VIDEO CODING USING MULTI-REFERENCE MOTION-ADAPTIVE TRANSFORMS BASED ON GRAPHS

Du Liu and Markus Flierl

KTH Royal Institute of Technology
Stockholm, SE 10044, Sweden
{dul, mflierl}@kth.se

ABSTRACT

The purpose of the work is to produce jointly coded frames for efficient video coding. We use motion-adaptive transforms in the temporal domain to generate the temporal subbands. The motion information is used to form graphs for transform construction. In our previous work, the motion-adaptive transform allows only one reference pixel to be the lowband coefficient. In this paper, we extend the motion-adaptive transform such that it permits multiple references and produces multiple lowband coefficients, which can be used in the case of bidirectional or multihypothesis motion estimation. The multi-reference motion-adaptive transform (MRMAT) is always orthonormal, thus, the energy is preserved by the transform. We compare MRMAT and the motion-compensated orthogonal transform (MCOT) [1], while HEVC intra coding is used to encode the temporal subbands. The experimental results show that MRMAT outperforms MCOT by about 0.6dB.

Index Terms— Motion-adaptive transform, motion-inherited graph, subspace constraint

1. INTRODUCTION

Transforms are widely used in today's data compression techniques such as High Efficiency Video Coding (HEVC) as an important tool for signal decorrelation [2]. For videos with both spatial and temporal correlation, the transforms, especially the discrete cosine transform (DCT), are commonly applied to the spatial domain to reduce the spatial redundancy. For temporal correlation, the standard approach is to use motion-compensated prediction.

Instead of using closed-loop prediction, we focus on temporal transforms that operate in an open-loop fashion. Due to the complexity of motion fields, the well-known motion-compensated lifting [3] struggles with unconnected, connected, and multi-connected pixels when performing the update step. To address this shortcoming, [4] and [5] propose modified update operators to be used in the lifting scheme. On the other hand, the class of motion-compensated orthogonal transforms (MCOT) does not use the lifting scheme. It is designed to process a sequence of pictures in a hierarchical way, while maintaining strict orthogonality for any motion field [1, 6].

To further investigate the temporal orthonormal transforms in energy compaction and coding, the idea of graph-based signal processing is helpful. For example, the graph can be used in multiresolution signal analysis [7–9]. It is also shown that the eigenvector matrix of a graph matrix is able to capture different frequencies of graph signals, e.g., [10–12]. For compression, a lifting scheme based on a graph for video coding is considered in [13] and the graph Laplacian eigenbasis is used in depth video coding [14]. Both techniques give promising results by considering the graph information. In addition, the work in [15] provides an analysis for the optimal case of using the graph Laplacian eigenbasis.

For our work, the compression performance is relevant. The transforms in [16] are constructed using the Laplacian eigenbasis of vertex-weighted graphs. The transform using the graph information gives an improved energy compaction result. For our motion-adaptive transforms, each transform outputs an energy-compacted lowband coefficient and a number of highband coefficients. However, this is limiting, since in practice there can be more than one reference pixel as in bidirectional and multihypothesis motion estimation. In this paper, we aim at extending the motion-adaptive transform such that it creates multiple lowband coefficients with the help of graphs.

The paper is organized as follows: Sec. 2 summarizes the motion-adaptive transforms with single reference. Sec. 3 constructs the multi-reference motion-adaptive transforms. Sec. 4 provides the experimental results.

2. MOTION-ADAPTIVE TRANSFORM

2.1. Scale Factors Accommodating Energy Compaction

The scale factors are used to track the energy compaction of lowband coefficients under the assumption of ideal motion

This work has been supported in part by the Swedish Research Council under the grant 2011-5841.

978-1-5090-0746-2/16/$31.00

[1]. Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ be a vector of intensity values connected by motion estimation. Ideal motion assumes that $x_1 = x_2 = \cdots = x_n = x'$, i.e., these n pixels are with equal intensity x'. Let T be an $n \times n$ transform matrix, and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ the output. We have

$$\mathbf{y} = T^T \mathbf{x}, \tag{1}$$

where y_1 is considered to be the lowband coefficient and $y_2, \ldots, y_n$ the highband coefficients.

