SOT can significantly improve the rate-distortion performance of embedded zerotree video codec, particularly at low bit rates.

Key-Words: Wavelet-based Coding, Video Compression, SPIHT, Wavelet Tree

1 Introduction

Since the pioneer work of Shapiro [1], embedded *zerotree* wavelet (EZW) coding has become a core technique for many wavelet-based image and video codecs. The set partitioning in hierarchical trees (SPIHT) coding algorithm proposed by Said *et al* [2] further promoted this technique. It not only improves the rate distortion performance compared with EZW, but also maintains a low computation complexity. The scalable nature of SPIHT makes it suitable for multi-rate and multimedia applications. It can also be easily extended to three-dimension (3D) applications due to its dimension-free nature. In [3], Kim *et al* performed such a work. The simulations in [3] demonstrated that despite a success at high bit rates compared with MPEG-2, the 3D zerotree wavelet video codec's performance would deteriorate as the bit rate decreases, especially for the inactive video applications at very low bit rates.

In principle, many exiting dimension-free wavelet coding techniques can be introduced to 3D zerotree video coder to improve its performance for very low bit rate case. However, there actually exit many special properties for video sequence, and these properties also can be utilized to improve the coder's performance. For example, there is a typical stationary character for inter-frame signals when video sequences having no scene cuts, differently from that of intra-frame signals. This difference makes the temporal wavelet transform to yield higher compaction efficiency than the spatial wavelet transform. As results, the information distribution and parent-child correlation of transformed residues will have much difference in different direction in 3D wavelet domain. As we know, the information distribution and parent-child correlation of transformed residues are just the two main factors effecting on the coding efficiency of embedded zerotree-based coder.

Based on this observation, we, in this paper, will propose a novel 3D spatio-temporal orientation tree. We refer to it as the scalable 3D spatio-temporal orientation tree (3D-SOT). By better capturing the foresaid properties, the proposed scalable 3D-SOT will be expected to significantly improve the coder's performance.

The rest of the paper is organized as follows. In following Section, we first present methods to perform the theoretical analysis on the information distribution and parent-child correlation in 3D wavelet-transformed domain, then based on experiments results on these properties with real 3D wavelet coefficients, we propose the scalable 3D-SOT. The performance evaluation is given in section 3 and the last section concludes the paper.

2 Three-Dimensional Spatio-Temporal Orientation Tree

2.1 3D Wavelet Coefficients' Properties

For a bitplane-based non-uniform quantization with deadzone 2Δ, given minimum bit-plane $b_{\min}$

(i.e. $\Delta = 2^{b_{\min}}$), the information $h(c_i)$ of a coefficient c_i can be calculated with following formula (2).

$$h(c_i) = -\log_2(p_{n_i}) + 1 - d(n_i) \quad (1)$$

where p_{n_i} is the probability of the $|c_i|$, $n_i = \max\{\lfloor \log_2 |c_i| \rfloor - b_{\min} + 1, 0\}$, $d(n)$ is the Dirac function. The right-side first term $-\log_2(p_{n_i})$ of formula (1) represents the coding cost of the coefficient c_i magnitude, and the second term $(1-d(n))$ representing the coding cost of its sign. According to the formula (1), the total information $H_{t,s}$ of spatio-temporal subband $S_{t,s}$, where subscript *t* and *s* represent the temporal-level and spatial-level respectively, can be calculated as follows

$$H_{t,s} = \sum_{c_i \in \mathbf{S}_{t,s}} h(c_i) \quad (2)$$

For a given spatio-temporal subband $S_{t,s}$, the parent-child correlation coefficient of the wavelet coefficients' magnitude is defined as follows

$$r_{t,s} \frac{\sum_{i \in S_{t,s}} \sum_{j \in O_i} |p_i \| c_{i,j}|}{\sqrt{\left(\sum_{i \in S_{t,s}} \sum_{j \in O_i} p_i^2\right)\left(\sum_{i \in S_{t,s}} \sum_{j \in O_i} c_{i,j}^2\right)}} \quad (3)$$

where p_i is the *i*-indexed element in $S_{t,s}$, $c_{i,j}$ is the *j*-indexed direct child of the p_i, O_i presents the index set of the direct child for the *i*-indexed element in $S_{t,s}$.

