ABSTRACT

Recent video coding standards such as high efficiency video coding (HEVC) and multi-view video coding (MVC) can obtain significant gain from hierarchical prediction structures. Hierarchical B structure utilizes future reference frames, which induces high encoding delay. Comparatively, hierarchical P structure brings low-delay coding in which no future references are involved. The only hurdle of the low-delay hierarchical P structure is the need of motion estimation in distant reference frames. To maintain a high coding efficiency, large search range for motion estimation was commonly adopted in distant reference pictures. Therefore, computational complexity will be increased dramatically. In this paper, motion vector composition is proposed in hierarchical P structure to shorten coding process. Simulation results reveal that the proposed scheme can achieve significant complexity savings.

Index Terms— Video coding, vector composition, low-delay hierarchical P structure, distant reference pictures

1. INTRODUCTION

Owing to the increased flexibility of H.264 [1], hierarchical prediction structures [2,3] have become popular in the new standards [4] such as HEVC and MVC. In H.264, there are two types of hierarchical prediction structures. They are hierarchical B (HB) and hierarchical P (HP) structures. A typical HB prediction structure [5] is shown in Fig. 1. In the figure, I/P-frames are key frames in the temporal base layer, denoted as $T_0$ which are firstly coded in a group of pictures (GOP). The non-key frames within the GOP are coded as B pictures. The B-frame in layer $T_1$ is coded with two reference frames in the lower temporal layer, $T_0$. After that, it becomes one of the reference frames to the frames in the next temporal layer, $T_2$ and so on. The HB structure applies bidirectional prediction which utilizes a reference from the future. As a result, the coding order for this HB structure has to be re-arranged such that future frames are coded before being reference frames. This burdens the associated coding delay and memory requirement [6]. Constraining delay is of great importance for real-time applications such as video conferencing and live event broadcasting. Therefore, the HB structure is not favorable for real-time applications.

In order to reduce the encoding delay, a HP structure, as shown in Fig. 2, employing only P-frames has been designed [2,3,7]. This structure does not employ motion-compensated prediction from future frames but provides the same degree of temporal scalability as the HB structure [3]. A typical HP structure with 4 hierarchy/temporal layers is illustrated in Fig. 2. The leading I-frame and the last P-frame are the key frames in the temporal base layer, $T_0$. The non-key frames within a GOP are coded as P frames, which are different from the HB structure [1]. Since no B-frames are inserted, the frames can be coded according to the display order, as stated in Fig. 2. It implies that the coding delay tends to zero. Thus, the HP structure is a low-delay coding structure and is very suitable for delay-sensitive applications. In the HP structure, encoding complexity is much higher than that of the classical “IPPP...” structure because motion estimation (ME) over remote reference frames does occur in the base layer and rate-distortion (R-D) performance deteriorates. Such problem can be solved by using a larger search range but the computational complexity will be increased significantly.

In this paper, we propose a scheme with motion vector (MV) composition in the HP coding. It aims at providing better R-D performance with an acceptable search range in ME. Furthermore, computational complexity is reduced. This paper is organized as follows. In Section 2, the motivation of this research work on HP coding is stated. Section 3 describes the proposed scheme with MV composition. Simulation results are presented in Section 4. Finally, some concluding remarks are given in Section 5.
3. PROPOSED SCHEME

Now, let us define some symbols for the sake of illustration of the proposed scheme. In the HP structure with the GOP size of \( L \), each current frame \( F_n \) to be coded has its corresponding reference frame \( F_r \) which has different prediction distances according to temporal layers, \( T_k \) where \( k=0,1,\ldots,\log_2 L \). The indices \( n \) and \( r \) are related as follows:

\[
r = n - 2^{(\log_2 L) - k}
\]

In a ME process between frames, let \( CB = (x, y) \) be the top left coordinate of the current block being coded and be a MV of a frame \( F_\beta \) referenced by a frame \( F_\alpha \). The sum of absolute differences (SAD) is calculated as

\[
f_{SAD}(CB,MV_{\alpha \rightarrow \beta}) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |F_\alpha(CB + (i,j)) - F_\beta(CB + MV_{\alpha \rightarrow \beta} + (i,j))|
\]

The proposed scheme involves MV composition as a new coding framework introducing the MV reuse by the MV sum. The proposed scheme consists of three core steps. First, all MVs between adjacent frames are estimated through ME with a relatively small search range. Second, we obtain composed MVs for lower layers using the computed MVs in the previous step. Third, the final MV can be obtained by performing refinement of the composed MV over a narrow search range.

3.1. ME in Consecutive Frames with a Smaller Search Range

For consecutive frames, each block of \( N \times N \) pixels is undergone conventional full-search ME to obtain a MV. By minimizing \( f_{SAD}(CB,MV_{\alpha \rightarrow \beta}) \) in (2) through the conventional full-search ME with a small search range (SR), MVs of all blocks in a frame \( F_t \) pointing to the previous frame \( F_{t-1} \) are obtained and they are denoted as a set of \( MV_{\text{SR}}^{t-1} \), which comprises MVs in the highest layer, \( T_1 \), indicated by the dotted arrows in Fig. 2.

