Generalized Lifting Prediction Optimization Applied to Lossless Image Compression

Joel Solé, *Member, IEEE*, and Philippe Salembier, *Member, IEEE*

***Abstract*—A useful tool to construct wavelet decompositions is the lifting scheme. The generalized lifting is an extension of the classical lifting scheme to introduce more flexibility and to permit the creation of new nonlinear and adaptive transforms. However, the design of generalized prediction and update steps is more involved. This letter proposes a generalized prediction design that minimizes the detail signal energy and entropy at the same time. Two algorithm variants are given. The fixed prediction uses the image class statistics to derive the optimal transform. If the statistics are unknown, the adaptive prediction extracts them from the image being coded. The resulting decompositions are applied to lossless image coding, reporting good results. The adaptive algorithm has no bookkeeping or side information requirements, yet its performance is close to the fixed prediction performance.**

***Index Terms*—Image compression, lifting scheme, nonlinear filtering.**

I. Introduction

THE LIFTING scheme [1] gives a suitable framework for developing space-varying wavelet filter banks. Its initial polyphase decomposition [or lazy wavelet transform (LWT)] allows flexible signal analysis of one channel in order to apply a good lifting filter to the other channel. The transform is reversible because of the lifting scheme structure itself. Many approaches [2]–[4] follow the idea, trying in different ways to exploit the correlation existing between both decomposition channels: the local shape or statistics of one signal is considered in order to apply a better filter to the other one.

Going one step further, the adaptation may be improved if the information given by the same channel to be filtered is taken into account. A point-wise adaptation is possible by using a criterion invariant to the filtering so that it can be recovered at the decoder, thus allowing the correct choice of the synthesis filter. The generalized lifting scheme follows this line of research. In this scheme [5], the sums in the lifting are generalized to include possibly nonlinear and adaptive operations. Furthermore, the proposal guarantees the transform reversibility by a simple injectivity criterion of a mapping.

This letter presents a prediction step design within the generalized lifting scheme. The prediction is optimal in the sense that it minimizes at the same time the detail signal energy and entropy given the pixel value probability conditioned to its neighbors pixel values. These statistics are extracted in two ways: from the image class database (if it is available) or from the image being coded. In both cases, good image coding results are reported in the Jpeg2000 [6] environment and in a 3-D version of the SPIHT [7] entropy coder.

This letter is organized as follows. Section II introduces the lifting scheme and Section III the generalized lifting. Section IV describes an optimal prediction given the probability of a sample conditioned on its neighbors values, and Section V discusses an adaptive algorithm if the probability is not previously known. Section VI gives some experimental results, and main conclusions are drawn in Section VII.

II. Lifting Scheme

The lifting scheme (LS) is a well-known method to create biorthogonal wavelet filters from other ones [see Fig. 1(a)]. The scheme input data are $\mathbf{x}_0$, which is divided into two subsignals: an approximation signal $\mathbf{x}$ formed by the even samples of $\mathbf{x}_0$ and a detail signal $\mathbf{y}$ formed by the odd samples of $\mathbf{x}_0$. Then, the lifting steps are applied as follows.

- Prediction lifting step P that predicts the detail signal samples using the approximation samples $\mathbf{x}$: $y'[n] = y[n] - P(\mathbf{x}[n])$. Vector $\mathbf{x}[n]$ denotes a subset of samples in $\mathbf{x}$ around location n.
- Update lifting step U that updates the approximation signal with the detail samples $\mathbf{y}'$: $x'[n] = x[n] + U(\mathbf{y}'[n])$.

The transform coefficients $\mathbf{x}'$ and $\mathbf{y}'$ are the output data. A multiresolution decomposition of $\mathbf{x}_0$ is built by the concatenation of the lifting decomposition blocks on the approximate subsignal, like the recursive filter bank tree-structure does.

The prediction and update operators may be a linear combination of $\mathbf{x}$ and $\mathbf{y}$, respectively, or any nonlinear operation (since by construction, the LS is always reversible). The inversion of the scheme is straightforward. The same prediction lifting step (PLS) and update lifting step (ULS) are employed, and only the sign of the addition is changed. Finally, subsignals are merged into the higher rate signal to recover the original data $\mathbf{x}_0$.