Based the formula (1) and (2), we perform the experiments on the information distribution with 3D wavelet-transformed coefficients of a QCIF (176 x144) gray sequence "Carphone" (285 frames), where a group of contiguous 16 frames is performed by a three-levels 1-D temporal decomposition with motion compensation, followed by three-levels 2-D spatial pyramid wavelet decomposition.

Fig.1 presents the distribution histogram of the averaged information for single coefficient across spatio-temporal subbands at the time of $b_{\min} = 4$. It shows that for a given quantization threshold, the averaged information distribution of single wavelet coefficient is not symmetrical across the spatio-temporal subbands. For example, the coefficients of $S_{0,i}$ have much more information than those of $S_{i,0}$, the counterpart of $S_{0,i}$.

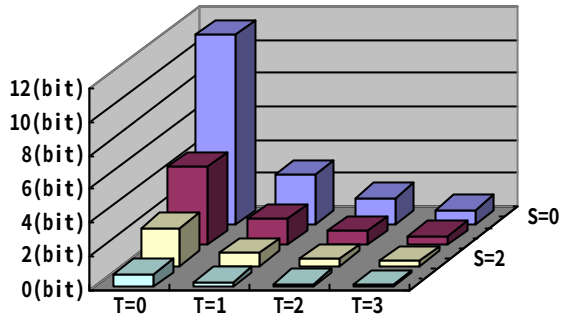


Fig.1 Distribution histogram of the averaged information for single coefficient across spatio-temporal subbands at $b_{\min} = 4$.

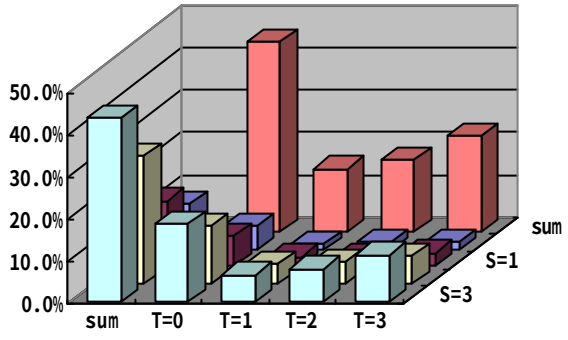


Fig.2 Distribution histogram of the total information for each spatio-temporal subband at $b_{\min} = 4$.

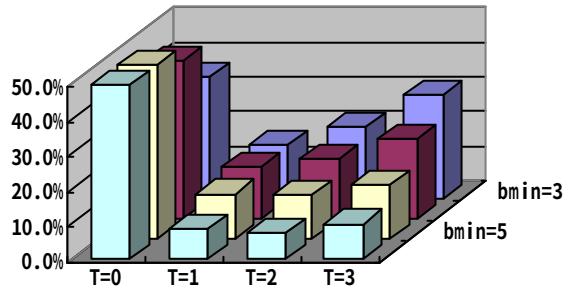


Fig3. Temporal boundary distribution distogram at different quantization threshold

The Fig.2 presents the distribution histogram of the total information of each spatio-temporal subband at the time of $b_{\min} = 4$. It shows that although the total information of each spatio-temporal subbband increases as meanwhile as

Table 1
Parent-Child Correlation of Adjacent Spatial-Scale with Same Temporal Location

s-Res. \ t-Res.	0	1	2
0	0.521	0.563	0.558
1	0.691	0.611	0.609
2	0.682	0.611	0.604

Table 2
Parent-Child Correlation of Adjacent Temporal -Scale with Same Spatial Location

s-Res. \ t-Res.	0	1	2
0	0.703	0.335	0.285
1	0.615	0.073	0.123
2	0.594	0.060	0.092

the spatial/temporal resolution increases due to increasing of number of coefficients, the temporal-approximation almost collects 50 percents information of the whole 3D wavelet coefficients. The Fig.3 shows that as the quantization threshold increasing, the contributing information gradually concentrates to the temporal-approximation subband.

The above 3D wavelet-transformed coefficients are also employed to test the parent-child correlation with formula (3). Table 2 and Table 1 present the statistical results respectively for the adjacent temporal-scale with same spatial location and the adjacent spatial-scale with same temporal location.

