ANNEALED LEARNING BASED BLOCK TRANSFORMS FOR HEVC VIDEO CODING

Saurabh Puri[⋆†] *Sébastien Lasserre*[⋆] *Patrick Le Callet*[†]

[⋆] Technicolor
975 avenue des Champs Blancs, CS 17616, 35576 Cesson-Sevigne Cedex, France
[†] IRCCyN, Ecole Polytechnique de l'Universite de Nantes
Rue Christian Pauc-BP 50609, 44306 Nantes Cedex 3, France

ABSTRACT

Most of the recent video compression standards employ the Discrete Cosine Transform (DCT) for transforming the residual signal in order to remove spatial correlation and to achieve higher compression efficiency. However, by careful adaptation of transforms to the video content, a better set of integer transforms can be obtained. This paper proposes a new on-the-fly block-based transform optimization technique which involves first the classification of the residual blocks based on the cost of encoding the block, and then the generation of new optimized transforms for each class. An annealing based learning technique is further proposed in this paper in order to improve the performance of the optimization algorithm. The algorithm is tested using the latest HEVC test software where an optimized set of transforms is learned on the first frame of the HEVC test sequences and then applied to the subsequent frames in a Random Access (RA) and All Intra (AI) configuration. The results shows that this method can gain over 2% in terms of Bjontegaard Delta (BD)-rate compared to standard HEVC encoder in AI configuration and nearly 1.5% in RA.

Index Terms— adaptive directional transforms, annealing, classification

1. INTRODUCTION

The video codecs in the past have employed the integer based Discrete Cosine Transform (DCT) as the dominant block transform as it approximates the Karhunen-Loève transform (KLT) for Gaussian-Markov fields which are supposed to be a decent approximation of natural residual images. Another important reason behind the high affinity towards DCT is that it can be efficiently implemented on hardware using fast computation algorithms [1]. In the latest video coding standard, named High Efficiency Video Coding (HEVC), a Discrete Sine Transform (DST) is employed for the 4×4 Luma residual block. As the DCT is not efficient in coding regions with high amount of variations, an extensive work in the past has been done in adapting the transforms to the video content which has led to three main approaches.

The first approach involves using systematic transforms which can perform better de-correlation of the residual signal compared to the DCT. This approach is used in [2] and [3] by employing a set of directional DCTs in order to exploit the directionality of the residuals.

In the second approach, the transforms obtained after the content adaptation are used and no additional signaling of transform indexes or transforms itself is done at the encoder. However, on the decoder side, this information is implicitly derived from the bit-stream. This approach is used by [4]-[5],[6] and [7]. In [4], the Mode-Dependent Direction Transform (MDDT) was proposed. The transforms are indexed by intra-prediction modes which means that no extra syntax is required in MDDT. Further, MDDT formed a part of the KTA codec [8], the test codec following the standard h.264/AVC. In [9] and [10], the MDDT algorithm is modified by introducing ℓ_0-norm regularized optimization in order to obtain robust learning algorithm and to enforce the sparsity-constraint in the optimization process. The paper [7] evaluates MDDT in HEVC. In [5], a different approach of adapting the transforms to the content has been proposed where the transform is obtained from a motion compensated residual block both at the encoder and decoder eliminating the need of encoding the adapted transform bases into the bit-stream.

The third approach is similar to the second except that the additional information is explicitly signaled by the encoder which leads to extra bit-rate. In [11], a set of adaptive transforms are used inside a rate distortion test loop and the best transform bases is selected using the R-D cost and explicitly indexed to the decoder. However, this was further modified by [12] by combining the rate distortion optimization algorithm with the MDDT in order to save the explicit signaling of the transform indexes. The work proposed in this paper employs the third approach and therefore, the adapted transforms are indexed explicitly to the decoder.

The work on adaptive transforms in the literature has mainly focused on developing algorithms for h.264/AVC. In the recent standard (i.e. HEVC), the residual energy has been further reduced by employing additional directional prediction modes for intra-prediction and by using higher sub-pixel

accurate filters for motion prediction. Therefore, the gain obtained by using directional transforms is less compared to previous standards [13]. However, the residuals may contain some statistical correlation within the same block or across various blocks of the same frame of a video sequence. This paper introduces a method to find transforms that are capable of exploiting such correlation.

