Conditional Motion Vector Refinement for Improved Prediction

Haricharan Lakshman[1], Christian Rudat[1], Matthias Albrecht[1],
Heiko Schwarz[1], Detlev Marpe[1] and Thomas Wiegand[1,2]
[1]Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany.
[2]Image Communication Chair, Technical University of Berlin, Germany.

***Abstract*— Adapting the resolution of motion compensated prediction in a video codec is considered in this paper. A new motion search and signaling scheme for increasing the accuracy of prediction without affecting the complexity of encoding is proposed. It involves using common information available to both encoder and decoder, e.g. current slice type, texture in reference pictures, etc as a cue to control motion vector accuracy and an efficient reuse of motion information for predicting subsequent blocks. An average bit rate reduction of around 2.5% for P-pictures and 0.5% for B-pictures is observed with no extra search compared to a fixed quarter-sample resolution.**

I. Introduction

Motion compensated prediction (MCP) using fractional-sample resolution is commonly used in video coding. The encoder transmits motion information, quantized prediction residuals together with other side information in the bitstream. In H.264/AVC [1], the motion vectors (MVs) have a quarter-sample resolution. A theoretical analysis of the efficiency of fractional sample filtering is provided in [2]. A structure in which the reference samples are first passed through interpolation filters to phase delay the signal followed by a Wiener filter is considered. The interpolation filters provide the required fractional shift while the Wiener filter smoothes the prediction signal to counter the noise in signals and the displacement estimation errors. However, in the current HEVC draft [3], an adaptive loop filter and a sample adaptive offset filter is introduced after the DPCM reconstruction stage, which reduce noise in the reference pictures. Therefore, the role of interpolation filtering is to mainly provide fractional shifts without introducing much attenuation to the high frequency components of the reference signals. In order to improve the interpolation filtering, techniques like increasing the tap length [5], adapting the filter coefficients [6], combined FIR and IIR filtering [7], etc have been proposed. Mostly, interpolation filters that induce the required phase delays for larger passbands need additional complexity for implementation.

An alternative approach is to keep the filter length constant, but to increase the number of available filters for MCP. Such a scheme would not increase the complexity of a decoder because only a single filter of the same length would be used for predicting each hypothesis. The larger set of filters could be used to provide more choice for the encoder in selecting a phase delay during MCP. In this regard, one-eighth sample resolution for MV has been explored in several works, e.g. [8], [9]. However, the reported gains due to a one-eighth sample motion resolution are limited and could even result in a degradation of RD performance in some cases due to the extra side information sent in the bitstream. Additionally, one-eighth resolution also requires extra search on the encoder side compared to the commonly used one-quarter resolution, hence resulting in an increase of encoder complexity.

In this paper, we propose a new technique by modifying the order of fractional sample motion estimation to achieve a resolution higher than one-quarter sample without increasing the complexity of motion estimation. A hierarchical motion search is commonly used to accomplish fractional sample search by first generating an integer sample MV, followed by half and quarter-sample MVs. In such a scheme, we use the half-sample search results to selectively search in a one-sixth sample grid instead of a one-quarter grid. This helps to keep the encoder complexity same as in the case of one-quarter sample search. Signaling a location in a full one-sixth grid however needs more motion information. Therefore, we operate the MV predictor in a regular quarter-sample resolution and first transmit the motion vector difference (MVD) in the quarter-sample resolution. Then we define a set of conditions using the current slice type, texture in the referenced pictures, number of hypothesis, etc, to control the transmission of MV refinement information that can yield one-sixth sample resolution. The new fractional positions are generated using FIR interpolation filters as in the quarter-sample case. Therefore, the set of possible amplitude responses and phase delays are increased, giving more opportunity for an encoder to choose the best operating point for coding a given block. The decoder follows the same steps in order to determine whether the refinement information exists for each MV, without any explicit signaling of MV resolution in the bitstream. The refinement information is further reused for subsequent prediction blocks having the same MV by using the Merge mode in HEVC [3]. The MCP for chroma is done synchronized to the luma MV, with the required chroma downsampling. We compare the proposed technique to a scheme which uses a fixed quarter-sample grid and measure the performance gains. Finally, the test is extended to include extra motion search and signaling to get an estimate of RD performance improvements when consuming more complexity at the encoder for the same architecture.