Table 1 demonstrates that there exits strong correlation between the parent-child coefficients of adjacent spatial-scale with same temporal location. Contrastively, the Table 2 shows the correlation for the parent-child coefficient of adjacent temporal-scale with same spatial location can be ignored except for the temporal approximation subband.

2.2 Scalable 3D-SOT

The character of information distribution of 3D wavelet-transform shows clearly that we should promote the coding priority of temporal-approximation subbands. And the weak parent-child correlations of the high frequency spatio-temporal subbands also show we can disregard their parent-child relationship when defining the 3D wavelet tree. Based on these observations, we define the following 3D wavelet tree.

Suppose that we have l_s levels 2-D pyramid decomposition in spatial domain and l_t levels 1-D Mallat decomposition in temporal domain on a GOP with dimensions of $M \times N \times F$, where *M, N, F* are horizontal, vertical, and temporal dimensions of coding unit. Thus, we have root video dimensions of $M_R \times N_R \times F_R$, where $M_R = M / 2^{l_s}$, $N_R = N / 2^{l_s}$ and $F_R = F / 2^{l_t}$. Then we define three sets as follows.

Definition 1 *A node represented by a pixel* (i, j, k) *is said to be a root node, middle node or a leaf node according to the following rule.*

If $i < M_R$ *and* $j < N_R$, *then* $(i, j, k) \in \boldsymbol{R}$

Else if $i \geq M/2$ *or* $j \geq N/2$ *then* $(i, j, k) \in \boldsymbol{L}$

Else $(i, j, k) \in \boldsymbol{M}$.

And, the sets $\boldsymbol{R}$, $\boldsymbol{M}$, *and* $\boldsymbol{L}$ *represent Root Middle, and Leaf respectively.*

Definition 2 *Let us denote* $\boldsymbol{O}(i, j, k)$ *as a set of offspring pixels of a parent pixel* (i, j, k). *Then a 3D wavelet tree is said to be a scalable 3D Spatio-temporal Orientation Tree according to the following rule.*

If $(i, j, k) \in \boldsymbol{L}$, *then* $\boldsymbol{O}(i, j, k) = f$;

Else $(i, j, k) \in \boldsymbol{M}$, *then*

$$\boldsymbol{O}(i, j, k) = \{(i+m, j+n, k+F_R) \mid m, n = 0,1\}$$

Else $(i, j, k) \in \boldsymbol{R}$ *then*

If $i\%2 = 0$ *and* $j\%2 = 0$ *then*

$$\boldsymbol{O}(i, j, k) = \begin{cases} \{(i+m, j+n, k+F_R) \mid m, n = 0,1\} & k < F_R \\ \{(2i+m, 2j+n, k) \mid m, n = 0,1\} & F_R \leq k < F/2 \\ f & \textbf{otherwise} \end{cases}$$

Else if $i\%2 = 1$ *and* $j\%2 = 0$ *then*

$$\boldsymbol{O}(i, j, k) = \{(i + M_R - m, j+n, k) \mid m, n = 0,1\}$$

Else if $i\%2 = 0$ *and* $j\%2 = 1$ *then*

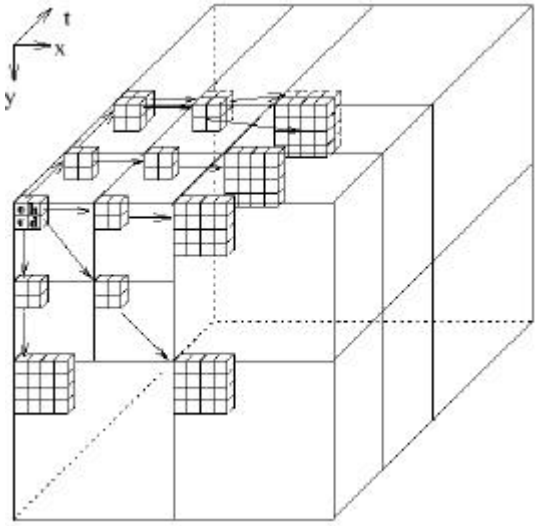
$$\boldsymbol{O}(i, j, k) = \{(i+m, j + N_R - n, k) \mid m, n = 0,1\}$$

