AN 8X8 IEEE-COMPLIANT LIFTING-BASED MULTIPLIERLESS IDCT STRUCTURE AND ALGORITHM

Lijie Liu and Trac D. Tran

Department of Electrical and Computer Engineering
The Johns Hopkins University, Baltimore, MD 21218
Email: {ljliu,trac}@jhu.edu

ABSTRACT

In this paper we propose a lifting-based 8x8 IDCT structure and its IEEE-1180 compliant approximation solution. Derived from an efficient Loeffler's 11-multiply IDCT structure, the proposed scheme comprises of butterflies and dyadic-rational lifting steps that can be implemented using only shift and add operations. Our approach also allows the computational scalability with different accuracy-versus-complexity trade-offs. Furthermore, the lifting construction allows a simple construction of the corresponding multiplierless forward DCT, providing bit-exact reconstruction if pairing with our proposed IDCT. Our high-accuracy solution provides a very close approximation of the floating-point IDCT. The experiments in MPEG-2 and MPEG-4 video coders under the worst-case assumptions show almost drifting-free reconstructions.

Index Terms— IEEE-1180, lifting, multiplierless, IDCT

1. INTRODUCTION

The discrete cosine transform (DCT) has found wide applications in image/video coding and processing, and has been the main transform employed in current compression standards such as JPEG, H26x, and MPEG family [1]. In order to reduce the computation complexity of DCT/IDCT, numerous fast algorithms have been proposed for image and video applications. Many of these algorithms are based on various sparse factorizations of the DCT matrix, e.g., Chen's [2] and Loeffler's algorithm [3], which still require multiplications with irrational parameters. In practice, various fixed-point approximations are very common, leading to algorithms sacrificing accuracy for lower computational complexity. In most of MPEG standards, different video decoders employ different IDCT fixed-point approximations, resulting in the mismatched reconstruction of the decoded frame relative to the encoder's model of the intended decoded picture. Subsequent inter-picture prediction from the aforementioned mismatched P frame makes error accumulation and generates the so-called *drifting* effects which can degrade the decoded video quality quickly, especially when the encoder has a minimal amount of intra-block refresh. As signal processing technology has advanced, the need for the degree of freedom in IDCT transform implementations in decoder has subsequently diminished. As a result, MPEG considers to develop a new voluntary standard specifying a particular fixed-point approximation to the ideal IDCT function, and recently has a call for proposals on fixed-point 8x8 IDCT and DCT Standards [4]. The call requires an approximation that is within a specified degree of precision relative to the ideal function definition of the IDCT. And the accuracy requirements are compliant to IEEE-1180 standard [5], which has specified a set of evaluation criteria for IDCT implementations.

One approach in designing integer transforms and realizing DCT /IDCT approximations is via lifting-based structures [6, 7]. Different with other fixed-point DCT/IDCT implementations, the lifting structure enables invertible, integer-to-integer mapping as well as in-place computation. The irrational lifting parameters can be approximated by dyadic rationals, leading to fast algorithms that can be implemented using only binary additions and shifting operations. In [6], the binDCT family derived from Chen's and Loeffler's DCT structure, is shown to have a 16-bit multiplierless implementation while enabling lossless compression via invertible integer-to-integer mapping and capable of achieving virtually similar lossy compression efficiency as the original DCT. However, none of the binDCT configurations can pass IEEE-1180 measurements because the main design focus of the binDCT is on low complexity for portable computing. Zelinski *et al* uses the adder numbers of the dyadic rationales to represent the arithmetic complexity, and proposes an automatic approach for minimizing such complexity under the constraint of a particular quality measure such as coding gain of the forward DCT [7]. Their major concern is in the forward DCT domain, and the approximation accuracy of IDCT has not been addressed.

In this paper we propose a lifting-based 8x8 IDCT approximation scheme, which is an extension of our binDCT family [6]. Our goal is to not only meet all of the compliance requirements in IEEE-1180, but also find out the accuracy we can achieve within the 32-bit precision constraint. The paper is organized as follows. Section II presents our lifting solution, and gives the dynamic range analysis which which turns out to be a critical implementation issue in fixed-point architectures. Results of IEEE-1180 compliant tests and drifting tests in MPEG-2 and MPEG-4 codec are presented in Section III. Finally, conclusions are given in Section IV.