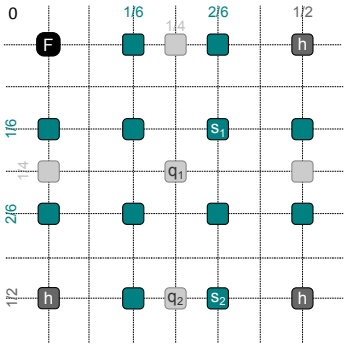

Fig. 1. *Fractional sample grid for motion representation. Integer, half and quarter-sample positions are denoted by F, h and q, respectively. One-sixth sample resolution is denoted by s.*

II. Refinement framework

Increasing the MV resolution has a direct impact on the number of positions to be searched by an encoder to determine a MV for predicting a given block of samples. It also increases the motion information to be transmitted in the bitstream. In this section we address these issues to make adaptive MV resolution more attractive for practical implementations.

A. Encoder search without complexity increase

Consider a fractional sample grid as shown in Fig. 1, depicting integer, half and quarter sample positions using the letters *F*, *h* and *q*, respectively. A popular approach for motion search in video encoders is the so called hierarchical scheme, in which, first the integer positions are searched in a window of samples in the reference picture. Without a loss of generality, let us consider *F* as the integer position yielding the best RD cost during the integer position search. After that, another search is carried out to evaluate the RD costs of the eight half-sample positions around *F*. Then, there are two possible cases: (a) integer position has the best RD cost, or/and (b) a half-sample position has the best RD cost. After this point, the adjacent quarter-sample positions, around the best found position are evaluated. Note that the quarter-sample positions in between integer and half-sample positions are tested in both cases (a) and (b). In this paper, we exploit the already available RD costs after the half-sample search to reduce the search space in a one-sixth resolution grid to the same number of search points as a quarter-sample grid. To this end, let us denote one-sixth sample positions in Fig. 1 as *s*. The approach we follow is to search only the nearest neighbor one-sixth sample positions from the best position after the half-sample search. For instance, if the best (x, y) position after the half-sample search is found to be $(\frac{1}{2}, 0)$, we then search eight positions $(\frac{2}{6}, 0)$, $(\frac{2}{6}, \frac{1}{6})$, $(\frac{1}{2}, \frac{1}{6})$, and so on.

Although quarter-sample positions, like $(\frac{1}{4}, 0)$, are not a subset of the one-sixth sample grid and therefore not included in the motion search, they are still used for MV representation. A one-sixth resolution MV is decomposed into the closest quarter-sample base MV and an additional refinement information. With reference to Fig. 1, position s_1 is represented as position q_1 in conjunction with a vertical and a horizontal refinement. In case of s_2, MV representation only needs a horizontal refinement to the base MV pointing to q_2, so vertical refinement is neither stored nor transmitted.

B. Texture based MV resolution control

The reference samples used for MCP influence the approximation error due to a finite precision quantization of MV. In order to characterize the influence, we use a first-order Taylor approximation of the signal $f(x)$ around a quarter-sample position x_0 and denote the MV refinement as Δx,

$$f(x_0 + \Delta x) \approx f(x_0) + \Delta x \cdot \frac{\partial f}{\partial x}(x_0) + \cdots \qquad (1)$$

When using a fixed quarter-sample resolution for x, the refinement Δx can be considered as zero. It can be seen that if the reference signal does not have a sufficiently large gradient $\frac{\partial f}{\partial x}$ at the position of interest x_0, the influence of MV quantization error is limited. This fact can be exploited to define rules known to both encoder and decoder to implicitly infer MV resolution for predicting a block given only the integer component of the MV. The decoder would then interpret the parsed fractional value according to the inferred MV resolution. However, such a scheme would require the decoder to evaluate the gradients for each block during MCP which adds to the computational complexity.