3.2. Optimal Motion Vector Selection and Composition

In lower layers, MV composition is conducted with the help of \( MV_{\text{SR}}^{t \rightarrow t-1} \) obtained above. We are the first to propose MV composition in HP coding as a new coding framework. To compute MV set, \( MV_{r \rightarrow \alpha} \) for lower layers, \( F_n \) with its corresponding reference \( F_r \) are composed by

\[
MV_{r \rightarrow n} = \sum_{i=0}^{N-1} \text{select} MV_{r \rightarrow i-1}^{\text{SR}}
\]

where \( \text{select} MV_{r \rightarrow i-1}^{\text{SR}} \in MV_{r \rightarrow i-1}^{\text{SR}} \).
For each MV composition step, \( selectMV_{t-1\rightarrow t-1}^{SR} \) is a MV for a selected block from the set of \( MV_{t-1\rightarrow t-1}^{SR} \). For instance, from (3), the composed MV for the fourth P frame in \( T_2 \) shown in Fig. 2 can be computed as

\[
MV_{t-1\rightarrow t-1} = selectMV_{t-1\rightarrow t-1}^{SR} + selectMV_{t-2\rightarrow t-2}^{SR} + selectMV_{t-3\rightarrow t-3}^{SR} + selectMV_{t-4\rightarrow t-4}^{SR}
\]

The success of a composed MV for large skip factors depending on the reliability of \( selectMV_{t-1\rightarrow t-1}^{SR} \) has been raised in our previous work [10] in video transcoding. This motivates the proposed HP coding scheme to have precise consecutive MV selection of \( selectMV_{t-1\rightarrow t-1}^{SR} \) for each block in coding with a distant reference. We found that the most popular vector selection technique in MV composition algorithm, forward dominant vector section (FDVS) [11], is inappropriate for cases when the temporal distances between the references and the current frame are far away in HP coding. The main philosophy of the proposed MV selection scheme and FDVS are illustrated in Fig. 3. Originally in FDVS, \( selectMV_{t-1\rightarrow t-1}^{SR} \) is defined as the MV of the dominant compensated block which gets the largest size of compensated area, i.e. \( MV' \) of block \( MB' \) in \( F_t \), shown in Fig. 3. It will further select \( MB' \) as the dominant block by the largest dot-bordered area in \( F_{t-1} \). This criterion for selection may lead to a worse coding performance when the number of composition steps increases. The reason is that the dot-bordered area in \( F_{t-1} \) under consideration comprises not only content relevant compensated area (grey region) of the target block in \( F_t \) but also a large portion of non-relevant area (white region) reducing the reliability of \( selectMV_{t-2\rightarrow t-2}^{SR} \).

To tackle this, our proposed MV selection algorithm firstly only uses the actually relevant dominant compensated area which is the grey area instead of the whole dot-bordered area in \( F_{t-1} \) to determine the next optimal MV, as depicted in Fig. 3. Secondly, the proposed scheme can entirely make use of the homogeneity of MVs in which compensated areas in blocks with the same MV are merged. For the example in Fig. 3, \( MB' \) and \( MB' \) in \( F_t \), which have the same MV, are merged for dominant block determination. Under our scheme, the final selected dominant MV for \( selectMV_{t-2\rightarrow t-2}^{SR} \) is picked by the crossed region in \( MB' \) of \( F_{t-1} \), not \( MB' \) from FDVS. The proposed scheme also keeps multiple composed MVs according to the compensated area size during MV composition until the reference \( F_t \) is reached.

3.3. Refinement after Motion Vector Composition

After MV composition, a small refinement vector \( \varepsilon \) is added to the resultant MV, which is expressed as

\[
\varepsilon = (\varepsilon_x, \varepsilon_y) \text{ where } \varepsilon_x, \varepsilon_y \in [-1,0,1]
\]

The refinement operation provides fine tunings for the MV composition. It adds an optimum \( \varepsilon_{opt} \) which can yield a minimum value of \( f_{SAD} \), such that

\[
\varepsilon_{opt} = \arg\min_{\varepsilon} f_{SAD}(CB, MV_{t-\rightarrow t} + \varepsilon)
\]

With (6), an final MV \( (MV'_{t-\rightarrow t}) \) with refinement is calculated as

\[
MV'_{t-\rightarrow t} = \sum_{t-t+1} selectMV_{t-1\rightarrow t-1}^{SR} + \varepsilon_{opt}
\]

From the above coding order, the forward coded MVs between adjacent frames are utilized in each composition step. Only ME among adjacent frames is necessary. In this case, a small search range is sufficient. The computational complexity thus can be significantly reduced by eliminating the need to search for temporal remote reference frames. It can be seen that the coding order of MV composition is also the same as the display order shown in Fig. 2, which does not impose extra delay in the coding process.