2. LIFTING-BASED IDCT STRUCTURE

2.1. Plane-Rotation-Based IDCT Factorization

An elegant factorization for eight-point IDCT was proposed by Loeffler *et al* [3]. The resulting structure is depicted on the left of Fig. 1. This structure contains a 4-point IDCT and it requires 11 multiplications (achieving the multiplication lower bound as proven in [8]) and 29 additions. This factorization is non-scaled and requires a uniform scaling factor of $1/\sqrt{8}$ at the end of flow graph to complete true 1-D IDCT transform. Hence, it does not require any modification of the quantization/inverse quantization stages if it is used in image/video applications. In the 2D separable implementation, the

This research is supported by the National Science Foundation under CAREER Grant CCR-0093262 and FastVDO, LLC.

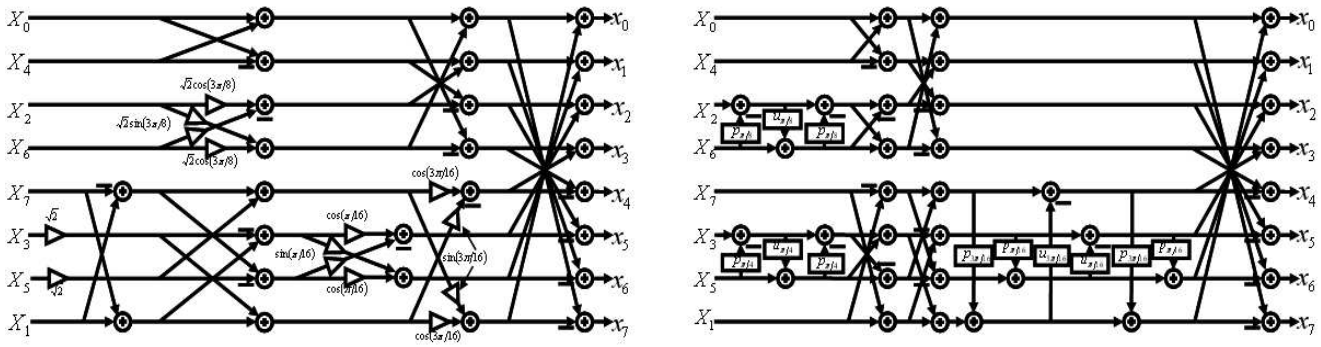

Fig. 1. Signal flow graph of eight-point IDCT. Left: Loeffler's factorization. Right: Proposed lifting-based IDCT approximation structure.

scaling factor becomes $1/8$ which is a simple 3-bit right-shift operation. Another advantage of the structure is that the two major plane rotations $\frac{\pi}{16}$ and $\frac{3\pi}{16}$ are close to the final output butterfly, which can delay approximation errors at the beginning of the flow graph, leading to high-accuracy IDCT approximation. As it is difficult to directly approximate $\sqrt{2}$ with high accuracy in Loeffler's structure, we factorize $\sqrt{2}$ factors according to (1) and (2). Therefore, the derived plane rotation-based IDCT factorization structure contains series of butterflies and four plane rotations, i.e., $\frac{\pi}{8}$, $\frac{\pi}{4}$, $\frac{\pi}{16}$ and $\frac{3\pi}{16}$.

$$\sqrt{2}\begin{bmatrix} cos(\frac{3\pi}{8}) & -sin(\frac{3\pi}{8}) \\ sin(\frac{3\pi}{8}) & cos(\frac{3\pi}{8}) \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} cos(\frac{\pi}{8}) & -sin(\frac{\pi}{8}) \\ sin(\frac{\pi}{8}) & cos(\frac{\pi}{8}) \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} cos(\frac{\pi}{4}) & -sin(\frac{\pi}{4}) \\ sin(\frac{\pi}{4}) & cos(\frac{\pi}{4}) \end{bmatrix} \tag{2}$$

2.2. Proposed Lifting-Based IDCT Structure

Any plane rotation can be decomposed into a cascade of three lifting steps as

$$\begin{aligned} \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix} &= \begin{bmatrix} 1 & -p \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ u & 1 \end{bmatrix}\begin{bmatrix} 1 & -p \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ p & 1 \end{bmatrix}\begin{bmatrix} 1 & -u \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ p & 1 \end{bmatrix}, \end{aligned} \tag{3}$$

where $p = \frac{1-cos(\alpha)}{sin(\alpha)}$ and $u = sin(\alpha)$. To invert a lifting step, we simply need to subtract out what was added in at the forward step. Hence, the original signal can still be perfectly reconstructed even if the floating-point multiplication results in the lifting steps are rounded to integers, as long as the same procedure is applied to both the forward and inverse routines. Therefore, perfect reconstruction is guaranteed by the lifting structure itself.

