Improvement of HEVC Inter-coding Mode Using Multiple Transforms

Pierrick Philippe
Orange, b<>com
pierrick.philippe@orange.com

Thibaud Biatek
TDF, b<>com
thibaud.biatek@tdf.fr

Victorien Lorcy
b<>com
victorien.lorcy@b-com.com

***Abstract*—Multiple transforms have received considerable attention recently, especially in the course of an exploration conducted by MPEG and ITU toward the standardization of the next generation video compression algorithm. This joint team has developed a software, called the Joint Exploration Model (JEM) which outperforms by over 25% the HEVC standard. The transform step in JEM consists in Adaptive Multiple Transforms (AMT) and Non-Separable Secondary Transforms (NSST) which are designed and adapted to the intra-coding modes. In inter-coding, only the AMT is allowed and it is restricted to a single set of five transforms. In this paper, adaptive transforms schemes suitable for inter-predicted residuals are designed and proposed to improve the coding efficiency. Two configurations are evaluated for the proposed designs, providing an average bitrate saving of roughly 1% over HEVC with unchanged decoding time.**

I. Introduction

HEVC/H.265 is the latest video coding standard [1], released in January 2013 as the successor of AVC/H.264 [2]. HEVC provides more than 50% of bandwidth reduction compared to AVC, for the same perceived visual quality. Consequently it is well adapted to larger resolutions such as Ultra-High-Definition (UHD) contents [3]. With next generation formats in focus, such as 360 degrees video, MPEG and ITU jointly established the Joint Video Exploration Team (JVET) in October 2015 to prepare the next generation of video coding standard, beyond HEVC. A Joint Exploration Model (JEM) has been developed and this software provides more than 25% of coding efficiency compared to HEVC in Random-Access configuration (RA) [4].

The JEM, initially built at the top of the HEVC Test Model 16.6 (HM-16.6) [5], introduces many new tools [6]. Among those new tools, the transform stage introduces the notion of transform competition through two stages.

The first stage, called Adaptive Multiple Transforms (AMT) [7], proposes a block-level flag that signals whether the classical DCT2 (Discrete Cosine Transform kernel of type II) is used. If not, additional indexes are transmitted to signal the selected horizontal and vertical transforms, in a list of trigonometric kernels [8] (DCT and DST of types I to VIII). It must be noted that the indexes point to transform sets that depends on the Intra Prediction Mode (IPM) for intra residuals while a single set is considered for inter-predicted residuals. The DCT8 and DST7, combined in horizontal and vertical directions are available there.

A second transform-stage can be added for intra-coded blocks, called Non-Separable Secondary Transforms (NSST). Those transforms are based on hypercube Givens rotations [9] applied on the lower frequency coefficients after the AMT transformation. The impact of these tools has been evaluated for the JEM, where they each provide around 2% of bit rate savings [10]. The impact of the transform-related tools undeniably represents a significant part of the coding gains in the JEM version (JEM5 at the time of writing this article).

Several technologies have been proposed to improve the transform stage. For example, in [11], an extension of the AMT transform set is proposed by introducing two additional transforms kernels. This enables bitrate savings of roughly 0.3%. In addition, an alternative transform set design has been proposed in [12] to reduce the required computational power by replacing the most expensive transform kernels, providing in the range of 50% of encoding complexity reduction for equivalent compression performance.

It can be noticed that none of these methods have been optimized and deeply investigated for the inter coded residuals. The AMT authorizes a wide variety of transform kernels for intra residuals while only DST7 (Discrete Sine Transform if type VII) and DCT8 (DCT of type VIII) are considered in inter. Regarding the secondary transform stage, NSST is only activated for intra slices.

Several approaches have been proposed in the litterature to improve the transform stage efficiency on motion-compensated residuals. In [13], the authors propose to adaptively rotate the DCT2 for inter residuals (ROT). In this approach, the DCT2 is multiplied by a cascade of rotational transforms, where the angles composing the global rotation are estimated thanks to a gradient-based searching algorithm. Then a syntax scheme is proposed to signal the angles and whether a region uses the rotational transforms or not, which leads to 3.9% of coding gains compared to AVC. In [14], inter frame prediction residuals are modeled under the assumption that image intensities follow a first-order Markov mode in the direction of the motion trajectory. An adaptive transform, which requires update at the decoding side, is in competition with the DCT2 and the results reveal that 2% gains are achieved compared to AVC. In [15], the application of graph-based transforms (GBT) is also explored on residuals generated with HEVC inter-mode, where GBT achieves substantial gains compared to the DCT2 and KLT. The subject of transform competition in the case of

ISBN 978-0-9928626-7-1 © EURASIP 2017

inter-coded residuals remains not well covered in the literature.

