MOTION HINTS COMPENSATED PREDICTION AS A REFERENCE FRAME FOR HIGH EFFICIENCY VIDEO CODING (HEVC)

Ashek Ahmmed[1,2]*, Miska M. Hannuksela*[2]*, and Moncef Gabbouj*[1]

[1] Department of Signal Processing, Tampere University of Technology, Tampere, Finland.
[2] Nokia Technologies, Tampere, Finland.

ABSTRACT

Segment-based temporal prediction combined with higher-order motion models have been studied as an alternative to conventional block-based translational inter prediction. One example of such studies is known as motion hints, where an affine motion model has been used. In this paper, we explore the applicability of motion hints with an elastic motion model in generating reference frames for conventionally coded B-frames. The presented design enables the re-use of existing codecs, such as HEVC, without modifications in low-level coding tools. Experimental results show that a savings in bit rate of up to 5.1% is achievable over standalone HEVC with increased computational complexity.

Index Terms— Motion hint, HEVC, elastic motion model, video coding.

I. INTRODUCTION

Conventional video coding paradigms employ block-based translational motion models. In such models, neighboring pixels are grouped together into square or rectangular blocks to form an artificial partitioning of the current frame i.e. the frame to be predicted. For each of these blocks, the motion modelling step tries to come up with an identical shape block, known as the predicted block, that closely resembles the target block by performing a search in the set of already coded frames, known as the reference frames. The greater the match between the current block and its prediction, the smaller the distortion is.

In the traditional setting of block-based translational motion model, all the pixels belonging to a block are assigned a single motion vector. This uniformity of motion within a block assumption is highly effective in minimizing the bit rate required to code the overall motion data associated to a frame and thereby provides a rate-distortion (RD) trade off; however it fails to generalize in blocks which are on moving object boundaries i.e. where motion discontinuities exist.

Experimental results and the latest standards have shown that by shifting from fixed sized blocks to variable sized blocks the coding inefficiency, incurred due to the failure of the block-based translational model to efficiently model the actual locations of discontinuities in the motion field, can be mitigated to some extent. As an example, several types of block partitions from 4×4 to 16×16 pixels are supported by the H.264/AVC standard [1] and the more recent standard the HEVC [2] allows a range of symmetric and asymmetric partitions with the maximum block size can go up as far as 64×64 samples. Carefully partitioning motion blocks with object boundaries, it is possible to segment the motion vector field into disjoint regions with each block possessing a smooth (typically constant) motion model [3–5].

With the aid of segmentation it is possible to obtain a more flexible partitioning. Milani *et al.* [6] segmented the current frame into irregularly-shaped regions by estimating them from the previously-coded frame. A predicting frame is generated by combining the best predictors for each region which are to be found among the previously-decoded pictures. This segmentation-based video coding system has shown to outperform the RD performance of H.264/AVC of 2 dB.

In this paper, we chose to employ the motion hints based video coding paradigm developed in [7, 8]. The innovative idea of motion hint is pioneered by Naman *et al.* in [9] and its applicability in generating inter-frame predictions with good subjective quality and high peak signal-to-noise ratio (PSNR) is shown in [10]. Motion hint provides a global description of motion over specific domains and is related to the foreground-background segmentation where the foreground and background motions are the hints. The inspiration behind motion hint is that while existing approaches of [6, 11, 12] segment the current frame, the motion hints based approach segment the reference frames and thereby avoids ascribing motion vectors to regions where content in the current frame is occluded in a reference frame. Forming the segmentation within the reference frames also provides the flexibility to use information from additional frames or even other media types e.g. depth maps as evidence within the segmentation estimation step.

With motion hints, there is no need to use the motion model to describe the object boundaries because the spatial structure of the already decoded reference frames can be exploited to infer appropriate boundaries for the intermediate ones. The *motion hints mode* using an affine motion model, is proposed for coding bi-predictive slice macroblocks (MBs)