The even-sampled channel is used to extract redundancy from the odd-sampled channel by means of the prediction step. The differences, which are the detail or wavelet coefficients, are left in this odd channel. Details tend to be small, which amounts to compression efficiency. Then, wavelet coefficients are used to update the even channel in order to obtain a coarse scale version of the input signal. The ULS can be seen as an anti-aliasing filter.

III. Generalized Lifting Scheme

This section introduces a generalization of the lifting scheme, which is similar to the classical lifting, except that the sums after the filters are embedded in a more general setting [see Fig. 1(b)].

Manuscript received January 8, 2007; revised March 11, 2007. This work was supported in part by the TEC2004-01914 Project of the Spanish research program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vesa Valimaki.

The authors are with the Department of Signal Theory and Communications, Technical University of Catalonia (UPC), 08034 Barcelona, Spain (e-mail: joel@gps.tsc.upc.edu; philippe@gps.tsc.upc.edu).

Digital Object Identifier 10.1109/LSP.2007.898348

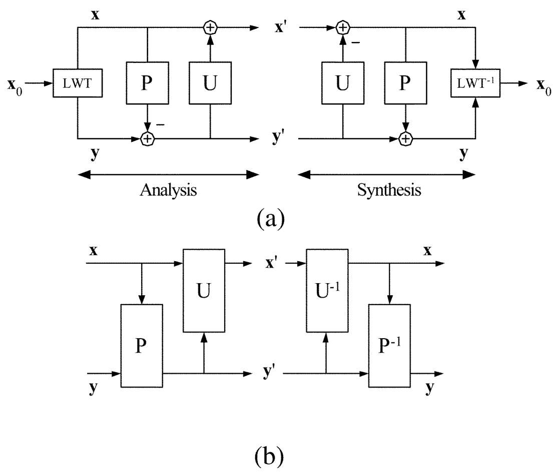

Fig. 1. (a) Lifting scheme and (b) the generalized lifting scheme.

For instance, the classical lifting prediction is a filter that generates a predicted value that is used to modify $y[n]$ through subtraction. In the *generalized lifting* (GL) scheme, the prediction step is viewed as a function that maps $y[n]$ to $y'[n]$, taking into account values from the approximation signal $\mathbf{x}$. The restriction of modifying $y[n]$ only through a sum has been removed, and so, the scheme allows more complex, possibly adaptive or nonlinear modifications. The same generalization can be done for the ULS, so the generalized lifting steps (GLSs) may be noted as $y'[n] = P(y[n], \mathbf{x}[n])$ and $x'[n] = U(x[n], \mathbf{y}'[n])$.

The following is a formal definition of a GLS. Let $\mathcal{A}$ be the set of functions a from $\mathbb{R}\times\mathbb{R}^k$ to itself, i.e., $a \in \mathcal{A} \Leftrightarrow a : \mathbb{R}\times\mathbb{R}^k \rightarrow \mathbb{R} \times \mathbb{R}^k$. Let $\mathcal{A}_0$ be the subset of $\mathcal{A}$ containing all functions that do only modify the first component, that is, for which the restriction to $\mathbb{R}^k$ is the identity: $\mathcal{A}_0 = \{a \subseteq \mathcal{A} | a_{|\mathbb{R}^k\rightarrow\mathbb{R}^k} = \mathbf{I}_k\}$. A GLS is defined as a function belonging to the subset $\mathcal{A}_0$. At the same time, if a reversible scheme is desired, the GLS should be an injective function of $\mathcal{A}_0$. The same statements apply to the generalized prediction and update steps.

The GL scheme as presented so far assumes that the values taken by $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}'$, and $\mathbf{y}'$ are real numbers. Quantization steps should be avoided in lossless compression. Then, the values taken by the GLS inputs and outputs are integers. We call this integer version the discrete GL. A bijective condition arises in a natural way in the discrete GL because the input and output discrete spaces have the same size, so mappings have to be one-to-one in order to have all the elements in each space related. Consider now the following framework for discrete grayscale images where each pixel is represented by 8 bits. Without loss of generality, sample values are assumed to range from -128 to 127. Let $\mathbb{Z}_{255}$ be the set of integers that belong to the interval [-128,127]. The discrete generalized update and prediction are mappings from the $\mathbb{Z}_{255} \times \mathbb{Z}_{255}^k$ space to itself that can only modify the first component.