Although, the ROT and GBT approaches previously mentioned are promising in term of performance, they require the transmission or update of the transform coefficients which can be an issue for hardware implementations: this typically prevents these transforms from fast implementations.

In this paper, an improved AMT scheme with adaptive transform set selection for inter-coded slices is proposed to resolve these issues. The proposed scheme extends the JEM using the same transform kernels, and dynamically adapts the transform sets used on inter residuals and provides an improved coding efficiency over HEVC.

This paper is organized as follows. The RDOT criterion is first introduced as a mean to select appropriate set of transforms for inter-coding. Then, the selection of the number of transforms is discussed and the coding performance obtained while the number of transforms is increased is presented. In the subsequent section, an adaptive transform set approach is discussed and evaluated.

II. Transform sets

A. Rate Distorsion optimized transforms

The Rate-Distortion Optimized Transforms (RDOT) have been introduced in [16] to efficiently learn transforms for a given set of residuals. In [17], the RDOT method is used to learn optimal sets of transforms for intra-predicted residuals in HEVC for the general case of non-separable transforms, then extended for separable transforms and Discrete Trigonometric Transforms (DTT). In this paper, set of DTTs is considered as a support to learn the transform sets used in the proposed design, in a fashion similar to the transforms adopted in JEM [7]. Hence, the RDOT learning aims at finding an optimal pair of vertical and horizontal transforms $\{\boldsymbol{A}_v, \boldsymbol{A}_h\}$, for a set of residuals $\{\boldsymbol{x}_i\}$ defined by solving the following minimization problem:

$$\{\boldsymbol{A}_v, \boldsymbol{A}_h\}_{opt} = \underset{\boldsymbol{A}_v, \boldsymbol{A}_h}{\operatorname{argmin}} \sum_{\forall i} \min_{\boldsymbol{c}_i} \left(||\boldsymbol{x}_i - \boldsymbol{A}_v^T \boldsymbol{c}_i \boldsymbol{A}_h||_2^2 + \lambda ||\boldsymbol{c}_i||_0 \right) \tag{1}$$

where $(\boldsymbol{A}_v, \boldsymbol{A}_h)$ the horizontal and vertical transforms and $\boldsymbol{c}_i$ the transformed and quantized residual. As demonstrated in [17], the Lagrangian multiplier λ depends on the quantization accuracy. In this paper, a transform set is learned based on inter-predicted residuals extracted from bitstreams coded with HEVC in RA configuration, for 70 sequences (with resolution varying from 240p to 2160p). Over 10 million of residuals blocks are considered at this stage.

For the purpose of this article, the learning process is performed to select a set of transforms for inter-predicted residual. Therefore the learning process is turned into a selection process of M pairs of vertical and horizontal transforms in the set of all possible discrete trigonometric transforms.

The learning design is illustrated in Algorithm 1 [17], [18]. For all possible transform sets, the residuals are clustered into classes related to each transform pairs according to the RDOT metric (Class$_m$). When a set minimizing the RDOT metric is reached, the convergence criterion is achieved and the current set is selected.

```
Data: Inter-predicted residuals $\boldsymbol{x}$ from a given size
Result: Set of $M$ pairs $\{A_{h,m}, A_{v,m}\}$
Initialization: random classification into $M$ classes;
while !convergence do
    for $m = 0$ to $M - 1$ do
        Select $\{\boldsymbol{A}_v, \boldsymbol{A}_h\}_{opt}$ for Class$_m$ using Eq 1.
    end
    foreach block $\boldsymbol{x}$ do
        for $m = 0$ to $M - 1$ do
            $\delta_m = ||\boldsymbol{x} - \boldsymbol{A}_{v,m}^T \boldsymbol{c} \boldsymbol{A}_{h,m}||_2^2 + \lambda ||\boldsymbol{c}||_0$
        end
        $m* = \operatorname{argmin}_m(\delta_m)$
        Class$_{m*}$.append($\boldsymbol{x}$)
    end
end
```

Algorithm 1: RDOT learning design

With the considered learning design, the transform sets are built independently for each block sizes. In a second pass, the obtained sets are homogenized to obtain a set of transforms common for all sizes, from 4x4 to 32x32 blocks.