Instead of gradient computation on a block-by-block basis, we estimate the texture content of an entire slice and transmit a flag in the bitstream. To determine whether a reference slice contains high texture or not, we approximate the gradients using a first-order difference, as given by Eq. 2. It is split into two summations, one of which calculates the horizontal differences between adjacent samples and the other the vertical ones. This result is then normalized by the number of samples used for the computation.

$$\begin{aligned} \sigma(f) &= \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=0}^{N} |f(i,j) - f(i-1,j)| \\ &+ \frac{1}{M \cdot N} \sum_{i=0}^{M} \sum_{j=1}^{N} |f(i,j) - f(i,j-1)| \end{aligned} \qquad (2)$$

A slice is then classified as containing high texture if the estimated gradient $\sigma(f)$ is greater than a pre-defined threshold. Then, a set of conditions, detailed in the Sec. III, is checked before transmitting refinement information to reduce the overhead due to extra side information.

III. Transmission of refinement information

In hybrid video codecs like the H.264/AVC and the draft HEVC, the estimated MVs are predicted using a MV predictor and the motion vector difference (MVD) is transmitted in the bitstream. If we increase the accuracy of the MV, the predictor has to be modified to support the higher resolution. One solution is to operate the MV predictor also in the high resolution grid. The rate overhead due to the higher MV resolution would then be shared between the possible integer and fractional positions. In this paper, we utilize a scheme in which integer and half-sample positions have no

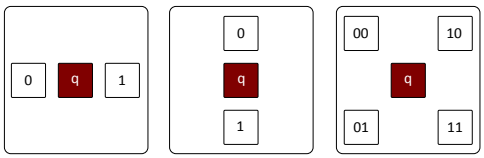

Fig. 2. *Refinement bits to be transmitted relative to the quarter-sample position q. Left: horizontal refinement only, Center: vertical refinement only, Right: horizontal and vertical refinement.*

rate overhead, but predictions requiring higher MV resolutions bear the overhead completely. The MV predictor is operated at a quarter-sample resolution and the refinement information is transmitted only for the positions landing on the one-sixth sample grid that are not part of the half-sample grid. Specifically, positions $\frac{0}{6}$ and $\frac{3}{6}$ coincide with the half-sample grid, but $\frac{1}{6}$, $\frac{2}{6}$, $\frac{4}{6}$ and $\frac{5}{6}$ do not. These new positions are mapped on to the nearest quarter-sample positions and used for MV prediction. Then, refinement information is transmitted that enables the decoder to recover the original one-sixth resolution position. MV coordinates are refined by sending one bit indicating to either increase or decrease their value from a quarter-sample position to the next one-sixth sample position as shown in Fig. 2. For positions that map on to quarter-sample postions both in horizontal and vertical directions, a total of two bits are transmitted.

Criteria for refinement

Adapting the MV resolution can entail two forms of rate overhead: rate to indicate current MV resolution, and rate for MV refinement, if needed. In this paper, we define a set of rules known to both the encoder and decoder to infer the MV resolution without any explicit block-by-block forward signaling. The set of conditions is formulated so as to maximize the cases in which the extra rate for refinement can actually yield improved prediction. A flow diagram of the decoding process including the criteria for infering the existence of refinement is depicted in Fig. 3. After decoding the MVD, the base MV is reconstructed by adding the MVD to the MV prediction value. The reconstructed MV components that point to integer or half-sample positions are left unaltered. For the MV components that point to one-quarter or three-quarter positions, the decoder infers the presence of refinement using the following conditions:

- In case of a P-slice, the absence of Bi-prediction significantly hampers the accuracy of MCP. Hence, the MV refinement information is always sent, as a substitute for the second MV.
- In case of a Bi-prediction, the MV refinement is sent for the predictions that access samples from a reference picture that contains high texture. The decoder does not need to estimate the texture content in the reference picture because a flag is sent in the bitsteam.
- In case of Bi-predictions not accessing high texture reference pictures, the MV refinement is sent only for pictures from a pre-defined reference picture list.
- For single hypothesis predictions in B-slices, MV refinement is not used.