Each of the four rotation angles $\{\pi/8, \pi/4, \pi/16, 3\pi/16\}$ previously mentioned is then converted to three lifting steps as in (3). The resulting lifting-based IDCT structure is illustrated on the right side of Fig. 1. To obtain fast implementation, we approximate the floating-point lifting coefficients by hardware-friendly dyadic values of the form $k/2^n$, which can be implemented by only shift and addition operations. Different accuracy versus complexity trade-offs can be achieved by adding or removing dyadic fractions in the approximation of the theoretical parameters, leading to scalable computational capability. As the high accuracy approximation to 64-bit float-point IDCT is our major concern, we select the dyadic lifting parameters listed in Table 1 such that our proposed lifting structure has the highest accuracy with the lowest complexity within 32-bit word length [9]. The accuracy performance is measured by IEEE-1180 criteria, which will be explained in details in Section III. The level of complexity is roughly measured as the total number of additions and bit-shift operations required. Specifically, the 1-D IDCT implementation requires 85 additions and 61 shifts, and 2-D IDCT totally needs total 1362 additions and 1106 shifts. The corresponding forward DCT approximation can be obtained by simply reversing the signal flow and inverting the polarity of lifting parameters.

2.3. Dynamic Analysis

In order to improve approximation accuracy, the input DCT vector coefficients $\mathbf{X}$ need to be up-scaled by certain K bits before they feed into IDCT. The value of K depends on the dynamic range of our IDCT scheme. A similar method in [6] is used to analyze the dynamic range. However, we could not directly assume $\mathbf{X}$ can be randomly generated. Instead, we calculate the maximum or minimum output of each subband by generating the worse-case inputs $\mathbf{X} = \mathbf{DCT_8^{II}x}$, where the integer vector $\mathbf{x}$ can be randomly assigned and $\mathbf{DCT_8^{II}}$ is the ideal type-II DCT matrix. As all lifting parameters are less than unity and implemented with addition and right-shift operations, they minimize the intermediate dynamic range. It can be verified that the absolute value of the worst intermediate result in each lifting steps is less than that of its final output.

For an 8-bit video signal, the input sample values to the DCT $\mathbf{x}$ after motion estimation and compensation are in the range of [-256, 255]. Hence, the 2D-DCT coefficients $\mathbf{X}$ can be shown to be within the 12-bit range $[-2048, 2047]$. The outputs of our 1D-IDCT approximation scheme in Fig. 1 would be still within $[-2048, 2047]$ without the $\sqrt{8}$ down-scaling. The maximum intermediate data have 13-bit range due to the internal butterflies. In the second pass of IDCT, the final IDCT outputs after 3-bit down shifts would be within 9-bit range $[-256, 255]$. Therefore, the dynamic range upper bound of our proposed structure is 13-bit for the 12-bit DCT inputs. That means, for the popular case that the input DCT coefficients are in the range of [-2048, 2047], $K = 3$ is the limit for 16-bit IDCT implementations while $K = 11$ is the upper limit for 24-bit and $K = 19$ is the upper limit for 32-bit architectures.

3. EXPERIMENTAL RESULTS

In this section, we first present IEEE-1180 test results to demonstrate how close the proposed lifting-based IDCT is to the floating-point

Table 1. Lifting parameters for high accuracy IDCT approximation.