Table I gives the transform sets obtained after the learning process, they contain from 1 to 9 transforms. According to the HEVC terminology, each TU (Transform Unit) will consider using one of those transforms for each inter-residual block. The number of transform per set is chosen to be $1 + 2^b$ to anticipate the signalization of the selected transform from the encoder to the decoder. A flag indicates whether the first transform is used, if not, an additional code on b bits is conveyed to signal the selected transform.

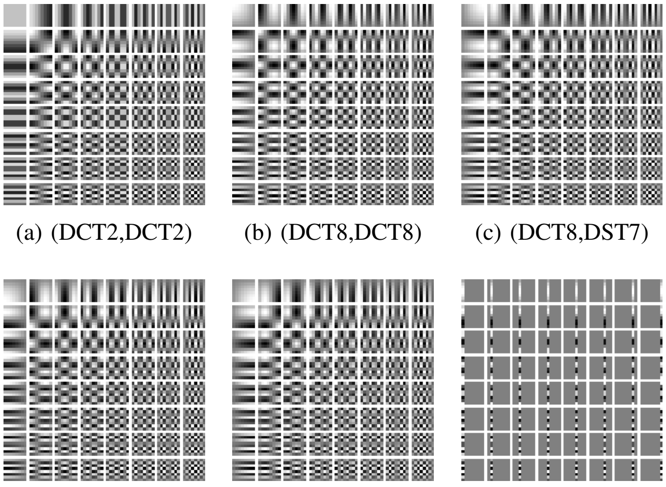
TABLE I: Transform sets obtained through the learning algorithm. E.g. set 3 includes T0, T1 and T4 DTT kernel pairs.

			Transform Set				
Index	Row	Col.	1	2	3	5	9
T0	DCT2	DCT2	o	o	o	o	o
T1	DST7	DST7	·	o	o	o	o
T2	DST7	DCT8	·	·	·	o	o
T3	DCT8	DST7	·	·	·	o	o
T4	DCT8	DCT8	·	·	o	o	o
T5	DST1	DST1	·	·	·	·	o
T6	DST7	-	·	·	·	·	o
T7	-	DST1	·	·	·	·	o
T8	DCT8	-	·	·	·	·	o

As can be seen, the learning algorithm teaches that the DCT2 transform, for both row and column directions, is confirmed as the optimal transforms when used solely. The DST7 and DCT8 are the most frequent transform kernels for transform sets up to 5 transforms (transform sets 2, 3 and 5).

It must be noted that the TrSet 5, although designed independently, matches the transform set as used in the JEM. For TrSet9, it is remarked that one single additional transform kernel (DST1) is added to those of HEVC (as DCT8 and DST7 are dual, i.e. identical as a vector basis reversal).

Some of the 2D transforms, are illustrated on figure (1a-1f). Figure (1a) displays the 2D-DCT2 as frequently encountered in video coding. Different combinations of DCT8 and DST7 are displayed (1b-1e). As can be noticed, each consider a particular spatial localization. Figure (1f) performs the transform decomposition on the vertical axis, as such it is appropriate for residual patterns with banded vertical textures.



(a) (DCT2,DCT2) (b) (DCT8,DCT8) (c) (DCT8,DST7)

(d) (DST7,DCT8) (e) (DST7,DST7) (f) A_v=DST1

Fig. 1: 8x8 Transform basis of 2D-transforms used in the proposed systems. The five first transforms are used in the transform set containing 5 transforms. (f) represents the 2D-transforms (T7) when only the DST1 acts in the vertical direction.

B. Coding performance with transform sets

Five transform sets have been determined in the learning process, this section deals with testing each of them in a coding environment.

The HEVC coding scheme is extended to allow the usage of the proposed multiple transforms. Consistent with the approach in [17], a flag indicates whether the legacy HEVC transform, (DCT2), is used. If not, an additional code is conveyed on 1,2,3 bits for respectively Transform Sets 2, 3, 5 and 9. These flag and code are coded at the HEVC Transform Unit syntax when the luma residual signal is significant (it contains one or more coefficients different from zero).