Else if $i\%2 = 1$ *and* $j\%2 = 1$ *then*

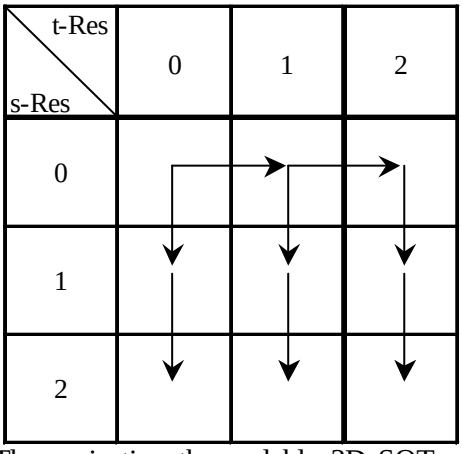
$$\boldsymbol{O}(i, j, k) = \{(i + M_R - m, j + N_R - n, k) \mid m, n = 0,1\}$$

where % is the mod operator.

Fig.4 depicts the parent-child relationships defined by the scalable SOT for the 21 subbands generated by a two-level temporal decomposition followed by a two-level pyramid spatial



(a) The parent- child relationships of scalable 3D-SOT



(b) The projection the scalable 3D-SOT onto 2-D spatio-temporal multiresolution plane

Fig. 4 the proposed scalable SOT (2-levels spatial decomposition plus 2-levels temporal decomposition, 21 subbands).

decomposition [4]. Each arrow originating from a root pixel and pointing to a group node (2×2, or $2 \times 2 \times 2$) shows the respective parent-child linkage.

It can be seen that the scalable SOT disconnect the parent-child relationships of spatial-temporal high-frequency subbands. It potentially sorts these subbands by their temporal frequency attribute. Furthermore, the scalable SOT allows a complete temporal decomposition due to its one-to-one temporal parent-child relationship mode of S_{00} and $S_{0,1}$.

3 Performance Evaluations

3.1 Improved 3D-SPIHT Video coding system

The SPIHT codes the wavelet-transformed data in two stages, i.e. the sorting and the refinement. For optimally sorting, it maintains three linked lists, known as the list of insignificant pixels (LIP), the list of significant pixels (LSP) and the list of insignificant sets (LIS). It first initializes the LIP with all pixels of the lowest subband, the LIS with all those pixels within the lowest subband and having descendents and LSP as an empty set.

In the sorting pass, every entry of LIP and LIS will be tested against a threshold. The corresponding significance will be recorded as 0 or 1. If a pixel in LIP is found to be significant, its sign is recorded accordingly, and it is then moved to the end of LSP. If a set in LIS is found to be significant, it is then partitioned into subsets (children and grandchildren). If a son-child is found to be significant, then it will be added to the end of the LSP, otherwise, it will be added to the end of the LIP. This split process will be continuously performed for each subset until it is insignificant. In the refinement pass, SPIHT outputs the refinement bits at current level of bit significance for all pixels in LSP. Through decreasing the current threshold by the factor of two, SPIHT repeats the above process until the desired bit rate is achieved.

The improved 3D-SPIHT has a similar coding structure to that of 3D-SPIHT. The difference that distinguishes it from SPIHT is that it organizes the bit stream in a layer-temporal-spatial-resolution successive mode. The improved 3D-SPIHT organizes all the nodes of the LIP, LSP and LIS with a uniform SOT. We refer it to a real SOT to distinguish from the virtual SOT in LIS of the original SPIHT since it keeps all the nodes instead of only the roots in LIS. To perform the sorting and the refinement, a state marker is used for each node to mark its position in the real SOT and, to indicate its pixel-coding state.

In addition to a pass used to record the updating process of the real SOT, The improved 3D-SPIHT employs two other coding passes to code these pixels in the real SOT for each spatial-temporal resolution. Each coefficient bit on the current bit plane is coded only by one of the two coding passes. The coding passes are called the significance propagation and the

magnitude refinement. The significance propagation pass includes only bits of coefficients that were insignificant (the significance bit has yet to be encountered). The magnitude refinement pass includes the bits from coefficients that are already significant (except those that have just become significant in the immediate significance propagation pass).