Parameters	Theoretical value	Dyadic values	Multiplierless representation
$p_{\pi/8}$	$\frac{1-cos(\pi/8)}{sin(\pi/8)}$	$\frac{3259}{16384}$	$y = w + (w >> 4) - (1 >> 12) - (1 >> 14)$, where $w = (1 >> 3) + (1 >> 4)$;
$u_{\pi/8}$	$sin(\pi/8)$	$\frac{50159}{131072}$	$y = (1 >> 2) + w - (w >> 10)$, where $w = (1 >> 3) + (1 >> 7)$;
$p_{\pi/4}$	$\frac{1-cos(\pi/4)}{sin(\pi/4)}$	$\frac{217167}{524288}$	$y = (1 >> 2) + w + (1 >> 7) + (w >> 10) - (1 >> 19)$,where $w = (1 >> 3) + (1 >> 5)$;
$u_{\pi/4}$	$sin(\pi/4)$	$\frac{46341}{65536}$	$y = w + (w >> 3) + (1 >> 8) + (w >> 13)$, where $w = (1 >> 1) + (1 >> 3)$;
$p_{\pi/16}$	$\frac{1-cos(\pi/16)}{sin(\pi/16)}$	$\frac{25819}{262144}$	$y = (1 >> 4) + w + (1 >> 10) - (w >> 8) - (1 >> 18)$,where $w = (1 >> 5) + (1 >> 8)$;
$u_{\pi/16}$	$sin(\pi/16)$	$\frac{25571}{131072}$	$y = w + (1 >> 7) - (1 >> 12) + (w >> 13)$, where $w = (1 >> 3) + (1 >> 4)$;
$p_{3\pi/16}$	$\frac{1-cos(3\pi/16)}{sin(3\pi/16)}$	$\frac{2485}{8192}$	$y = w - (w >> 5) + (w >> 9)$, where $w = (1 >> 2) + (1 >> 4)$;
$u_{3\pi/16}$	$sin(3\pi/16)$	$\frac{145639}{262144}$	$y = (1 >> 1) + w + (w >> 6) + (w >> 11)$, where $w = (1 >> 4) - (1 >> 7)$;

Table 2. IEEE-1180 test results of the proposed IDCT approximation.

Input Range	Scaling value	1000000 iterations						
		ppe $<= 1$	$pmse$ $<= 0.06$	$omse$ $<= 0.02$	pme $<= 0.015$	ome $<= 0.0015$	pep	Pass test?
$[-256, 255]$	$K = 3$	1	4.44e-001	9.36e-002	4.44e-001	7.12e-003	9.36%	No
	$K = 6$	1	3.93e-002	9.30e-003	3.93e-002	6.14e-004	0.93%	Yes
	$K = 10$	1	2.81e-003	5.95e-004	2.81e-003	3.65e-005	0.06%	Yes
	$K = 18$	1	3.04e-004	1.78e-004	3.10e-005	-2.94e-006	0.02%	Yes
$[-384, 383]$	$K = 3$	1	2.96e-001	6.09e-002	2.96e-001	4.67e-003	6.09%	No
	$K = 6$	1	2.64e-002	6.18e-003	2.64e-002	4.04e-004	0.62%	Yes
	$K = 10$	1	1.73e-003	4.16e-004	1.73e-003	2.73e-005	0.04%	Yes
	$K = 18$	1	2.76e-004	1.72e-004	5.50e-005	-1.09e-006	0.02%	Yes
$[-512, 511]$	$K = 3$	1	2.23e-001	4.54e-002	2.23e-001	3.53e-003	4.54%	No
	$K = 6$	1	1.99e-002	4.64e-003	1.99e-002	3.19e-004	0.46%	Yes
	$K = 10$	1	1.43e-003	3.40e-004	1.43e-003	1.80e-005	0.03%	Yes
	$K = 18$	1	3.22e-004	1.78e-004	2.90e-005	2.72e-006	0.02%	Yes

IDCT. Then, we discuss the results of near-dc test and perfect reconstruction (PR) test. Finally, we show drifting test results in MPEG-2 and MPEG-4 coders under the worst-case drifting assumption.

3.1. IEEE-1180 Test

IEEE-1180 provides a set of specific criteria to measure the compliance of 8x8 IDCT to the ideal IDCT[5]. In the IEEE-1180 test, an 8x8 block of integers is randomly generated and fed into double-precision floating-point forward DCT. The output DCT coefficients are then passed through 64-bit floating-point IDCT and the proposed fixed-point IDCT, respectively. The accuracy is measured based on the reconstructed integers from these two IDCTs. Specifically, the peak pixel-wise error (ppe), peak mean-squared error (pmse), overall mean-square error (omse), peak mean error (pme), and overall mean error (ome) need to compute for the pseudo-random input blocks generated at 10000 and 1000000 iterations. The randomly generated block integers should cover the following five ranges, i.e., $[-5, 5]$, $[-256, 255]$, $[-300, 300]$, $[-384, 383]$ and $[-512, 511]$ with positive and negative sign. One fixed-point IDCT could be considered to be compliant with IEEE-1180 standard if only it satisfies the conditions of $ppe <= 1$, $pmse <= 0.06$, $omse <= 0.02$, $pme <= 0.015$ and $oms <= 0.0015$ for all the input ranges.