The remainder of this paper is organized as follows. In Section 2, a detailed mathematical description of the optimization process as well as the implementation details along with the changes done in the codec are presented. The experimental results are illustrated in the section 3. Finally, the paper is concluded in section 4.

2. ALGORITHM CONCEPTION AND IMPLEMENTATION

Here, the mathematical basis of the optimization process used to design the newly adapted orthonormal transforms is detailed. The algorithm is inspired from the work described in [14]. In this paper, the algorithm learns on a set of residual blocks (in contrast to blocks from natural images in [14]) in order to classify the residuals depending on the different transforms and, at the same time, generates optimized non-separable transforms. Moreover, the algorithm is further adapted to the current block-based coding engine (HEVC) and the true rate of coding the transformed residual block in contrast to the $\ell 0$-norm described in [14] is used for transform optimization. This helps in the accurate computation of the rate and hence, a better adaptation to the content.

Let $\mathbf{r}^j$ (where $j = 1, \ldots, J$) be the j^{th} residual vector in a particular video sequence. Each of the residual can be transformed using K different transforms $\mathbf{T}_k$ (where $k = 1, \ldots, K$). These transforms are initialized with a set of oriented DCT-like non-separable transforms [15]. The optimization process is split into two parts i.e. classification of the residual blocks into K different classes S_k associated to each transform $\mathbf{T}_k$ using equation (1), and generation of the new set of transforms $\mathbf{T}'_k$ for a particular class S_k by minimization of the distortion on class S_k using equation (2).

$$label\{\mathbf{r}^j\} = \text{argmin}_k \left\{ \underbrace{\|\mathbf{r}^j - \mathbf{T}_k\mathbf{Q}^{-1}(\mathbf{c}_k^j)\|^2}_{\text{I}} + \lambda(\underbrace{R(\mathbf{c}_k^j)}_{\text{II}} + \underbrace{R_T}_{\text{III}}) \right\} \quad (1)$$

$$\mathbf{T}'_k = \text{argmin}_{\mathbf{H}} \left(\sum_{j \epsilon S_k} \|\mathbf{r}^j - \mathbf{H}\mathbf{Q}^{-1}(\mathbf{c}_k^j)\|^2 \right) s.t. \mathbf{H}^T\mathbf{H} = \mathbf{I} \quad (2)$$

In equation 1, term I denotes the distortion computed between the residual and the reconstructed residual obtained after the inverse quantization and inverse transform of encoded coefficients $\mathbf{c}_k^j$. Term II is the rate of encoding the coefficients $\mathbf{c}_k^j$ with $\mathbf{c}_k^j = \mathbf{Q}(\mathbf{T}_k^{\mathbf{T}} r^k)$ where $\mathbf{Q}$ is the quantization function and term III is the cost of encoding the transform index for each transform block.

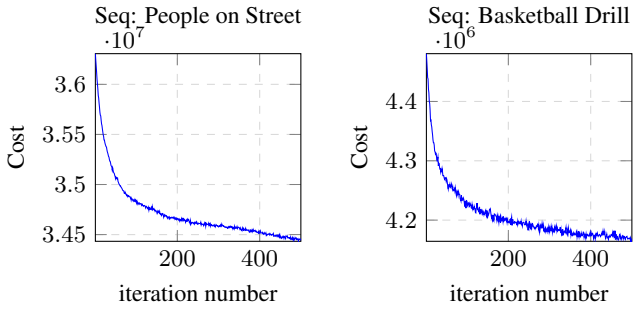

Fig. 1: Cost Convergence vs Iteration

Equation (2) is solved in similar way as described in [14]. Note that the new transform are imposed to be orthonormal. After the new set $\mathbf{T}'_k$ of transforms is obtained, the classification using equation (1) is performed again by replacing $\mathbf{T}_k$ by $\mathbf{T}'_k$. This defines an iterative process that converges or is stopped after a certain number of iterations. Figure 1 shows the convergence of the rate distortion cost for two different sequences during learning phase.

As our method utilize the HEVC rate of encoding the coefficients for transform optimization, it is different from the algorithms developed in [11] and [14] where, the transforms are learned offline and the rate of coefficient coding is approximated by $\ell 0$-norm.