As results, the improved 3D-SPIHT can yield embedded multiresolution bit stream [5]. For the optimizing process restricted in a smaller range, the rate distortion performance of the improved 3D-SPIHT would be decreased a little compared with that of the original 3D-SPIHT.

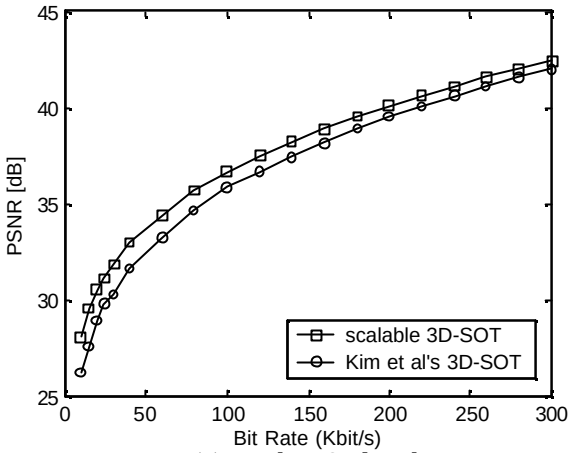
To improve the RD property, the intra-frame context-based adaptive arithmetic coding is used to perform the lossless entropy coding. The scheme of JPEG2000 is employed to reduce the context [6]. The arithmetic coder flushes the output bits at the end of each sub-block, and updates the context-model at the start of coding for each temporal band on each bit-plane layer.

3.2 Simulation results

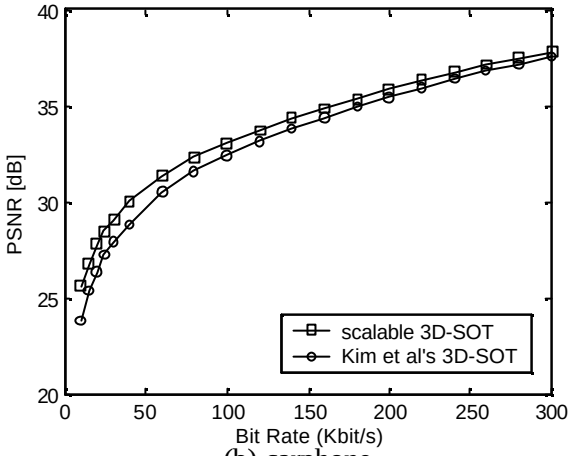
The two QCIF gray video sequences "Carphone" and "Mother & Daughter" are chosen to evaluate the performance of the proposed scalable 3D-SOT. The "carphone" sequence has a moving background; and the "mother & daughter" has relative quiet background. The input GOP of 16-frames is decomposed into 3D subbands by a 4-level or 3-level temporal transform with Haar wavelet for the proposed scalable 3D-SOT or the 3D-SOT of [3], respectively, followed by 3-levels Mallat spatial transform with Daubechies 9/7 filters. The resulting wavelet-packet coefficients are coded with the foresaid improved 3D-SPIHT coding algorithm.

We compared the proposed scalable 3D-SOT with the 3D-SOT defined by Kim et al. in [3]. Fig.5 plots the mean PSNR versus the bit rate on the 3D-SPIHT. It is apparent that the scalable SOT results in much improvement over Kim's SOT at very low bit rates, e.g. an advantage of ~2dB at 10kbit/s for both "Carphone" and "Mother & Daughter".

Due to the strong correlation between adjacent frames, the temporal wavelet transform compresses