In the following, a small example with two coefficients and a Haar transform is used to illustrate the use of scale factors. If we compact the energy of x_1 and x_2 into one lowband coefficient y_1, i.e.,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{2}x' \\ 0 \end{bmatrix}, \tag{2}$$

the output of the lowband coefficient $y_1 = \sqrt{2}x'$ becomes a scaled x' with a factor $\sqrt{2}$.

Let the scale counter m_i $(i = 1, \ldots, n)$ be the number of pixels that are compacted to the ith lowband coefficients. The original intensity values $x_1, \ldots, x_n$ have scale counters of zero. The scale factor c_i is determined by $c_i = \sqrt{m_i + 1}$, representing the compacted energy of the ith lowband coefficient. The scale factors of the original intensities are all ones, since the original scale counters are all zeros.

From the example in (2), the scale counter of y_1 is 1 as one pixel energy is compacted to y_1, and no scale counter is associated with y_2 as y_2 is a highband coefficient. The scale factor of y_1 is $\sqrt{2}$, which is the same as the factor of x' in (2). Similarly, if we compact the energy of n pixels of $\mathbf{x}$ to one lowband coefficient, the scale counter of this lowband coefficient will be $n - 1$, and the corresponding scale factor is $\sqrt{n}$. The scale counters and the scale factors are only determined by the motion information. They do not require extra information to be encoded.

2.2. Subspace Constraint for Motion-Adaptive Transform

Now, let us consider $\mathbf{x}$ as a vector consisting of n lowband coefficients connected under ideal motion assumption. These n lowband coefficients can be expressed by an original intensity x' with n scale factors $c_1, \ldots, c_n$, respectively, i.e., $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T = [c_1 x', c_2 x', \ldots, c_n x']^T = x'\mathbf{c}$, where $\mathbf{c} = [c_1, c_2, \ldots, c_n]^T$ is the vector of scale factors.

We construct an orthonormal transform matrix T that perfectly compacts the energy of $\mathbf{x}$ to a lowband coefficient. Let $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$ be the basis vectors of T. Using (1), we have

$$y_i = x'\mathbf{t}_i^T\mathbf{c}, \text{ for } i = 1, \ldots, n. \tag{3}$$

The lowband coefficient $y_1 = x'\mathbf{t}_1^T\mathbf{c}$ is designed to capture the total energy of the signal $\mathbf{x}$, thus, $\mathbf{t}_1$ needs to be collinear with $\mathbf{c}$,

$$\mathbf{t}_1 = \frac{\mathbf{c}}{\|\mathbf{c}\|_2}. \tag{4}$$

Then, $y_1 = x'\sqrt{\mathbf{c}^T\mathbf{c}}$ contains the total energy of $\mathbf{x}$. Since $\mathbf{t}_1$ represents one dimension in the n-dimensional space, and all the other basis vectors $\mathbf{t}_2, \ldots, \mathbf{t}_n$ are orthogonal to $\mathbf{t}_1$, the highband coefficients $y_2, \ldots, y_n$ are all zeros. With this, the transform T is able to compact the energy perfectly. We refer to the constraint of $\mathbf{t}_1$ in (4) as the subspace constraint [17].

2.3. Motion-Adaptive Transforms

As the basis vector $\mathbf{t}_1$ of the transform is determined according to the scale factors, the remaining $n - 1$ basis vectors are left to be constructed. Different sets of $n - 1$ basis vectors lead to different highband coefficients. The following summarizes the construction of three transforms proposed in previous work.

The motion-compensated orthogonal transform (MCOT) is a Haar-like transform constructed from a sequence of rotation matrices H_i [1]. Each rotation matrix H_i compacts the energy of two coefficients into one lowband coefficient and one highband coefficient. The transform matrix is the product of these rotation matrices, and $n - 1$ rotation matrices are needed to process n coefficient, i.e., $T_{MCOT} = (H_1 H_2 \cdots H_{n-1})^T$.