The above algorithm is implemented in the latest HEVC test software (version HM15.0) in order to generate a set of adaptive transforms. A total of $K = 4$ transforms ($\mathbf{T}_k^{8\times8}$) of size 64×64 and $L = 4$ transforms ($\mathbf{T}_l^{4\times4}$) of size 16×16 are used as initial directional non-separable transforms aligned in the directions ranging from 0° to 180°. These transforms are expected to model the directionality of the residual signal and therefore, can be used for classification of the residuals with similar statistics into separate classes.

The HM encoder is modified such that first, the directional transforms of size 64×64 and 16×16 (denoted as non-DCTs in rest of the paper) are learned on 8×8 and 4×4 size residual blocks respectively of the first frame of each sequence using the above described iterative algorithm. Then, the final optimized transforms are tested in brute-force fashion on each of the residual of size 8×8 and 4×4 and the best candidate transform is indexed to the decoder using an additional syntax. The changes are applied only to the luma component. The chroma components are transformed using the conventional DCT. As the non-DCTs concentrate the coefficient energy in decreasing order on an average from the first coefficient to last, the coefficients obtained from non-DCTs are reordered such that they are always scanned in the horizontal direction using HEVC adaptive scanning order.

In addition to the best candidate transform index, the basis vectors of non-DCTs need to be transmitted to the decoder for successful decoding. The methods used to handle the side information is detailed in the next two sub-sections.

Associated Syntax Encoding

The implementation on the encoder side involves selection of best transform candidate for each residual blocks of size 8×8 and 4×4. This makes it mandatory to encode the transform direction index flag for each transform block having non-zero values. A 1-bit flag is encoded to indicate the usage of DCT or non-DCT for a particular residual block. In case a non-DCT flag is set, the oriented transform used is indexed using additional $\log_2(K)$ bits where K is the number of non-DCT transforms. A context is attached to the flag indicating DCT and non-DCT.

As the usage percentage of the non-DCTs is not uniform across sequences, a probability table ($\mathbf{p}_k$) is learned starting with an equi-probable initial value. Each element of $\mathbf{p}_k$ defines the number of occurrences of the index in a sequence. At each iteration, the $\mathbf{p}_k$ is updated and the rate of transform index can be estimated as

$$R_T = -\log_2(\mathbf{p}_k)$$

This table is used in the next iteration for selection of best transform candidates inside the rate-distortion search operation of HM encoder. This approach is advantageous in rejecting the transforms which are not adaptive to the content and are not used at all during the optimization process. This also saves the overhead of encoding the transform itself. The downside of this approach is that the less used transforms are penalized with high index cost right from first iteration and this results in faster rejection of less used transforms which might not lead to optimal results. In order to avoid that, an annealing parameter, ϵ is introduced inside the cost function, C_k as shown in equation 3.

$$C_k = \|\mathbf{r}^j - \mathbf{T}_k\mathbf{Q}^{-1}(\mathbf{c}_k^j)\|_2^2 + \lambda(R(\mathbf{c}_k^j) + \epsilon R_T) \quad (3)$$

The value of ϵ is increased gradually from an initial value 0 to 1 at each iteration. At start, the rate of index coding, R_T is strongly penalized by annealing parameter, ϵ in order to allow the transforms to learn without constraint of index coding and gradually, this rate term is introduced into the optimization equation 3.

Transform Coding

The algorithm introduced in this paper requires transmitting the newly learned transform basis vectors on the decoder side. This accounts for an additional overhead which affect the final performance gain achieved by utilizing these adaptive transforms. In a scenario where all the basis vectors of a non-DCT transform need to be encoded, the overhead of transmitting K 64×64 transforms and L 16×16 transforms at precision of 10 bits can be computed as $16 \times 16 \times 10 \times L + 64 \times 64 \times 10 \times K$. For K = L = 4, this is estimated to be around 170K bits.

This overhead can be significantly reduced by sending only few basis vectors to the decoder side especially at high QP where after quantization, most of the high frequency coefficients are zero. The set of transforms can be generated from these few basis vectors using completion algorithm as described in [15]. Moreover, the precision of the transform basis can be reduced at low bit-rates where the loss in precision has a minute affect on the overall distortion on the decoder side. As described in section 2, the probability table learned during the learning process provides an indication of most used transforms for a particular sequence and QP. Therefore, this overhead can be reduced further by dropping the transforms that are marginally used.

The coding of basis vectors has not been addressed in this paper and will be addressed in future works. However, we have added the estimated overhead cost of 170K bits while presenting the results.