The performances are evaluated in the Common Test Conditions (CTC), as defined by the JVET group. The test set includes 25 video clips with resolutions from 240 lines to 4096x2160 pixels [19]. The coding configuration, is Random Access, as such an intra picture is inserted approximatively every second, the intermediate frames are coded with a hierarchical B frames structure with a GOP size of 16. Both HEVC implementation and the proposed coding schemes are evaluated in this configuration, both codecs are based on the latest HEVC reference model (HM16.6).

	Transform Set			
Resolution	**2**	**3**	**5**	**9**
A1 (4K)	-0.3%	-0.6%	-0.8%	-0.9%
A2 (4K)	-0.2%	-0.4%	-0.5%	-0.6%
B (1920x1080)	-0.4%	-0.7%	-1.0%	-1.2%
C (832x480)	-0.2%	-0.4%	-0.5%	-0.8%
D (416x240)	-0.3%	-0.6%	-0.7%	-1.0%
F (various)	0.1%	0.0%	0.0%	-0.2%
Average	**-0.2%**	**-0.5%**	**-0.6%**	**-0.8%**
Add. Complexity	**5%**	**9%**	**15%**	**28%**

TABLE II: Coding performances of obtained with the transform sets expressed in bit rate savings compared to HEVC (a negative number indicates gains).

Table II presents the results expressed with the BD-rate metric commonly used in video coding [20]. The percentage expresses the relative bit rate decrease over the HEVC which serves as the anchor for this study. The estimation is estimated over a bit rate range driven by a quantization parameter Qp from 22 to 37.

It can be noticed that the gains increase as the number of transforms increase, from -0.2% of bit savings to -0.9% for the Transform Set 9. One also notice that the added encoding complexity with respect to HEVC also increases with the number of transforms, up to 28%.

These results highlight the performance of transform competition in the context of inter-coding solely, as the transform competition is enabled only for inter predicted residuals. The coding gains are lower than the ones obtained in the case of AMT for Intra: one source of explanation comes from the fact that inter-coding includes a significant number of blocks perfectly predicted for which there is no residual. Those blocks do not take profit from the additional transforms.

It can also be noticed, notably on the content of Class F, that comprises screen content scenes with mostly static scenes, that increasing the number of transforms has no effect on the coding performance: the potential gain vanishes as the rare coded residuals taking benefit from this increase is counterbalanced by the transform signaling.

III. Adaptive Transform Sets

On the one hand, the benefit from an increased number of transforms is justified for inter-predicted residuals as stated in the previous section. On the other hand, the possibility for the encoder to limit its number of transforms seems also motivated because in some cases, such as easily predictable regions (i.e. motionless and immobile areas), a flat residual is more probable and thus a DCT2 could be sufficient.

Consequently, to further increase the compression efficiency, this paper proposes to dynamically adapt the number of transforms per Coding Tree Unit (CTU, i.e. per 64x64 pixel blocks).

To summarize, the advantages of an adaptive transform set design are the following:

- Enlarge transforms sets when necessary : reduced distortion in the R-D trade-off as complex residuals take advantage of the multiple transforms
- Reduce transforms sets when necessary : reduced bitrate in the R-D trade-off by avoiding wasting bits signaling the transforms when unnecessary.

TABLE III: Conditional probabilities between the current transform set and the transform set from the co-located block

$TS_{Cur} \backslash TS_{Col}$	0	1	2	Code
0	92%	51%	42%	0
1	5%	28%	30%	10
2	3%	21%	28%	11
bps	1.11	1.69	1.85	

A. Principles of Adaptive Transform Sets (ATS)

In the proposed design, the encoder is allowed to modulate the number of transforms used in inter-prediction mode. To enable enough flexibility to the encoder, it is proposed to dynamically adjust the transform set at the CTU level, signaled in a differential way.

The five transform sets defined in table I are directly used in this ATS design. The first transform set is basically a disabled-AMT mode (DCT2 only) while the four other sets used DCT2 plus 1, 2, 4 or 8 transforms.