IV. Optimization of the Generalized Prediction

Once the framework has been established, the problem is to design useful GLS for specific applications. This section focuses on the design of a prediction step. To this end, it is useful to define the concept of a column. For $\mathbf{x}[n] = (x[n-n_1] \ldots x[n-n_k])^T$ fixed, the set of all possible values of $y[n]$ describes a column in the $\mathbb{Z}_{255} \times \mathbb{Z}_{255}^k$ space. Let such a column be denoted by $C_{\mathbf{i}\in\mathbb{Z}_{255}^k}$ in the following:

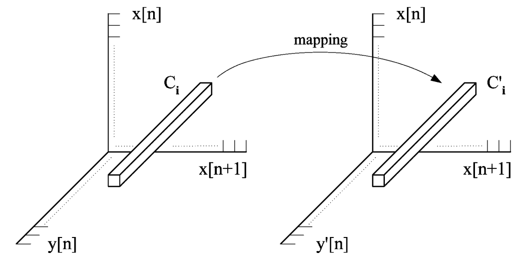

Fig. 2. Discrete mapping from the $\mathbb{Z}_{255} \times \mathbb{Z}_{255}^2$ space to itself. The lifting step is reversible if the mapping from each column $C_{\mathbf{i}\in\mathbb{Z}_{255}^2}$ to itself is bijective.

$$C_{\mathbf{i}\in\mathbb{Z}_{255}^k} = \{y[n], x[n-n_1] = i_1, \ldots, x[n-n_k] = i_k\}. \quad (1)$$

The generalized prediction only modifies the component $y[n]$, so it maps each $C_{\mathbf{i}\in\mathbb{Z}_{255}^k}$ to itself. Fig. 2 illustrates this for the case $k=2$. The mapping of $\mathbb{Z}_{255} \times \mathbb{Z}_{255}^k$ to itself should be bijective for all columns in order to have a reversible scheme.

The PLS design is formulated as an optimization problem that depends on the signal probability density function (pdf). As stated, the transform is reversible if every column mapping is bijective. The column prediction mappings are independent because columns partition the $\mathbb{Z}_{255} \times \mathbb{Z}_{255}^k$ space. Accordingly, every column mapping $P_{\mathbf{i}}(\cdot)$ is independently designed from each other

$$y'[n] = P\left(y[n], \mathbf{x}[n]\right) = \bigcup_{\forall \mathbf{i}\in\mathbb{Z}_{255}^k} P_{\mathbf{i}}\left(y[n]\right)|_{\mathbf{x}[n]=\mathbf{i}}. \quad (2)$$

Given $\mathbf{i} \in \mathbb{Z}_{255}^k$, the transform relates every input value $y[n] \in \mathbb{Z}_{255}$ one-to-one to every output value $y'[n] \in \mathbb{Z}_{255}$. Therefore, output values for each column $\mathbf{i}$ are related to input values simply through a permutation matrix $\mathbf{P_i}$, so a prediction step P may be seen as the union of permutation matrices.

Entropy coders benefit from several characteristics of wavelet coefficients. Specifically, they tend to increase their performance when the detail signal coefficients energy is minimized. For compression purposes, it is also interesting to reduce the output entropy. A reasonable goal is to design a mapping that minimizes the expected energy of the detail signal. Such an optimal prediction is

$$P_{opt} = \arg\min_P \mathbb{E}[y'^2] = \bigcup_{\forall \mathbf{i}\in\mathbb{Z}_{255}^k} \arg\min_{P_{\mathbf{i}}} \mathbb{E}[y'^2|\mathbf{x}=\mathbf{i}]. \quad (3)$$

The second equality in (3) is due to the independence between columns. The design of the prediction function reduces to finding the optimal column mapping $P_{\mathbf{i}}(\cdot)$ (or permutation matrix $\mathbf{P_i}$) for all columns. As explained in [8], the output expected energy for a column may be expressed as

$$\mathbb{E}[y'^2|\mathbf{x}=\mathbf{i}] = \left((-128)^2 \ldots (127)^2\right) \times \mathbf{P_i}\left(Pr(y=-128|\mathbf{x}=\mathbf{i}) \ldots Pr(y=127|\mathbf{x}=\mathbf{i})\right)^T. \quad (4)$$

Note that $Pr(.)$ stands for the probability function. The energy expectation in (4) is minimized when the permutation matrix relates input values of high probability with small energy output values. Proposition 1 in the Appendix confirms this statement. This mapping also minimizes the output entropy, as proposition 2 shows.