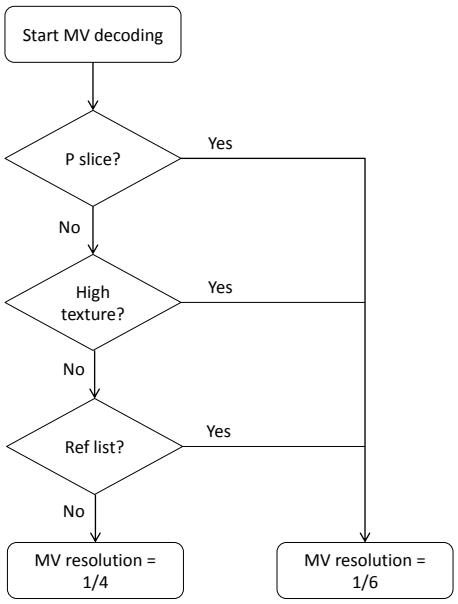

Fig. 3. *Flow diagram of the decoding process for determining the presence of MV refinement for one-quarter and three-quarter positions. Integer and half-sample positions are left unaltered by the proposed algorithm.*

The current version of HEVC draft [3] features a tool for inter frame prediction called *block merging* [10] which is an approach to reduce the signaling redundancy in inter prediction. Block merging enables the motion information (MVs, reference picture indices, etc.) to be shared between the adjacent prediction blocks. Using the same technique, the MV refinement information introduced in this paper is also shared between all the merged blocks for an efficient reuse of the side information. Note that in the proposed conditional refinement scheme, MVD of the current block has to be added to the MV predictor to reconstruct the base MV, in order to continue parsing the current block refinement and the MVD of the other blocks.

IV. Simulation Results

The RD performance of the proposed conditional MV refinement is evaluated using the test sequences defined in the recommended test conditions issued by the JCT-VC [4]. Three constraint sets restricting the coding structures are used for evaluation: (a) GOP8 random access: structural delay not larger than 8-pictures, (b) Low delay B: inter prediction using up to 2 hypothesis and no picture reordering in decoder, (c) Low delay P: inter prediction using 1 hypothesis and no picture reordering in decoder. In all coding structures, the encoder can decide between inter and intra modes for each prediction block. The software base of our experiments is the HEVC Test Model 4 (HM 4.0) [3] of the ongoing HEVC standardization. All sequences from the HEVC dataset are used for testing and 50 frames per sequence are coded.

We compare the proposed conditional MV refinement scheme with HM 4.0 anchor which has a quarter-sample MV grid. The improvements are measured in terms of the average bit rate difference over all rate points using the Bjøntegaard Delta (BD) rate metric [11]. The simulation results for the

TABLE I

High Efficiency coding scenario: BD rate comparison of the proposed conditional MV refinement and a fixed quarter-sample MV resolution.

High Effiicency	Y-BD bit rate %		
	LowDelay P	LowDelay B	GOP8
Class A - 2560 × 1600	-	-	-0.2
Class B - 1920 × 1080	-1.7	-0.4	-0.6
Class C - 832 × 480	-2.5	-0.4	-0.2
Class D - 416 × 240	-3.7	-0.4	-0.5
Class E - 1280 × 720	-2.0	-0.4	-
Average	**-2.5**	**-0.4**	**-0.4**

TABLE II

Low Complexity coding scenario: BD rate comparison of the proposed conditional MV refinement and a fixed quarter-sample MV resolution.

Low Complexity	Y-BD bit rate %		
	LowDelay P	LowDelay B	GOP8
Class A - 2560 × 1600	-	-	-0.2
Class B - 1920 × 1080	-1.8	-0.5	-0.6
Class C - 832 × 480	-2.5	-0.8	-0.5
Class D - 416 × 240	-3.7	-1.3	-1.0
Class E - 1280 × 720	-1.4	-0.4	-
Average	**-2.4**	**-0.7**	**-0.6**

high efficiency coding can be found in Tab. I and the low complexity coding in Tab. II. BD rate savings can be observed in all configurations and sequence classes. Savings up to 3% are observed in the high efficiency configuration using hierarchical B pictures (GOP 8 random access and low delay coding) and up to 5% in corresponding low complexity configurations. The rate savings in the low delay P coding structure is more due to the absence of Bi-prediction. Here, inter prediction benefits more from the MV refinement information with an average bit rate savings of around 2.5% and peak savings of around 14%. Some sequences like PartyScene and BQSquare from the HEVC dataset show significant performance gains, whereas the gains for other sequences are moderate.