B. Transform Set Signaling

The transform set index is signaled at the top of the CTU in a differential way. Indeed, it has been observed, especially, that the probabilty of having a given transform set index in a temporal layer is strongly correlated to the transform set index value of the colocated (same position) CTU in the lower temporal layer. Thus, it is wise to signal the transform set using a code based on the conditional information, as shown in table III. Using that method, the average cost for the current transform set is reduced to 1.11 bits on average when the collocated transform set includes only the DCT2. Note that the first bit of the table is encoded using a CABAC code. Consequently, there is an efficient signaling for sequences where fewer transforms are required.

C. Performance and Encoding complexity consideration

For the adaptive transform set, the encoder successively encodes each CTU for each transform set, therefore one of the main impact of the proposed design is its increased complexity. Indeed multiple redundant passes for the partitioning and prediction are reiterated.

To accelerate the encoding decisions, two acceleration tricks are considered. First, an early-termination method is implemented to break the Rate-Distortion Optimized (RDO) encoding if a CTU does not contain any residual for the first Transform Set, it is judged unnecessary to explore alternate transform sets. In addition, another technique can be implemented to reuse the quad-tree partitioning derived using a particular transform set for another one. In this case, transform sets are tested from the richer in terms off transforms (e.g. from transform set 9).

The ATS schemes are evaluated through several configurations: the simpler configurations use two transform sets, e.g. 1,2 can code a CTU either with the DCT2 or using the pair of transforms as selected for transform set (refer to table I). The number of transforms is progressively increased up to 9.

	ATS Configuration					
Resolution	**{1,2}**	**{1,3}**	**{1,5}**	**{1,9}**	**{1,5,9}**	**{1,2,3,5,9}**
A1 (4K)	-0.5%	-0.9%	-1.1%	-1.2%	-1.3%	-1.3%
A2 (4K)	-0.4%	-0.6%	-0.8%	-0.8%	-0.9%	-1.0%
B (1920x1080)	-0.6%	-0.9%	-1.2%	-1.4%	-1.5%	-1.6%
C (832x480)	-0.5%	-0.7%	-0.8%	-1.0%	-1.2%	-1.3%
D (416x240)	-0.5%	-0.7%	-0.9%	-1.2%	-1.3%	-1.4%
F (various)	-0.1%	-0.2%	-0.2%	-0.5%	-0.5%	-0.5%
Average	-0.4%	-0.7%	-0.9%	-1.0%	-1.1%	-1.2%
Add. Complexity	89%	94%	96%	105%	182%	296%

TABLE IV: Coding performances obtained with Adaptive Transform Set configurations

The ATS systems with 2 transform sets ({1,2}, {1,3},{1,5},{1,9}) have coding gains progressing from 0.4% to 1% on average, although the added complexity is significantly higher to the one of the AMT systems reported in table II. The ATS systems with a larger number of transforms demonstrate that additional gains that can be obtained when the number of transforms is precisely adapted to the nature of the CTU. Two configurations are investigated {1,5,9} and {1,2,3,5,9}. Although, the coding gain increases up to 1.2% the added complexity seems to discard this approach.

Both the AMT and the ATS systems ensure that the decoding complexity is kept practically unchanged compared to the HEVC, as roughly the same number and sizes of inverse transforms are applied.

The main drawback for the ATS approach remains the added complexity, although it permits to significantly outperform the AMT systems (gains progress from 0.8% to 1.2%). This is why the relationship between transform selection and partitioning should be better understood to reduced the redundancies in the encoding process.

IV. Conclusion

In this paper, an adaptive transform set design, using trigonometric kernels, is proposed to improve the coding efficiency of inter-predicted residuals in HEVC.

To further increase the performance, transforms sets, called AMT, with from 1 to 9 transforms are designed, in a rate-distortion sense. This design confirms the value of DCT2 for inter-coding when used solely and also confirms that the DST7/DCT8 are efficient kernels for these residuals. The AMT with 5 transforms are identical with the one derived independently for the ITU/MPEG JEM software. The design is conservative with the HEVC transforms as a single transform kernel (the DST1) is added, and the decoding complexity remains the one of this standard. Under strict testing conditions, it is shown that AMT can provide from 0.2% up to 0.8% with a reasonable increase of the encoder complexity.

The Adaptive Transform Sets are also introduced to further increase the coding gains. Thanks to ATS bit rate gains that were in the range of 0.8% for the AMT reach 1.2% at the expense of a significant complexity increase at the encoding side. Hence, the further work is required to reduce the complexity in the most performing configuration to increase the attractivity of such solution.