The permutation matrix optimizing (4) that relates input conditional probabilities with output energies is used in the discrete sample space to relate each input with the corresponding output. Assuming that the pdf is known, a column map is created by constructing a vector with input values sorted by their probability in descending order. The first element of this vector, which is the most probable input sample for the given context, is assigned (mapped) to a 0 output value (the minimum energy output). Following, the output value of -1 is assigned to the vector second element (corresponding to the input values of second highest probability), 1 is assigned to the third element, 2 to the fourth, and so on. In practice, a PLS is performed by column mappings, which are look-up-tables (LUT) that reorder input values according to their probabilities. These LUT are a practical representation of the permutation matrices and permit a fast transform but paying the cost of an LUT storage per image class at the coder and decoder.

V. Adaptive Prediction Design

This section describes a modification of the optimized prediction that avoids the necessity of the previous knowledge of the pdf and, thus, also avoids the storage of an LUT for every image class. Indeed, in this approach, the LUT may be different from level to level and for each filtering direction, which may result in compression gains w.r.t. one fixed LUT per image class. The drawback is the computation cost of an adaptive pdf estimation, which is not required in the previous "fixed" prediction.

The pdf estimation should be updated at each sample n in a way that permits the coder and the decoder reaching the same results, i.e., a synchronized iterative estimation. Therefore, the prediction is adapted to image statistics. Indeed, the pdf may be independently estimated at each resolution level reaching finer optimization than using a fixed LUT for the whole decomposition.

Nonparametric density estimation methods are suited for this application because they model data without making any assumption about the form of the distribution. Kernel-based methods (which is a subclass of the nonparametric methods) construct the estimation by locating weighted kernel-functions at the index position of the samples. Experiments using different kernel shapes and bandwidths have been carried out, leading to similar results for a wide range of values.

The delta function has been chosen as the kernel. It is the simplest kernel and amounts to the computation of the histogram. The delta kernel is the choice because its results are not worse w.r.t. other kernels, and it has two interesting properties for our purpose. First, the histogram pdf estimation converges to the optimal pdf that minimizes the detail signal energy for the image at the given resolution level and filtering direction. Second, in practice, the choice of delta avoids an explicit pdf estimation that other choices would not allow: since at each sample, only one histogram bin is modified, it is only necessary to reorder that bin in the vector that relates input probabilities with output values. In consequence, the time-consuming pdf re-estimation and the sorting pass of probabilities for constructing the input-output vectors are avoided.

TABLE I
Bit-Rates in Bits per Pixel for Four-Level Decompositions. Mean Values for the Mammography and SST Sets and for Two Synthetic Images Using Jpeg2000. Below, Results for the MRI Set Using a 3-D SPIHT

bpp in Jpeg2000	**5/3 wavelet**	**Fixed Pred.**	**Adap. Pred.**
Mammography	2.444	2.302	2.333
SST	2.874	2.326	2.356
Cmpnd1	2.082	—	1.352
Chart	3.088	—	3.038
bpp for SPIHT 3D	**5/3 wavelet**	**Fixed Pred.**	**Adap. Pred.**
MRI Head	3.667	3.635	3.508

An initial pdf estimation is required when no data are available. Different initial pdf may be considered. An interesting approach is to use the LUT of the image class at hand and then adaptively refine the pdf on the fly for the specific image being coded. For the experiment section, the chosen initial pdf is the one corresponding to the average pdf of the images in our database. This conditioned pdf turns out to be simple and structured. It has a maximum at the mean value of the neighbors and decreases monotonically and symmetrically on both sides. At a given sample, the pdf estimation is done by adding the initial pdf (i.e., the "average" pdf) with the histogram of all samples seen until the current one. The initial pdf sums up to one, while each processed sample adds one to the histogram, thus giving much less importance to the prior and speeding up the adaptation. The estimated pdf is then used to optimize the prediction for the current sample.

VI. Experiments

The fixed and adaptive optimized predictions are derived for $k = 2$ (i.e., two neighbors are considered for the prediction) and compared to the 5/3 wavelet via lifting scheme. This wavelet is employed in the Jpeg2000 standard for lossy-to-lossless compression purposes, and it also considers two neighbors to compute the PLS. No ULS follows the proposed predictions. The image is first filtered vertically and the approximation signal is filtered horizontally, resulting in a three-band decomposition. In the adaptive case, the pdf is estimated at each resolution level vertically and horizontally.