3. RESULTS

In this section, results computed on HEVC test sequences are presented from the experiments conducted for two different test cases.

In the first test, the algorithm described in section 2 is applied to the first frame of a sequence where, the number of iterations is set to 100 and number of 64×64 and 16×16 transforms are set to 4. The final optimized transforms set is tested on the first frame. Table 1 shows the result for Adaptive Directional Transforms (ADT) and Annealing based Adaptive Directional Transforms (AADT). In case of ADT, the cost of transform index (R_T) in equation 1 is kept zero during the learning process in order to avoid any implicit restriction on transform learning. For AADT, annealing parameter as described in section 2 is applied to the index cost (3) where the value of ϵ is increased from 0 to 1 uniformly for 100 iterations. Results in Table 1 show that a very simple annealing strategy can provides better gain (around 1.4% in this case) which proves the merits of jointly optimizing the coefficient encoding cost and the index encoding cost. Even with annealing, the gains obtained using this learning algorithm are not enough to compensate for the overhead of transmitting the basis vectors to the decoder. This overhead has less impact on the bit-rate of high resolution sequences whereas, for low resolution sequences, the overhead can be significant portion of the bit-rate.

In the second test, the learned transforms obtained using the proposed algorithm are tested on HEVC test sequences in random access (RA) coding mode where the GOP size is set to 8 and intra frame is encoded every half a second. The overhead cost of transform coding is reduced considerably as the same set of transforms is used for the whole sequence. Table 2 illustrates the gain in case of RA coding mode. Additionally, the BD-rate gain obtained on intra-only frames is shown in the table 2 under all intra (AI). Improvement is observed for almost every test sequence. The overall performance gain on intra frames (AI) is more compared to the gain

Class	Sequence	ADT	AADT
A	PeopleOnStreet	-3.07	-4.28
	Traffic	-2.62	-3.43
BD Rate		-2.85	-3.85
B	BasketballDrive	-2.03	-2.99
	BQTerrace	-1.81	-2.68
	Cactus	-1.83	-3.12
	ParkScene	-2.14	-3.68
BD Rate		-1.95	-3.12
C	BasketballDrill	-8.16	-10.33
	BQMall	-2.62	-3.60
	PartyScene	-3.11	-4.79
	RaceHorses	-2.51	-5.58
BD Rate		-4.10	-6.08
E	FourPeople	-2.43	-3.78
	Johnny	-1.15	-2.55
	KristenAndSara	-2.13	-3.36
BD Rate		-1.90	-3.23
Total		-2.70	-4.07

Table 1: BD rate gain on first frame for ADT and AADT

Class	Sequence	AADT (RA)	AADT (AI)
A	PeopleOnStreet	-1.10	-3.65
	Traffic	-1.83	-2.70
BD Rate		-1.46	-3.18
B	BasketballDrive	0.29	-0.15
	BQTerrace	-1.01	-1.40
	Cactus	-0.63	-1.55
	ParkScene	-0.92	-1.94
BD Rate		-0.57	-1.26
C	BasketballDrill	-4.72	-7.54
	BQMall	-0.38	-0.52
	PartyScene	-1.21	-1.56
	RaceHorses	0.05	-0.87
BD Rate		-1.57	-2.62
E	FourPeople	-2.20	-2.24
	Johnny	-1.58	-1.16
	KristenAndSara	-2.11	-1.91
BD Rate		-1.96	-1.32
Total		-1.40	-2.10

Table 2: BD rate gain with overhead bits for RA and AI mode

in RA coding mode. Also, it is observed that the sequences with high directional structure (e.g. Basketball Drill) provide better performance. This proves that by carefully adapting to the content of a sequence using proposed learning approach, much better BD-rate gains are obtained. Figure 2 illustrates the areas of the sequence Basketball Drill at QP 22 where non-DCTs are chosen in place of conventional DCT/DST. It clearly shows that the non-DCTs are extensively chosen across the whole frame. Table 3 shows the usage statistics of non-DCTs for class C sequences. Moreover, it is observed that for sequences with fast motion, performance drops comparatively faster when the set of adaptive transforms are applied on the subsequent intra frames of a sequence. For illustration, the BD-rate gain on intra-frames of two different sequences is shown in Figure 3.