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

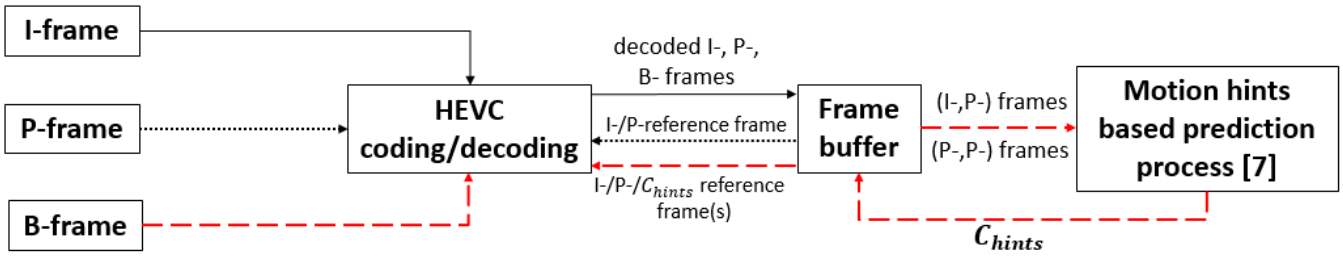

Fig. 1. Block diagram of the coding/decoding framework that uses the motion hints based prediction [7] as a reference frame, along with the usual temporal reference(s), for the B-frames.

and incorporated into the JM reference software [13] for H.264/AVC in [14]. Experimental results showed that the hybrid coder coded more than 50% of the motion discontinuity MBs using the *motion hints mode* in low bit rate cases and that coder achieved a RD gain of 1.11 dB, or equivalently 17.05% savings in bit rate, over the true JM coder that does not use the *motion hints mode*, when both low and high bit rate scenarios are considered.

An elastic motion model equipped with 2-D discrete cosine basis functions was used by Pickering *et al.* [15], to describe complex motion vector fields. Although there are alternative basis functions exist e.g. B-splines, polynomials, harmonic functions, radial basis functions, wavelets, they chose discrete cosines because of their ability to represent smooth motion fields in a sparse way. It was shown in [16] that this elastic motion model can outperform the affine model in compensating complex deformed motion.

This paper introduces two major contributions compared to the earlier works in [7, 8, 14]: *(i)* we investigate the RD performance of motion hint based coding relative to the latest video coding standard, HEVC, while the earlier works were based on H.264/AVC. Hence, this paper assesses the benefits of motion hints compared to the state-of-the-art temporal prediction offered by HEVC. *(ii)* we integrate an elastic motion model into the motion hints scheme, while the earlier works used an affine motion model.

The rest of this paper is organized as follows: in section II we describe the architecture of the proposed codec. Experimental results are reported in the following section. Finally, in section IV, we present our conclusions from these results.

II. STRUCTURE OF THE CODING/DECODING ARCHITECTURE

For simplicity, we use terms I-, P-, and B-frames similarly to their conventional interpretation e.g. in MPEG-2. I-frame for an intra-coded frame, P-frame for an inter-coded frame using reference frames preceding the P-frame in frame output order, and B-frame for an inter-coded frame that can use any available reference frames. B-frames are not used as reference frames for P-frames. However, the use of these terms does not limit the generality of the proposed method. For example, multiple reference frames or bi-prediction can be used in P-frames.

Figure 1 shows a simplified block diagram of the proposed coding/decoding architecture. The I- and P-frames are coded/decoded conventionally with HEVC. Each pair of adjacent I- or P-frames is input to the motion hints based inter-frame prediction process [7], which generates a hints-based reference frame (C_{hints}) for each B-frame. The B-frame coding/decoding then takes place using HEVC, where in addition to the usual temporal reference frames, the hints-based reference frame can also be used.

Further details of the coding/decoding architecture are provided in the following subsections. In II-A, the derivation of motion hints and the derivation of the hints-based reference frame are described. Section II-B explains how the proposed method makes use of the elastic motion model. Section II-C gives an insight into how the hints-based reference frame is used in conjunction with the temporal references to predict the current frame.

II-A. Bi-directional motion hints compensated reference frame

The first frame of each GOP is coded using the standard Intra mode of HEVC i.e. as an I-frame and transmitted to the decoder. The next frame, usually more than one time instance ahead, is coded using the temporally-predictive mode i.e. as a P-frame and the motion vectors along with the prediction residual signal are coded into a binary bit stream and transmitted to the decoder. Now these two already coded frames are fed into the bi-directional segmentation-based motion compensated prediction process that employs motion hints to generate reference frames for the intermediate ones.

In its core the motion hints based video coding process has a bi-directional inter-frame prediction component that requires the knowledge of spatial structure of the reference frames to predict that of the intermediate frame(s). This motion hints segmentation is an inverse problem and in most of the cases neither the shape nor the motion are known initially. The prediction algorithm attempts to solve this

problem by clustering the reliable block-based motion vectors between the previously decoded reference frames R_i and R_j . Four different reliability criteria were discussed in [7] and among them the kurtosis-based reliability approach was used in [8, 14] to generate the initial block-based motion hints segmentation.