The second transform is the DCT-based rotation (DBR), which is constructed by rotating the DCT basis such that the first basis vector $\mathbf{b}_1 = \frac{1}{\sqrt{n}}[1, \ldots, 1]^T$ of the DCT meets the subspace constraint [17]. As $\mathbf{b}_1$ and $\mathbf{t}_1$ span a plane and determine a rotation from $\mathbf{b}_1$ to $\mathbf{t}_1$, the higher order basis vectors of DCT can be rotated in the same direction and hold the relative positions.

The third transform is obtained from the eigenvector matrix of the Laplacian of a vertex-weighted graph (VWL) [16]. The graph is inherited from the motion information. The weights on the graph are given by the scale factors, such that the Laplacian eigenbasis has one eigenvector that is the same as the subspace constraint $\mathbf{t}_1$. This transform requires the eigen-decomposition of a vertex-weighted graph. However, the advantage is that it incorporates the underlying motion structure into the transform. The DBR can be viewed as an approximation of VWL, as the DBR does not require eigen-decomposition, and as the energy compaction is close to that of the VWL.

Since all the proposed transforms share the same the subspace constraint $\mathbf{t}_1$, the corresponding lowband coefficients produced by these transforms are equal. However, the highband coefficients differ for these transforms.

3. MULTI-REFERENCE MOTION-ADAPTIVE TRANSFORM (MRMAT)

In this section, we first construct the multi-reference motion-adaptive transform (MRMAT). Then, we introduce the process of applying MRMAT on a given graph.

3.1. Construction of MRMAT

Each transform discussed in Sec. 2.3 compacts the energy to only one lowband coefficient, and only this lowband coefficient can be used again in other transforms for energy compaction. This limits the case of using the transform if there are multiple coefficients to be used in another transform or if there are multiple references for motion estimation, e.g., bidirectional or multihypothesis motion estimation. In the following, we construct the transform that allows multiple references and create multiple lowband coefficients.

The main concept of creating multiple lowband coefficients is to first, compact the energy of the input signal to one coefficient, and second, redistribute the energy from one coefficient to multiple coefficients equally. The energy should be conserved in general, thus, the transforms from the two steps need to be orthonormal.

Again, we let $\mathbf{x}$ be the n-dimensional input vector, and $\mathbf{y}$ the output vector of the energy compacting transform with y_1 as the lowband coefficient. Assume there are k $(1 \leq k < n)$ energy-redistributed coefficients $\widetilde{\mathbf{x}}_k = [\widetilde{x}_1, \ldots, \widetilde{x}_k]$. Let U_k be the transform used in energy redistribution, we consider $\widetilde{\mathbf{x}}_k = U_k^T \mathbf{y}_k$, where $\mathbf{y}_k$ denotes the first k elements of $\mathbf{y}$.

From (1), the inverse process of energy compaction is given by $\mathbf{x} = T^{T^{-1}}\mathbf{y} = T\mathbf{y}$ as T is orthonormal. This inverse process can be viewed as redistributing the energy back to n coefficients. Using the same idea, we simply let $U_k^T = T_k$ to redistribute the energy from one to k coefficients, where T_k is the transform that compacts k input coefficients into one lowband cofficient. Thus, we have

$$\widetilde{\mathbf{x}}_k = T_k \mathbf{y}_k. \tag{5}$$

Since T satisfies the subspace constraint determined by $\mathbf{c}$, similarly, the first basis vector of T_k needs to satisfy the subspace constraint determined by the scale factors of $\widetilde{\mathbf{x}}_k$. The full matrix of T or T_k can be constructed using the transforms introduced in Sec. 2.3.