Fig. 2: Usage of conventional DCT/DST (red) and non-DCTs (yellow) for sequence Basketball drill (QP 22)

BaskeballDrill		BQMall		PartyScene		RaceHorses	
4×4	8×8	4×4	8×8	4×4	8×8	4×4	8×8
65%	75%	58%	20%	63%	8%	72%	26%

Table 3: Usage Statistics of non-DCTs

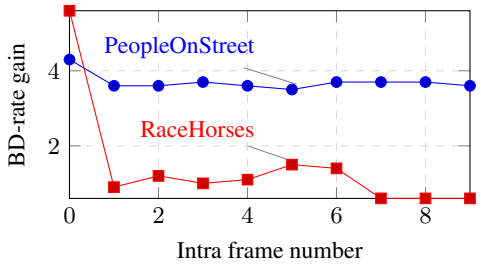

Fig. 3: BD-rate gain at each intra frame

4. SUMMARY AND CONCLUSIONS

A framework proposed in this paper exploits the residual signal statistics in order to adapt the transforms to the content using an on-the-fly block based learning algorithm. It is observed that these new optimized transforms obtained after the classification of residuals based on the encoding cost can provide considerable gain in terms of the BD-rate.

For the optimization, a set of non-separable 2D transforms are used and the true rate of HEVC codec is being used to impose a rate constrain instead of imposing sparsity constraint as done extensively in the literature. The annealing scheme improves the stability of the iterative algorithm and its performance. More efficient annealing schemes may be found to obtain even better convergence and RD performance. However, the proposed method is computationally demanding and requires coding of side information at encoder side. The decoder side complexity remains comparable to HEVC.

The work further may be carried in reducing the side information and at the same time finding a low complexity framework with similar or better gains. Further gains may be drawn by employing better initialization of transform bases and dynamically varying the number of transforms.

5. REFERENCES

[1] C.W. Kok, "Fast algorithm for computing discrete cosine transform," *IEEE Transactions on Signal Processing*, pp. 757–760, 1997.

[2] Bing Zeng and Jingjing Fu, "Directional discrete cosine transforms-a new framework for image coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 3, pp. 305–313, 2008.

[3] Robert A Cohen, Sven Klomp, Anthony Vetro, and Huifang Sun, "Direction-adaptive transforms for coding prediction residuals," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 185–188.

[4] Yan Ye and Marta Karczewicz, "Improved h. 264 intra coding based on bi-directional intra prediction, directional transform, and adaptive coefficient scanning," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 2116–2119.

[5] Moyuresh Biswas, Mark R Pickering, and Michael R Frater, "Improved h. 264-based video coding using an adaptive transform," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 165–168.

[6] Miaohui Wang, King Ngi Ngan, and Long Xu, "Efficient h. 264/avc video coding with adaptive transforms," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 16, no. 4, pp. 933, 2014.

[7] Feng Zou, Oscar C Au, Chao Pang, and Jingjing Dai, "Rate distortion optimized transform for intra block coding for hevc," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*. IEEE, 2011, pp. 1–4.

[8] K Sühring, G Heising, D Marpe, et al., "Kta (2.6 r1) reference software, 2009," .

[9] Joel Sole, Peng Yin, Yunfei Zheng, and Cristina Gomila, "Joint sparsity-based optimization of a set of orthonormal 2-d separable block transforms," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 9–12.

[10] Osman Gokhan Sezer, Robert Cohen, and Anthony Vetro, "Robust learning of 2-d separable transforms for next-generation video coding," in *Data Compression Conference (DCC), 2011*. IEEE, 2011, pp. 63–72.

[11] Xin Zhao, Li Zhang, Siwei Ma, and Wen Gao, "Video coding with rate-distortion optimized transform," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 1, pp. 138–151, 2012.

[12] Long Xu, King Ngi Ngan, and Miaohui Wang, "Video content dependent directional transform for intra frame coding," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 197–200.

[13] Siwei Ma, Shiqi Wang, Qin Yu, Junjun Si, and Wen Gao, "Mode dependent coding tools for video coding," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 6, pp. 990–1000, 2013.

[14] Osman Gokhan Sezer, Oztan Harmanci, and Onur G Guleryuz, "Sparse orthonormal transforms for image compression," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 149–152.

[15] Ivan W Selesnick and Onur G Guleryuz, "A diagonally-oriented dct-like 2d block transform," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2011, pp. 81381R–81381R.