Table 2 lists the IEEE-1180 results of the proposed lifting-based IDCT for three K values, which represents 32-bit, 24-bit and 16-bit implementation, respectively. Due to the limited space, Table 2 only includes the results for the three input ranges, i.e., $[-256, 255]$, $[-384, 383]$ and $[-512, 511]$ with the positive sign at 1000000 iterations. In order to show pixel-wise errors more clearly, the percentage of pixel-wise errors (pep) is also included in the table. From these results, we can see that within the 32-bit word length constraint, the proposed lifting-based fixed-point IDCT solution delivers super high-accuracy approximation ($omse = 1.78e - 04$ for $K = 18$); and 24-bit implementation also leads a very high accuracy approximation ($omse = 5.95e - 004$ for $K = 10$). The percentage of pixel-wise error clearly shows the proposed algorithm is very accurate (less than 0.06% mismatch errors) for 24-bit and 32-bit implementations. Moreover, $K = 6$ is the minimal up-scaling bits for our implementations in order to pass IEEE-1180 tests fully. Although the 16-bit implementation of $K = 3$ cannot pass IEEE-1180 tests, it discloses an interesting observation that only high accurate lifting parameters would not certainly lead to high accuracy approximation. The relative error analysis are addressed in our other papers [10]. In [9], we also have another 16-bit implementation solution of our scheme with different dyadic values, which can pass IEEE-1180 test for the input range $[-256, 255]$.

3.2. Near-DC Test and PR Test

Near-DC test is to measure DC leakage problem in one transform, i.e., the bandpass and highpass subbands should have no DC leakage. This means that these high-frequency subbands should have at least one vanishing moment. The zero DC leakage can prevents the annoying checkerboard artifacts that can occur if high-frequency bands are severely quantized. It is an important measurement for fixed-point IDCT transform [4]. In near-dc test, the input $8x8$ DCT coefficients are set to be zeros except DC coefficient dc and the highest frequency DCT coefficient h. dc is set to be one of integers in the range of $[-2048, 2047]$. If dc is even, $h = 1$; otherwise, $h = 0$. Then, pixel-wise errors are computed between the reconstructed values from the proposed IDCT and those from double-precision floating-point IDCT. As long as the scaling bit K is greater than 4, our proposed lifting-based IDCT can pass near-dc test, i.e., pixel-wise errors are not larger than 1 [9].

Due to the closely-matched nature of the forward-inverse lifting, the mismatch between our lifting-based IDCT and DCT is mitigated. In fact, perfect reconstruction (no mismatch at all) for the integer space in the range $[-256, 255]$ is achievable when the DCT coef-

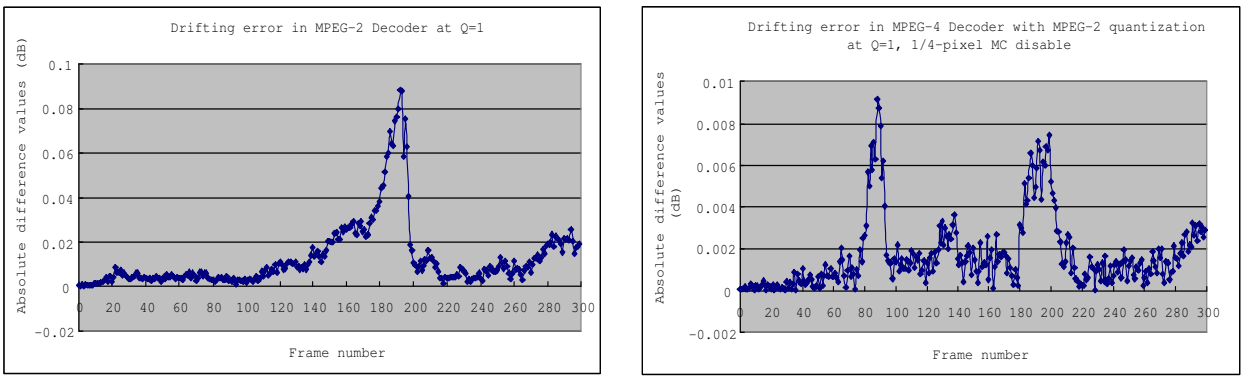

Fig. 2. IDCT drifting tests for 'Foreman' CIF sequence of 300 frames. Left figure: drifting errors in MPEG-2 decoder at $Qp = 1$; Right figure: drifting errors in MPEG-4 decoder with MPEG-2 quantization module used at $Qp = 1$ and 1/4-pixel motion compensation disable.

ficient range is extended to $[-8192, 8191]$ (14-bit representation is needed instead of 12-bit). This bit expansion comes from the fact that there are still butterflies left in our structure shown in Fig. 1. And each of the butterflies carries an expansion factor of $\sqrt{2}$.