Now, we need to compute the scale factors of $\widetilde{\mathbf{x}}_k$. To redistribute the lowband energy equally to the k coefficients, these scale factors need to be updated equally. Let $\mathbf{m} = [m_1, \ldots, m_n]^T$ and $\mathbf{c} = [c_1, \ldots, c_n]^T$ be the scale counters and scale factors associated with $\mathbf{x}$, respectively. Let $\widetilde{\mathbf{m}}_k = [\widetilde{m}_1, \ldots, \widetilde{m}_k]^T$ and $\widetilde{\mathbf{c}}_k = [\widetilde{c}_1, \ldots, \widetilde{c}_k]^T$ be the scale counters and scale factors associated with $\widetilde{\mathbf{x}}$, respectively. The scale counters consider the energy shifted from $x_{k+1}, \ldots, x_n$ to $\widetilde{x}_1, \ldots, \widetilde{x}_k$. The update of scale counters is then given by

$$\widetilde{m}_i = m_i + \frac{1}{k}\left(n - k + \sum_{j=k+1}^{n} m_j\right), \text{ for } i = 1, \ldots, k, \tag{6}$$

and the update of scale factors follows $\widetilde{c}_i = \sqrt{\widetilde{m}_i + 1}$.

In conclusion, in the first step in (1), T^T is determined by $\mathbf{c}$ to compact energy. The highband coefficients $y_{k+1}, \ldots, y_n$ are obtained by T^T. In the second step in (5), T_k is determined by $\widetilde{\mathbf{c}}_k$ to redistribute energy, and the energy from y_1 is distributed to references $\widetilde{x}_1$ to $\widetilde{x}_k$ for further processing.

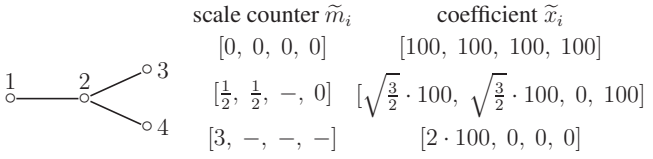

Fig. 1. An example of a tree structure for the multi-reference motion-adaptive transform. The scale counters $\widetilde{m}_i$ $(i = 1, \ldots, 4)$ and the coefficients $\widetilde{x}_i$ are given after each step. The initial $\widetilde{m}_i$ are zeros and the initial $\widetilde{x}_i$ are 100s under ideal motion assumption. The first MRMAT operates on three coefficients 1, 2, and 3. The scale counters $\widetilde{m}_1$ and $\widetilde{m}_2$ are updated according to (6), and the compacted energy in $\widetilde{x}_i$ is represented as $\sqrt{\widetilde{m}_i + 1} \cdot 100$. No scale counter is associated with the 3rd coefficient as it becomes a highband coefficient. The second MRMAT operates on coefficients 1, 2, and 4. The energy is finally compacted to the 1st coefficient, and $\widetilde{m}_1$ is updated to 3.

3.2. Process of MRMAT

The graph inherited from the motion information is always in a tree structure, where the root is the reference in the first frame of each group of pictures (GOP), and the connection of two nodes is determined by motion estimation. We apply MRMAT along each path in the motion inherited graph.

For example, Fig. 1 depicts an example of a motion-connected tree. The 1st node is connected to one node, meaning that the first pixel is used as motion reference for the second pixel. Similarly, the second pixel is used as reference for the third and fourth pixel. In the first transform, the 1st, 2nd, and 3th pixel from three frames are processed as they lie on the same path. The energy of these three pixels is distributed to the 1st and the 2nd coefficients using MRMAT. Then, the 1st and the 2nd pixel are used again in the next transform. In this second transform, the 1st, 2nd, and 4th coefficient that lie on the second path are processed and produce one lowband coefficient for the 1st node. As a result of this sequence of transforms, we obtain one lowband coefficient and three highband coefficients, and the size of each transform matrix is determined by the number of coefficients that lie on the current path.

This successive process is different from the process we performed in previous work, where all motion-connected coefficients have been transformed only once. Transforming all the coefficients in one step may lead to discontinuities among neighboring coefficients in one subband, in particular, if the neighboring output coefficients are captured by basis vectors with different frequency properties. The successive process of MRMAT is able to reduce such discontinuities.