The clustering step yields the initial foreground and background shapes of R_j and once the shapes are available their associated motion hints i.e. $M_{\text{fg}}^{(R_i \to R_j)}$and $M_{\text{bg}}^{(R_i \to R_j)}$can be estimated. These hints are then used to warp R_i to generate two predictions of R_j and from them the deduced two sets of motion hints compensated prediction errors, in conjunction with a color-based segmentation of R_j generates a refined segmentation map over R_j . With improved shapes, the accuracy of the motion hints $M_{\text{fg}}^{(R_i \to R_j)}$and $M_{\text{fg}}^{(R_i \to R_j)}$improves. This observation leads to an iterative hints estimation-shape refinement strategy employed to find the motion hints segmentation of R_j . The inverse of the converged motion hints $M_{\text{fg}}^{(R_i \to R_j)}$and $M_{\text{bg}}^{(R_i \to R_j)}$are taken as the initial values for the motion hints $M_{\text{fg}}^{(R_j \to R_i)}$and $M_{\text{bg}}^{(R_j \to R_i)}$respectively. The same iterative strategy is then replicated to find the motion hints segmentation of R_i . The motion hints $M_{\text{fg}}^{(R_i \to R_j)}$, $M_{\text{bg}}^{(R_i \to R_j)}$, $M_{\text{fg}}^{(R_j \to R_i)}$and $M_{\text{bg}}^{(R_j \to R_i)}$, are collectively referred to as the reference motion hints set herein. This approach can produce representative enough motion hints segmentation if the motion flow between R_i and R_j follows a smooth trajectory, as was shown in [7].

The motion hints $M_{\text{bg}}^{(R_i \to R_j)}$and $M_{\text{bg}}^{(R_j \to R_i)}$are utilized to determine the uncovered regions in R_j and R_i , which are then added to the existing backgrounds of R_i and R_j respectively to generate new backgrounds. After this the motion hints based prediction process scales the available four reference motion hint fields to come up with versions of these fields from R_i to the current frame C and R_j to C . More specifically, the four motion hint fields namely $M_{\text{fg}}^{(R_i \to C)}$, $M_{\text{bg}}^{(R_i \to C)}$, $M_{\text{fg}}^{(R_j \to C)}$, and $M_{\text{bg}}^{(R_j \to C)}$are obtained through projection of the estimated foregrounds and modified backgrounds of R_i and R_j onto C respectively. These newly estimated hints are quantized and then employed to warp R_i and R_j to generate four different predicted frames, namely $M_{\text{fg}}^{(R_i \to C)}\big(R_i\big)$, $M_{\text{fg}}^{(R_j \to C)}\big(R_j\big)$, $M_{\text{bg}}^{(R_i \to C)}\big(R_i\big)$ and $M_{\text{bg}}^{(R_j \to C)}\big(R_j\big)$. Finally, by fusing these predictions using appropriate weighting, $C_{\textit{hints}}$, the motion hints based reference frame for C is formed. For more details of all the steps summarized here please refer to [7].

II-B. Elastic motion model for the hints

In the works [7, 8, 14], the 6-parameter motion model i.e. the affine model was used to represent all the motion hints. For this work, we describe the hints by the elastic motion model that uses 2-D discrete cosine basis functions. The discrete basis functions used for motion modelling are defined as follows:

$$\begin{aligned}\phi_k(x_i, y_j) &= \phi_{k+\frac{P}{2}}(x_i, y_j) \\ &= \cos\Big(\frac{(2x_i+1)\pi u}{2M}\Big)\cos\Big(\frac{(2y_j+1)\pi v}{2N}\Big)\end{aligned}$$

where M and N are the horizontal and vertical dimensions of the block whose motion to be estimated. Since motion hints provide global description of motion i.e. their support is the whole frame so in our application M and N corresponds to the horizontal and vertical components of a frame respectively and the variables u and v refers to their associated discretized frequencies with $u, v = 0, \ldots, s-1; s = \sqrt{\frac{P}{2}}$. Here, the variable P represents the number of elastic motion parameters.

Having defined the basis functions, two frames $I(x_i, y_j)$ and $J(x_m, y_n)$ can be related by the following coordinate transformation:

$$\begin{aligned}x_m &= x_i + \sum_{k=1}^{\frac{P}{2}} m_k \phi_k(x_i, y_j) \\ y_n &= y_j + \sum_{k=\frac{P}{2}+1}^{P} m_k \phi_k(x_i, y_j)\end{aligned}$$

where $\{m_k\}_{k=1}^{P} = M^{I \to J}$ is the associated elastic motion hint estimated using standard gradient-based image registration techniques [17]. For our experimental analyses, we set $P = 8$ i.e. each elastic motion hint model has 8 parameters. The fractional part of the elastic hint is kept to the accuracy of $1/16$-th of a pixel and the hints are coded using the Exponential Golomb coding technique [18]. Each elastic hint has 8 parameters so in total $4 \times 8 = 32$ motion parameters are transmitted to the decoder to generate the reference frame $C_{\textit{hints}}$.