3.3. Drifting Tests

Drifting is the effectively-random deviation of decoders from the values that are modelled in the encoder. Drifting tests are carried on the coders of MPEG-2 and MPEG-4. In the encoders, double-precision floating-point DCT and IDCT are used, and all pictures are coded as P-frames except the first one is coded as I-frame. To evaluate the drifting effects better, we consider the extreme worse-case by disabling the intra macroblock refresh in encoder and setting quantization step size Qp to be 1. At the decoder side, the proposed lifting-based IDCT and double-precision floating-point IDCT are used to reconstruct the sequences, respectively. PSNR values are then computed and their absolute difference values are used to evaluate the drifting effects. Fig. 2 illustrates the experiment results of Foreman CIF sequence (total 300 frames) in MPEG-2 and MPGE-4 coders for our 32-bit fixed-point IDCT implementation at $K = 18$. In MPEG-2 decoder, the average drifting is 0.0043dB among the first 131 frames, and 0.01258dB among the total 300 frames. In MPEG-4 with MPEG-2 quantization module and quarter-pixel motion compensation disable, the average drifting is 0.001369dB among the first 131 frame, and 0.001714dB among the 300 frames. These results clearly confirm that the proposed IDCT transform has very high accurate approximation and leads to very little drifting effects.

4. CONCLUSION

In this paper, a non-scaled lifting-based multiplierless IDCT approximation structure is proposed, which allows bit-exact reconstruction given that the range restriction on the DCT coefficients is extended by 2 more bits. The proposed structure comprises of regular stages, convenient for pipelining, and allows the computational scalability with different accuracy-versus-complexity trade-offs. A very high-accuracy fixed-point solution is also presented, which provides a very close approximation of the floating-point IDCT. The algorithm delivers very low overall mean-square errors in the e-004 range. Such super high-accuracy level leads to almost drifting-free reconstruction when pairing with a 64-bit floating-point IDCT in the encoder. To our best knowledge, it is the highest accuracy that we can achieve in our IDCT fixed-point approximation structure within the 32-bit constraint. Due to the limited space, rate-distortion analysis of our proposed scheme will be further addressed in our other papers.

5. REFERENCES

[1] J. Mitchell, W. Pennebaker, C. Frogg, and D. LeGall, *MPEG Video Compression Standard*, Chapman and Hall, New York, 1996.

[2] W. Chen, C. Harrison, and S. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Communications*, vol. COM-25, pp. 1004–1011, 1977.

[3] C. Loeffler, A. Lightberg, and G. Moschytz, "Practical fast 1-d DCT algorithms with 11 multiplications," in *Proc. IEEE Int. CASSP*, 1989, vol. 2, pp. 988–991.

[4] G. Sullivan and A. Luthra, "Call for proposals on fixed-point 8x8 IDCT and DCT standard," *MPEG input contribution N7099*, April 2005.

[5] IEEE CAS Standards Committe, "IEEE standard specification for the implementation of 8x8 invrese discrete cosine transform," *IEEE Standard 1180-1990*, December 1990.

[6] J.Liang and T. D. Tran, "Fast multiplierless approximations of the DCT with the lifting scheme," *IEEE Trans.on Signal Processing*, vol. 49, pp. 3032–3044, Dec. 2001.

[7] A. C. Zelinski, M. Puschel, S. Misra, and J. C. Hoe, "Automatic cost minimization for multiplierless implementations of discrete signal transforms," in *Proc. IEEE CASSP*, 2004, vol. 5, pp. 221–225.

[8] E. Feig and S. Winograd, "On the multiplicative complexity of discrete cosine transform," *IEEE Trans. Information Theory*, vol. 38, pp. 1387–1391, July 1992.

[9] T. D. Tran, L. Liu, and P. Topiwala, "Improved high-accuracy and low-complexity multiplierless DCT/IDCT based on the lifting scheme," *MPEG input contribution M13728*, July 2006.

[10] Lijie Liu and Trac D. Tran, "Rate distortion performance of two lifting-based IDCT schemes," in *Conference on Information Sciences and Systems*, Mar. 2007.