4. EXPERIMENTAL RESULTS

In the experiments, we evaluate the coding performance for the QCIF sequences *Foreman* and *Bus*, each with 128 frames. We compare the Motion-Compensated Orthogonal Transform (MCOT) [1], the vertex-weighted Laplacian (VWL) [16], and the multi-reference motion-adaptive transform (MRMAT).

The transform is applied along temporal direction. The MCOT is a product of a sequence of Haar-like transforms. The MRMAT is a process including energy compaction and energy redistribution. The transform matrices constructed for MRMAT use the vertex-weighted Laplacian (VWL) as introduced in Sec. 2.3, as it gives a better energy compaction when compared to the other two transforms in previous work [16].

The transform does not limit the block size in motion estimation or the GOP size, as the graph can represent a general motion structure. The graphs are defined by 16×16 block motion with a search range of ± 64. The GOP size is set to eight, thus, after the temporal transform, we obtain one temporal lowband and seven temporal highbands. The compared transforms use the same set of motion vectors, thus, the comparison is targeting the efficient coding of the subbands.

Since the lowband is energy compacted, the coefficients have a large range. We scale down the lowband coefficients by dividing them with the corresponding scale factors. The temporal subbands are then coded using HEVC model HM16.7 with intra coding [18]. The encoder uses the Main 10 profile due to the possibility that the temporal subbands may have a bit depth larger than eight bits. Other encoding parameters remain the same as recommended in the HEVC configuration file. We measure the peak signal to noise ratio (PSNR) of the luma (Y) signal. The U and V components are set to a constant value before encoding.

The rate allocation needs to be determined among the subbands. To find the optimal rate allocation, we encode every subband with all possible quantization parameters (QP). For the kth subband, we choose the parameter that minimizes the cost J_k,

$$J_k^* = \min_i \left(D_{k,i} + \lambda R_{k,i} \right), \tag{7}$$

for a given λ, where $D_{k,i}$ is the distortion of subband k with ith QP and $R_{k,i}$ the rate of the corresponding subband and QP.

Figs. 2 and 3 depict the luma PSNR vs. bitrate performance for *Foreman* and *Bus*. Since the transforms are orthonormal, the distortion is not amplified and it can be measured without performing the inverse transform. In Fig. 2, MRMAT outperforms MCOT by about 0.6dB, and VWL by about 0.2dB. We observe in both figures that the proposed MRMAT outperforms the other two. That is, the energy compaction and redistribution steps allow an improved compression of subbands.

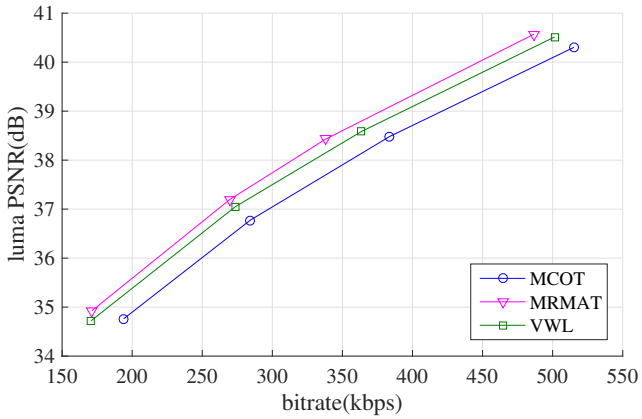

Fig. 2. Comparison of PSNR vs. rate for QCIF *Foreman* with 128 frames at 30fps.