We extend the proposed conditional MV refinement scheme to analyze the case when additional complexity is available at an encoder. With the number of search points similar to that of a hierarchical $\frac{1}{8}$th sample motion estimation, the search scheme proposed in this paper can be employed to achieve a resolution of $\frac{1}{12}$th of a sample. It follows the same structure as in Sec. II, with $\frac{1}{12}$th MV refinement after a quarter-sample search. But it entails more rate for motion information than a one-sixth resolution MV in order to indicate the exact position in the higher resolution grid. After the transmission of a quarter-sample MVD, a flag is transmitted to indicate whether the MV has refinement or not. The eight possible $\frac{1}{12}$th sample refinement positions adjacent to the quarter-sample indicated by the MVD are distinguished using three bits. As described in Sec. II-B, MV refinement is bypassed if the reference picture is considered as low texture. Results for conditional $\frac{1}{12}$th resolution MV tests can be found in Tab. III. For P-pictures, average bit rate savings of around 3.6% and a peak saving of 19% is observed compared to HM 3.2 [12] as a reference. The encoder complexity increase is mainly due to the fact that the decision to transmit $\frac{1}{12}$th sample refinement is taken in the RD optimization loop for each block partition.

TABLE III

High Efficiency coding using $\frac{1}{12}$th sample MV Refinement compared to a quarter-sample MV resolution, with extra search and RD optimization.

High Efficiency	Y-BD bit rate %		
	LowDelay P	LowDelay B	GOP8
Class A - 2560 × 1600	-	-	-0.5
Class B - 1920 × 1080	-2.7	-0.6	-0.8
Class C - 832 × 480	-4.0	-0,6	-0.8
Class D - 416 × 240	-5.5	-0.7	-1.1
Class E - 1280 × 720	-2.2	-0.2	-
Average	**-3.6**	**-0.6**	**-0.8**

V. Conclusion

In this paper we descibed a method to increase the resolution of the motion vectors to one-sixth of a sample without introducing any extra search compared to a quarter-sample resolution. The found high resolution MV is transmitted as a regular quarter-sample MV augmented with refinement information. A set of conditions based on the information available to the decoder is defined to reduce the overhead due to the extra side information. Moderate bit rate savings for B-pictures and high savings for P-pictures are reported. Hence, conditional MV refinement offers an interesting way to improve the prediction without much changes to interpolation filter length, reference picture memory accesses and computational complexity.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard", *IEEE Trans. Circuits Syst. Video Technol.,* vol. 13, pp. 560–576, Jul. 2003.

[2] B. Girod, "Motion-compensating prediction with fractional-pel accuracy," IEEE Tran. on Comm, vol. 41, pp. 604 –612, Apr 1993

[3] JCT-VC, "WD4: Working Draft 4 of High-Efficiency Video Coding," JCT-VC F803, Torino 2011.

[4] F. Bossen, "Common test conditions and software reference configurations", JCT-VC F900, Torino 2011.

[5] W.-J. Han, et al.,"Video coding technology proposal by Samsung (and BBC)," *JCT-VC A124*, Dresden, Apr. 2010.

[6] T. Wedi, "Adaptive Interpolation Filters and High-Resolution Displacements for Video Coding'," *IEEE Trans. CSVT.*, vol. 16, no. 4, Apr. 2006.

[7] H. Lakshman, H. Schwarz, T. Blu, and T. Wiegand, "Generalized Interpolation for Motion Compensated Prediction," *Int. Conf. Img. Proc., Brussels, Sep 2011.*

[8] T. Wedi, "Hybrid Video Coding Based on High-Resolution Displacement Vectors," *Proc. VCIP*, Jan 2001

[9] W.-J. Chien, "Summary Report of Adaptive Motion Vector Resolution', *JCT-VC D362*, Jan. 2011.

[10] S. Oudin, P. Helle, J. Stegemann, C. Bartnik, B. Bross, D. Marpe, H. Schwarz, T. Wiegand, "Block merging for quadtree-based video coding," IEEE International Conference on Multimedia and Expo (ICME), 2011, Jul 2011.

[11] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," ITU-T VCEG-M33, Apr. 2001.

[12] JCT-VC, "WD3: Working Draft 3 of High-Efficiency Video Coding," JCT-VC E603, Geneva 2011.