(a) mother & daughter

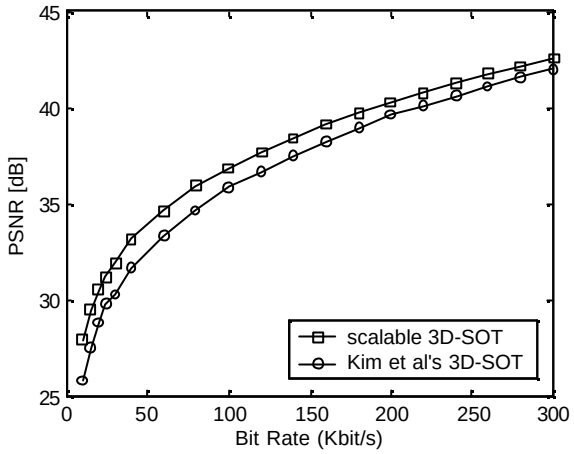


(b) carphone

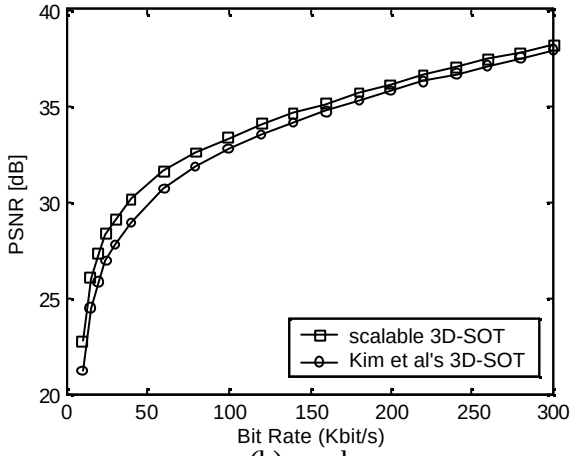
Fig. 5 The mean PSNR versus the bit rate on 3D-SPIHT

most of the energy to the temporal approximation subband. The typical one-to-one parent-child temporal linkage structure of the scalable 3D-SOT in temporal-approximation subband makes it possible to perform complete temporal-decomposition, thus the energy of video can be further compacted. At the same time, the scalable 3D-SOT can yield higher coding priority for $S_{t,s} = S_{m,n}, m < n$ subband than that of the counterpart $S'_{t,s} = S_{n,m}$ of $S_{m,n}$, thus decreasing the coding cost of pixels in 3D-SOT at the early encoding stage.

However, due to the decreasing of branches of wavelet tree, the scalable 3D-SOT simultaneously results in the increasing of coding cost of wavelet tree information. As the bit rate increases, the increased information originating from using the tree of the scalable SOT possibly counteracts these gains.



(a) mother & daughter



(b) carphone

Fig. 6 The mean PSNR versus the bit rate on MC 3D-SPIHT

When applied to motion-compensated three-dimensional wavelet video coding system, the proposed scalable 3D-SOT can be expected to improve the codec's performance due to the further compaction of motion- compensated (MC) temporal transform to the energy left in the temporal high-frequency subbands. The results curves on MC 3D-SPIHT in Fig.6 confirm this assumption.

4 Conclusion

Based on the observations on the information distribution and the parent-child correlation of magnitude for 3D wavelet coefficients, we proposed a novel 3D wavelet tree. Simulations show that the proposed scalable 3D-SOT can significantly improve the RD performance for zerotree-based 3D wavelet video coder, particular at very low bit rates.

References

[1] J. M. Shaprio, Embedded image coding using zerotrees of wavelet coefficients, *IEEE Trans. on Signal Processing*, Vol.41, No.12, 1993, pp.3445-3462

[2] A. Said and W. A. Pearlman, A new, fast, and efficient image codec based on set partitioning in hierarchical trees, *IEEE Trans. on Circuit and Sys. Video Tech.*, Vol.6, No.3, 1996, pp.243-250

[3] B.-J.Kim , Z. Xiong and W. A. Pearlman, Very low bit-rate embedded video coding with 3D set partitioning in hierarchical trees (3D SPIHT), *IEEE Trans. on Circuit and Sys. Video Tech.*, Vol.10, No.8, 2000, pp.1374-1387

[4] Z.-P. Zhang, G.-Z. Zhang and Y.-W. Yang, High performance scalable video compression with embedded multiresolution MC-3DSPIHT, *Proceedings-IEEE International Conference on Image Processing (ICIP),* Vol.3, p.III/721-III/724, *2002*

[5] Z.-P. Zhang, Three-dimensional wavelet transform-based scalable video compression and error-resilient coding, Ph.D thesis of Xi'an Jiaotong University, 2002

[6] M. Bolick (editor), JPEG 2000 Part I Final Committee Draft Version 1 .0, *ISO/IEC JTC 1/SC 29/WG1 N1646R.* 2000.