References

[1] G.-J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 22, no. 12, pp. 1649–1668, December 2012.

[2] T. Wiegand, G.-J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 13, no. 7, pp. 560–576, July 2003.

[3] T.-K. Tan, R. Weerakkody, M. Mrak, N. Ramzam, V. Baroncini, J.-R. Ohm, and G.-J. Sullivan, "Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 26, no. 1, pp. 1051–8251, January 2016.

[4] X. Li and K Suerhing, "JVET-E0003: JVET Common test conditions and software reference configurations," Tech. Rep., JVET AHG report: JEM software development (AHG3), San Diego, February 2016.

[5] C. Rosewarne, B. Bross, K. Sharman, and G.-J. Sullivan, "JCTVC-U1002: High Efficiency Video Coding (HEVC) Test Model 16 (HM 16) Improved Encoder Description," Tech. Rep., Joint Collaborative Team on Video Coding (JCT-VC), Warsaw, October 2015.

[6] J. Chen, E. Alshina, G.-J. Sullivan, J.-R. Ohm, and J. Boyce, "JVET-E1001: Algorithm Description of Joint Exploration Test Model 5," Tech. Rep., Joint Video Exploration Team (JVET), Geneva, February 2017.

[7] X. Zhao, J. Chen, M. Karczewicz, L. Zhang, and X. Li, "Enhanced Multiple Transform for Video Coding," *Data Compression Conference (DCC)*, March 2016.

[8] V. Britanak, P.C. Yip, P. Yip, and K.R. Rao, *Discrete Cosine and Sine Transforms: General Properties, Fast Algorithms and Integer Approximations*, Academic, 2007.

[9] A. Said, X. Zhao, M. Karczewicz, H. Egilmez, V. Seregin, and X. Li, "Highly Efficient Non-Separable Transforms for Next Generation Video Coding," *Picture Coding Symposium (PCS)*, December 2016.

[10] E. Alshina, A. Alshin, K. Choi, and M. Park, "JVET-B0022: Performance of JEM1.0 Tools Analysis by Samsung," Tech. Rep., Joint Video Exploration Team (JVET), San Diego, February 2016.

[11] V. Lorcy and P. Philippe, "JVET-C0022: Proposed Improvements to the Adaptive Multiple Core Transform," Tech. Rep., Joint Video Exploration Team (JVET), Geneva, June 2016.

[12] T. Biatek, V. Lorcy, P. Castel, and P. Philippe, "Low-Complexity Adaptive Multiple Transforms for post-HEVC Video Coding," *Picture Coding Symposium (PCS)*, December 2016.

[13] Z. Gu, W. Lin, B.-S. Lee, and C.-T. Lau, "Rotated Orthogonal Transform (ROT) for Motion-Compensation Residual Coding," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4770–4781, December 2012.

[14] H. J. Leu, S. D. Kim, and W. J. Kim, "Statistical modeling of inter-frame prediction error and its adaptive transform," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 519–523, April 2011.

[15] H.-E. Egilmez, A. Said, Y.-H. Chao, and A. Ortega, "Graph-Based Transforms for Inter Predicted Video Coding," *IEEE International Conference on Image Processing (ICIP)*, September 2015.

[16] O.-G. Sezer, O. Harmanci, and O.-G. Guleryuz, "Sparse Orthonormal Transforms for Image Compression," *IEEE International Conference on Image Processing (ICIP)*, October 2008.

[17] A. Arrufat, "Multiple Transforms for Video Coding," *PhD Thesis, INSA Rennes*, December 2015.

[18] A. Arrufat, P. Philippe, K. Reuze, and O. Deforges, "Low-Complexity Transform Competition for HEVC," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

[19] K Suerhing and X. Li, "JVET-B1010: JVET Common test conditions and software reference configurations," Tech. Rep., Joint Video Exploration Team (JVET), San Diego, February 2016.

[20] Gisle Bjøntegaard, "VCEG-M33: Calculation of Average PSNR Differences Between RD-Curves," Tech. Rep., Video Coding Experts Group (VCEG), Austin, April 2001.

ISBN 978-0-9928626-7-1 © EURASIP 2017