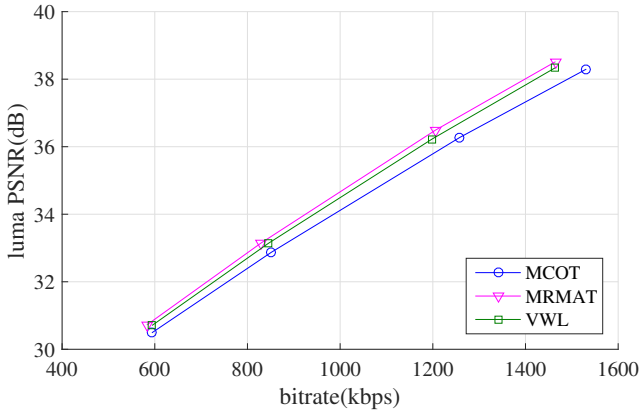

Fig. 3. Comparison of PSNR vs. rate for QCIF *Bus* with 128 frames at 30fps.

5. CONCLUSIONS

Aiming at extending the motion-adaptive transforms from producing one lowband coefficient to multiple lowband coefficients, we proposed a multi-reference motion adaptive transform in this paper. The new motion-adaptive transform incorporates the scale factors into the construction of the transform. The main concept of the proposed MRMAT includes energy compaction and energy redistribution. The energy compaction step uses a motion-adaptive transform to compact the energy to one lowband coefficient. The energy redistribution step then distributes this energy using the transpose of a dimension-reduced motion-adaptive transform. The MRMAT is applied successively along the paths of a given graph, thus, the structure of the motion-inherited graph is considered during the transform. In the experiments, MRMAT is used as temporal transform and the temporal subbands are encoded using HEVC intra coding. The experimental results show that the proposed MRMAT outperforms MCOT by about 0.6dB.

6. REFERENCES

[1] M. Flierl and B. Girod, "A motion-compensated orthogonal transform with energy-concentration constraint," in *Proc. IEEE International Workshop on Multimedia Signal Processing*, Oct. 2006, pp. 391–394.

[2] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[3] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 3, pp. 1793–1796.

[4] C. Tillier, B. Pesquet-Popescu, and M. van der Schaar, "Improved update operators for lifting-based motion-compensated temporal filtering," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 146–149, 2005.

[5] B. Girod and S. Han, "Optimum update for motion-compensated lifting," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 150–153, 2005.

[6] M. Flierl and B. Girod, "Half-pel accurate motion-compensated orthogonal video transforms," in *Proc. IEEE Data Compression Conference*, Mar. 2007, pp. 13–22.

[7] M. Gavish, B. Nadler, and R. R. Coifman, "Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning," in *Proc. International Conference on Machine Learning*, 2010, pp. 367–374.

[8] V. N. Ekambaram, G. C. Fanti, B. Ayazifar, and K. Ramchandran, "Multiresolution graph signal processing via circulant structures," in *Proc. IEEE Digital Signal Processing and Signal Processing Education Meeting*, 2013, pp. 112–117.

[9] S. K. Narang and A. Ortega, "Perfect reconstruction two-channel wavelet filter banks for graph structured data," *IEEE Trans. on Signal Processing*, vol. 60, no. 6, pp. 2786–2799, 2012.

[10] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[11] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[12] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. on Signal Processing*, vol. 62, no. 12, pp. 3042–3054, June 2014.

[13] E. Martínez-Enríquez, F. Díaz-de-María, and A. Ortega, "Video encoder based on lifting transforms on graphs," in *Proc. IEEE International Conference on Image Processing*, Sept. 2011, pp. 3509–3512.

[14] W.-S. Kim, S. K. Narang, and A. Ortega, "Graph based transforms for depth video coding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012, pp. 813–816.

[15] C. Zhang and D. Florêncio, "Analyzing the optimality of predictive transform coding using graph-based models," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 106–109, Jan. 2013.

[16] D. Liu and M. Flierl, "Motion-adaptive transforms based on the Laplacian of vertex-weighted graphs," in *Proc. IEEE Data Compression Conference*, 2014, pp. 53–62.

[17] D. Liu and M. Flierl, "Graph-based rotation of the DCT basis for motion-adaptive transforms," in *Proc. IEEE International Conference on Image Processing*, Sept. 2013, pp. 1802–1805.

[18] *JCT-VC HEVC reference software version HM 16.7*, available at `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.7/`.