II-C. Motion hints based prediction as a reference frame

The hints-based reference frames are used for prediction similarly to inter-layer reference frames in the multiview extension of HEVC [19]. In practice, the hints-based reference frames are included in the reference picture lists for inter prediction of B-frames in addition to the normal temporal reference frames. There are three major consequences of this design: First, the encoder can make a RD optimized decision on prediction unit basis whether to use conventional reference frames or hints-based reference frames. Second, the hints-based prediction can be locally adapted with conventional motion vectors and motion-compensated prediction from the hints-based reference frame. Third, the generation

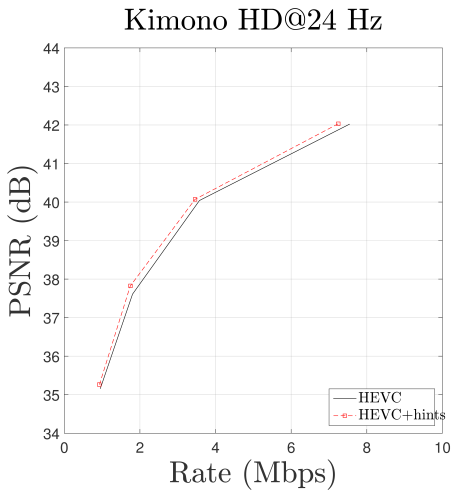

Fig. 2. Rate distortion performance of different coding strategies on the *Kimono* (1920×1080) sequence. A bit rebate of 5.10% is achieved by using the motion hints based references, along with the usual temporal references, for the B-frames.

of hints-based reference frames can be implemented as an inter-layer process into a multi-layer HEVC encoder/decoder architecture. Consequently, existing multi-layer HEVC implementations can be re-used as such in this design, and no block-level changes to the HEVC encoding/decoding processes are needed.

III. EXPERIMENTAL ANALYSIS

The RD performance of the proposed coder was investigated, on 4 different CIF sequences *Foreman*, *Stair* [20] where a walking person is followed by another person with a hand-held camera, *Carphone*, and *Handheld Smartphone* [21] where a moving person mimics video conferencing on a hand-held smartphone and on the full HD *Kimono* sequence.

The first 101 frames of each sequence was coded by the HM 16.5 reference software [22] for HEVC. The HM encoder was configured to have the GOP structure IBBBP i.e. GOP size = 4, Intra period = 12. Four different quantization parameter values (QP = 22, 27, 32, 37) were used. Only the I- and P-frames were used as references for other frames. Each pair of coded (I,P) or (P,P) frames were fed into the bi-directional motion hints based prediction process to generate the hints based predictions,C_{hints}, of the intermediate frames. To use these frames as references together with the usual references, which are the I- and/or P- frames, for the corresponding B-frames, all the C_{hints}frames were grouped together and supplied to the SHM 7.0 reference encoder [23] for the scalable extension of HEVC, as a base layer with inter-layer non zero motion estimation enabled. In the enhancement layer the original video sequences were used to perform quality scalability for the B-frames only.

Table 1. The Bjøntegaard delta gains obtained from the test sequences over standalone HEVC by using the motion hints based reference frames.

Sequence	Delta rate
Foreman	-1.58%
Stair	-3.75%
Carphone	-1.00%
Handheld Smartphone	-1.78%
Kimono	-5.10%

Therefore, the average PSNR for a test sequence, in the case where the C_{hints}frames are used as references, was calculated using the PSNR of I- and P-frames from the HM encoder and the PSNR of B-frames from the SHM encoder. In the same way, the overall bit rate was calculated and added to the bit rate for coding the motion hints.

Figure 2 shows the RD curve for the *Kimono* test sequence and all the obtained results are summarized in Table 1. The results are reported using the Bjøntegaard Delta [24] measurement method that computes the bit rate and average PSNR differences between two RD curves obtained from the PSNR measurement when encoding a content at different bit rates. All the used sequences showed encouraging gains, especially at high bit rates. The maximum savings in bit rate and gain in PSNR were achieved for the *Kimono* sequence. A smaller savings was noticeable for the *Carphone* sequence. The background of this sequence goes through multiple motions which is difficult to capture using a single motion model.