The optimal prediction mapping resulting from the "average" pdf is very similar to the 5/3 wavelet prediction mapping [9]. Therefore, there is more potential compression gain with respect to the 5/3 wavelet for those images belonging to a class with a pdf that significantly differs from the average pdf. The experiments consider such images. The test data include: a set of biomedical images (mammography), a set of remote sensing images (sea surface temperature), and synthetic images. Table I shows the bit-rate results for four-level decompositions. The proposed methods perform better than the 5/3 in all cases, with a small performance gap between the fixed and adaptive prediction. For synthetic images (which cannot be treated as an image class with a common pdf), the adaptive prediction gives compression rates up to 80% better than the 5/3. The two examples given in the table come from the official Jpeg2000 test set.

An MRI set of images of a human head is decomposed through the three dimensions and encoded with a 3-D version of the SPIHT. The fixed prediction employs the "average" pdf (note that this is not the MRI set pdf) and beats the 5/3 by a little margin. Results are improved by the adaptive prediction.

Both algorithms have distinct computational cost and memory requirement characteristics. The fixed prediction LUT implementation is computationally very efficient. However, its memory requirement is high: for the previous experiments with $k = 2$ and 256 gray-level images, an 8 Mbyte LUT per class is needed. The adaptive prediction keeps track of one LUT on run-time. This LUT contains the initial pdf prediction, and it is modified with each incoming sample by carrying out the following steps: 1) search the gray-level sample value within the column given by the context, 2) output the position of the value in the column, 3) update the gray-level value count by 1, and 4) sort the column updating the value position according to the new count. Therefore, the computational cost is much higher compared to the fixed prediction.

VII. Conclusions

The GLS defines the lifting step as a mapping between spaces. Its discrete version reduces the whole step design to a column-to-column mapping. The image pdf is employed to optimize the discrete generalized prediction. Propositions 1 and 2 show that the optimized prediction attains the minimal detail signal energy as well as the minimal entropy at the same time.

The optimized prediction can be applied to all the images in the class if the pdf is previously extracted from a training set of images. This fixed prediction is derived for the mammography and SST sets with considerable coding gain w.r.t. the 5/3 wavelet in the Jpeg2000 environment. The fixed prediction's main drawback is the LUT storage at the coder and decoder side. This fact suggests the creation of an adaptive optimized generalized prediction. The adaptation algorithm is simple and offers good convergence properties, which amounts to compression results only slightly worse than those of the fixed version. Additionally, the adaptive prediction is applicable to images that do not belong to any class with a common pdf. For example, the adaptive generalized scheme clearly outperforms the 5/3 wavelet for the synthetic images or the MRI volume set.

To sum up, from the initial formulation, two image coding approaches are derived that are useful and applicable in different situations: when the image statistics are known (such as in the biomedical or remote sensing image applications), an optimal fixed scheme can be used, and when they are unknown (as in the synthetic images case), an adaptive scheme has to be used.

Appendix
Minimum Energy/Entropy Mappings

Proposition 1: Let $\mathbf{s}, \mathbf{r} \in \mathbb{R}^n$, where the elements of $\mathbf{r} = (r_1 r_2 \ldots r_n)^T$ are sorted $r_1 < r_2 < \ldots < r_n$, and the elements of $\mathbf{s}$ are all different. Let $\mathbf{P}$ be an $n \times n$ permutation matrix and $\mathbf{P}_o$ be the $n \times n$ permutation matrix such that $\mathbf{s}_o^T = \mathbf{s}^T\mathbf{P}_o = (s_{o1}\ s_{o2}\ \ldots\ s_{on})$ and $s_{o1} > s_{o2} > \ldots > s_{on}$. Then, $\mathbf{P}_o$ is optimal in the sense that $\forall \mathbf{P} \neq \mathbf{P_o}$, $f^\star = \mathbf{s}^T\mathbf{P}_o\mathbf{r} < \mathbf{s}^T\mathbf{P}\mathbf{r}$. That is

$$\mathbf{P}_o = \arg\min_{\mathbf{P}} \mathbf{s}^T\mathbf{P}\mathbf{r}.$$

Remarks: The demonstration shows that the objective value $f = \mathbf{s}^T\mathbf{P}\mathbf{r}$ is not a minimum for any $\mathbf{P} \neq \mathbf{P}_o$. The proof is included in [9].