4. SIMULATION RESULTS AND DISCUSSIONS

In this section, we present some simulation results to evaluate the performance of adopting various vector selection algorithms of MV composition in HP coding. All the simulations were based on the H.264 JM17.2 codec [9]. The test sequences used in the simulations include “Bus (CIF), “Football (CIF), “Stefan (CIF), “Rush hour (720p), and “Spinacalendar (720p)”. All sequences have 100 frames. In HP coding, a quantization parameter (QP) is set for each \( k \) temporal layer (\( QP_k \)), as in

\[
Q_{P_k} = Q_{P_{base}} + k \text{ where } k = 0, 1, 2, 3
\]

Here, \( QP_k \) depends on the QP for the base layer denoted as \( QP_{base} \). For R-D performance analysis, \( QP_{base} \) was set to 20, 24, 28, and 32. The bitstreams were encoded with the HP structure by different algorithms. The evaluation in Bjontegaard (BD) measurement of proposed HP coding.
scheme in various vector selection algorithms in MV composition versus full-search algorithm was conducted. The vector selection algorithms in MV composition between consecutive frames include Forward Dominant Vector Selection (FDVS) [11], its enhanced version (E-FDVS) [12], median algorithm (MEDIAN) [13] and the proposed one with 4 candidates. To evaluate various vector selection algorithms in HP coding effectively, comparisons were focus frames in $T_0$ and $T_1$ involving distant references in a GOP. The BD measurement was done in two groups and listed in Table 2 and 3, respectively, as

- For CIF: full-search with SR=±8 ($FS_{SR=8}$) versus MV composition algorithms with $selectMV_{t-1-m}$; then, further versus $FS_{SR=16}$.
- For 720p: full-search with SR=±16 ($FS_{SR=16}$) versus MV composition algorithms with $selectMV_{t-1-m}$; then, further versus $FS_{SR=32}$.

From Table 2, all MV selection and composition algorithms with $selectMV_{t-1-m}$, for CIF sequences in $T_0$ and $T_1$ outperform $FS_{SR=8}$ by bitrate reduction and quality increment. A similar trend appears in all MV composition with $selectMV_{t-1-m}$ for 720p sequences when they are compared with $FS_{SR=16}$ in Table 3. This implies that MV composition algorithms can enhance ME in $T_0$ and $T_1$ when temporal remote reference frames are used, and are successfully adopted in HP coding.

Moreover, among the MV composition algorithms in HP coding, the proposed scheme provides higher coding efficiency than others. The performance in both BD-PSNR and BD-Bitrate of the proposed one for CIF sequences gets closer remarkably to $FS_{SR=16}$ as shown in Table 2. For 720p sequences, the proposed scheme even outperforms $FS_{SR=32}$ by average of 2% BD-Bitrates reduction and 0.2 BD-PSNR increment in Table 3. The reason for the proposed scheme outperforming others is that it only utilizes related partition of compensated area to the target MB for MV selection. The proposed scheme enlarges the relevant partition as possible in every MV composition step by merging homogeneous area while ensuring the resultant MV is highly correlated to the target MB in the current frame, which cannot be achieved by FDVS, E-FDVS and MEDIAN.

The number of search points per macroblock ($num_{st}$) required for ME is listed in Table 4, which is the unit of computation to perform the results independent of platforms. The calculation for $FS_{SR=8}$ was calculated as

$$num_{st} = (2x + 1)^2$$  \hspace{1cm} (9)

Table 4 shows all MV composition algorithms have almost the same $num_{st}$ in comparison to FS with the same search range in consecutive frames. To compare the complexity of MV composition with $selectMV_{t-1-m}$ to $FS_{SR=16}$ and $selectMV_{t-1-m}$ to $FS_{SR=32}$, MV composition algorithms can greatly reduce $num_{st}$ while similar or sometimes better coding performance is achieved as listed in Table 2 and 3.

### 5. CONCLUSION

A coding scheme for HP coding, especially for distant references in lower temporal layers, has been proposed in this paper. The scheme involves heuristic MV selections, compositions, and refinements after re-formations. It successfully avoids large search window in ME while coding efficiency is maintained. Simulation results show that the proposed scheme with MV composition performs well in HP structure coding, in which our proposed vector selection scheme has been proven to outperform other MV composition algorithms. Under the same searching conditions, our scheme can even perform better as compared with the full-search algorithm in lower temporal layers.

### 6. ACKNOWLEDGMENT

The work described in this paper is partially supported by the Centre for Signal Processing and a grant from the Internal Competitive Research Grant, Department of EIE, HKPolyU (PolyU G-YL20). Tsz-Kwan Lee acknowledges the research studentships provided by the University.
7. REFERENCES