IV. CONCLUSION

In this paper, we investigated the feasibility of using the motion hints based predictions equipped with elastic motion model as reference frames for the B-frames. Experimental results showed an overall significant improvement in bit rebate of up to 5.10% over standalone HEVC if these frames are used as additional reference frames, which enables the re-use of HEVC implementations without low-level modifications..

V. REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.

[2] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 22, pp. 1649-1668, 2012.

[3] R.U. Ferreira, E.M. Hung, R.L. Queiroz, and D. Mukherjee, "Efficiency improvements for a geometric-partition-based video coder," *16th IEEE International Conference on Image Processing (ICIP)*, pp. 1009-1012, 2009.

[4] A.A. Muhit, M.R. Pickering, and M.R. Frater, "A fast approach for geometry-adaptive block partitioning," *Picture Coding Symposium (PCS)*, pp. 1-4, 2009.

[5] R. Mathew and D.S. Taubman, "Scalable Modeling of Motion and Boundary Geometry With Quad-Tree Node Merging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 2, pp. 178-192, 2011.

[6] S. Milani and G. Calvagno, "Segmentation-based motion compensation for enhanced video coding," *IEEE International Conference on Image Processing (ICIP)*, pp. 1685-1688, 2011.

[7] A. Ahmmed, R. Xu, A.T. Naman, M.J. Alam, M.R. Pickering, and D.S.Taubman, "Motion segmentation initialization strategies for bi-directional inter-frame prediction," *IEEE MMSP*, pp. 58-63, 2013.

[8] A. Ahmmed, M.J. Alam, M.R. Pickering, R. Xu, A.T.Naman, and D.S.Taubman, "Motion hints based inter-frame prediction for hybrid video coding," *Picture Coding Symposium (PCS)*, pp. 177-180, 2013.

[9] A.T.Naman, D. Edwards, and D. Taubman, "Efficient communication of video using metadata," *18th IEEE International Conference on Image Processing (ICIP)*, pp. 581-584, 2011.

[10] A.T.Naman, D. Edwards, and D. Taubman, "Inter-frame prediction using motion hints," *20th IEEE International Conference on Image Processing (ICIP)*, pp. 1792-1796, 2013.

[11] M.T. Orchard, "Predictive motion-field segmentation for image sequence coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 1, pp. 54-70,1993.

[12] J. Kim, A. Ortega, P. Yin, P. Pandit, and C. Gomila, "Motion compensation based on implicit block segmentation," *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 2452-2455, 2008.

[13] JM Reference Software for H.264/AVC. Available at: `http://iphome.hhi.de/suehring/tml/`

[14] A. Ahmmed, M.J. Alam, A.T.Naman, M.R. Pickering, and D.S.Taubman, "Motion hints mode for macroblock coding in bi-predictive slices," *Picture Coding Symposium (PCS)*, pp. 55-59, 2015.

[15] M.R. Pickering, M.R. Frater, and J.F. Arnold, "Enhanced Motion Compensation Using Elastic Image Registration," *IEEE Int. Conf. on Image Processing (ICIP)*, pp. 1061-1064, 2006.

[16] A.A. Muhit, M.R. Pickering, M.R. Frater, and J.F. Arnold, "Video Coding Using Elastic Motion Model and Larger Blocks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 5, pp. 661-672, 2010.

[17] Simon Baker and Iain Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework," *Int. J. Comput. Vision*, vol. 56, no. 3, pp. 221-255, 2004.

[18] I. E. G. Richardson, *The H.264 Advanced Video Compression Standard*, Wiley Publishing, 2010.

[19] M.M. Hannuksela, Y. Yan, X. Huang, and H. Li, "Overview of the Multiview High Efficiency Video Coding (MV-HEVC) standard," *IEEE Int. Conf. on Image Processing (ICIP)*, Sept. 2015.

[20] G. Zhang, J. Jia, W. Hua, and H. Bao, "Robust bilayer segmentation and motion/depth estimation with a handheld camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 603617, 2011.

[21] Handheld Smartphone sequence in CIF format. Available at: `https://sites.google.com/site/ashekahmmed/software/`

[22] HM Reference Software for HEVC. Available at: `https://hevc.hhi.fraunhofer.de/`

[23] SHM Reference Software for SHVC. Available at: `https://hevc.hhi.fraunhofer.de/shvc`

[24] G. Bjøntegaard, Calculation of average PSNR differences between RD curves, Technical Report VCEG-M33, ITU-T SG16/Q6, Austin, Texas, USA, 2001.