Definition: Let $\mathbf{v} \in \mathbb{R}_+^n$. Function $\hbar(.)$ is defined as $\hbar(\mathbf{v}) = -v_1\log(v_1) - v_2\log(v_2)\ldots - v_n\log(v_n)$.

Proposition 2: Let $\mathbf{s}, \mathbf{r} \in \mathbb{R}_+^n$, where the elements of $\mathbf{r} = (r_1\ r_2\ \ldots\ r_n)^T$ are sorted $r_1 > r_2 > \ldots > r_n$, and the elements of $\mathbf{s}$ are all different. Let $\mathbf{P}$ be an $n \times n$ permutation matrix and $\mathbf{P}_o$ be the $n \times n$ permutation matrix such that $\mathbf{s}_o = \mathbf{P}_o\mathbf{s} = (s_{o1}\ s_{o2}\ \ldots\ s_{on})^T$ and $s_{o1} > s_{o2} > \ldots > s_{on}$. Then, $\mathbf{P}_o$ is optimal in the sense that $\forall \mathbf{P} \neq \mathbf{P}_o$, $f^\star = \hbar(\mathbf{r} + \mathbf{P}_o\mathbf{s}) < \hbar(\mathbf{r} + \mathbf{P}\mathbf{s})$. That is

$$\mathbf{P}_o = \arg\min_{\mathbf{P}} \hbar(\mathbf{r} + \mathbf{P}\mathbf{s}).$$

Remarks: The proof is also included in [9]. These results can be extended to any number of couples of equal elements and any number of equal elements. By symmetry, results still hold if $\mathbf{s}$ is the sorted vector and $\mathbf{r}$ the vector to be permuted. Finally, the same optimal rearrangement of vectors elements by a permutation matrix can be done if neither vector is sorted. Proofs are straightforward.

If there is more than one vector $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$, then $f = \hbar(\mathbf{r} + \mathbf{P}_1\mathbf{s}_1 + \ldots + \mathbf{P}_m\mathbf{s}_m)$ is minimized when permutation matrices $\mathbf{P}_i$ with $1 \leq i \leq m$ align the values of each $\mathbf{s}_i$ in the same fashion as in the two-vector case (proposition 2). If $\sum_j(r_j + \sum_i s_{i,j}) = 1$, then the resulting column vector $\mathbf{r} + \mathbf{P}_1\mathbf{s}_1 + \ldots + \mathbf{P}_m\mathbf{s}_m$ is a probability distribution and $f = \hbar(\mathbf{r} + \mathbf{P}_1\mathbf{s}_1 + \ldots + \mathbf{P}_m\mathbf{s}_m)$ its entropy.

Acknowledgment

The authors would like to thank J. Konrad for his helpful comments.

References

[1] W. Sweldens, "The lifting scheme: A custom-design construction of biorthogonal wavelets," *Appl. Comput. Harmon. Anal.*, vol. 3, no. 2, pp. 186–200, 1996.
[2] R. L. Claypoole, G. Davis, W. Sweldens, and R. G. Baraniuk, "Nonlinear wavelet transforms for image coding," in *Proc. 31st Asilomar Conf. Signals, Systems, Computers*, Nov. 1997, vol. 1, pp. 662–667.
[3] N. V. Boulgouris, D. Tzovaras, and M. G. Strintzis, "Lossless image compression based on optimal prediction, adaptive lifting, and conditional arithmetic coding," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 1–14, Jan. 2001.
[4] A. T. Deever and S. S. Hemami, "Lossless image compression with projection-based and adaptive reversible integer wavelet transforms," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 489–499, May 2003.
[5] J. Solé and P. Salembier, "Adaptive discrete generalized lifting for lossless compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 2004, vol. 3, pp. 57–60.
[6] ISO/IEC, ISO/IEC 15444-1: JPEG 2000 Image Coding System, 2000.
[7] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.
[8] J. Solé and P. Salembier, "Prediction design for discrete generalized lifting," in *Proc. Advanced Concepts for Intelligent Vision Systems*, Sep. 2004, pp. 319–324.
[9] J. Solé, "Optimization and generalization of lifting Schemes: application to lossless image compression" Ph.D. dissertation, Univ. Politècnica de Catalunya, Barcelona, Spain, Jun. 2006. [Online]. Available: http://gps-tsc.upc.es/imatge/_Joel/.