One-Pass Mode and Motion Decision for Multilayer Quality Scalable Video Coding

Meng Xu, Zhan Ma, *Member, IEEE*, and Yao Wang, *Fellow, IEEE*

***Abstract*— This paper presents a novel low-complexity motion estimation and mode decision algorithm for encoding multiple quality layers following the H.264/scalable video coding standard, considering both coarse grain scalability (CGS) and medium grain scalability (MGS). The proposed algorithm conducts motion estimation and mode decision only at the base layer (BL) and enforces the higher layers to inherit the motion and mode decisions of the BL. In order for the decision made at the BL to be nearly optimal for all layers, we use the highest layer reconstructed frame as the reference frame for motion estimation and set the Lagrangian multipliers according to the quantization parameter of the current and higher layers. We also propose a simple early skip/direct decision to further boost the encoding speed. Mode decision and motion estimation is conducted at a higher layer only if the layer below it uses the skip/direct mode for a block. Significant complexity reduction can be achieved because the mode and motion estimation is performed at most once for each macroblock. Because the mode and motion information only needs to be transmitted once, we also achieve a slightly better rate–distortion (R–D) performance for typical videos. Experiments have shown more than 2× (up to 5×) speedup for a three-layer encoder against the conventional R–D optimized reference software JSVM on both CIF and HD sequences, and for both CGS and MGS, with the tradeoff of the coding efficiency measured by the Bjontegaard delta rate.**

***Index Terms*— Fast motion and mode decsion, CGS, MGS, H.264/SVC.**

I. Introduction

SCALABLE video holds a great potential for networked video applications where a single bit stream can be adapted at the network gateway or proxy according to the constraints from the underlying network and/or heterogeneous clients. Scalable video stream typically includes one base layer (BL) at limited quality and one or more enhancement layers (ELs) providing improved quality.

We choose the H.264/SVC [1], [2] or simply SVC to demonstrate our ideas because SVC is already adopted in the commercial market such as Vidyo conferencing system, Google+, etc. What we have proposed here could benefit the product immediately. On the other hand, it requires substantial efforts to migrate these ideas to the recently finalized scalable extension of High-Efficiency Video Coding (HEVC) [3], which is deferred for future research.

SVC is developed on top of the well-designed H.264/AVC with additional inter-layer coding tools to exploit the inter-layer correlations. As in prior video coding standards, H.264/AVC and SVC also adopt the block based coding structure, where each non-overlapped block with 16×16 pixels is the basic coding unit called macroblock (MB). It can be further split into smaller partitions [4]. In a conventional implementation of SVC encoder (including the SVC reference software JSVM [5]), at each layer, the mode of each block is determined by exploring all H.264/AVC compatible modes, as well as additional SVC modes if applicable. This makes an L-layer encoder having the complexity more than L-times of a single layer encoder as observed in [6]. In addition, SVC reference implementation optimizes the mode at the current layer by only considering this layer, and thus will not guarantee the global optimality for all layers.

Generally, SVC optimization has been pursued along two avenues. One is improving the SVC coding efficiency at the expense of increasing encoder complexity. In this direction, Schwarz and Wiegand [7] have studied the joint optimization for all layers, yielding the best performance of SVC (i.e., +10% bit rate overhead for two-layer scenario against the H.264/AVC single layer encoder). However, this introduces even higher encoding complexity (i.e., multiple times of the reference software), making it impractical for coding structures with more than two layers. Another algorithm is later proposed by Li *et. al* [8], which has reduced complexity than [7] but is still more complex than the reference software, with about 7% encoding time increase. Another work [9] has tried to refine the Rate-distortion (R-D) model and Lagrangian multiplier λ, considering the distribution of inter-layer prediction residual in addition to the quantization parameter (QP). Relating λ as the function of QP and residual distribution has also been studied in [10] for single layer video coding.

Another avenue for improving SVC is to reduce the computational complexity for practical application purpose. This typically comes with the loss of the coding efficiency. Most of these works, such as [11]–[17], exploit the inter-layer mode correlation to reduce the mode candidates at the EL. The mode decision algorithm proposed in [18] relies on multiple thresholds rather than the R-D optimization approach. Note that all

Manuscript received January 21, 2015; revised May 28, 2015 and July 24, 2015; accepted July 25, 2015. Date of publication July 29, 2015; date of current version August 14, 2015. This work was supported by the National Science Foundation for Young Scholar of Jiangsu Province, China, under Grant BK20140610, and in part by the CCF-Tencent Open Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joan Serra-Sagrista. *(Corresponding author: Z. Ma.)*

M. Xu and Y. Wang are with the Department of Electrical and Computer Engineering, Polytechnic School of Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: mxu02@students.poly.edu; yw523@nyu.edu).

Z. Ma is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: mazhan@nju.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2462747

these methods try to reduce the complexity at ELs, without considering the mode decision process jointly for all the layers.

Different from all prior works, we *accomplish both complexity reduction and coding efficiency improvement* through the following innovative ideas:

- *Multilayer Mode Inference (MMI):* We perform the one-pass motion estimation and mode decision (1-MEMD) at one layer only, and force an upper layer to inherit the mode and motion (including partitioning) information from the lower layer. This not only bypasses the computationally intensive motion estimation and mode decision at upper layers, but also reduces the bit rate for signaling the motion and mode information. Typically 1-MEMD is performed at the base layer. However, if the BL satisfies the early skip/direct (ESD) threshold, 1-MEMD will be done at a higher layer, as explained below.
- *One-Pass Motion Estimation and Mode Decision Using Reference Decoupling and Lagrangian Multiplier Refinement:* With the goal to make the motion and mode decision at the base layer to be near optimal for all the layers, we use the reconstructed reference frame at the finest layer to perform motion estimation at the base layer. However, to avoid potential error drift for CGS, we use the reconstructed reference frame at the current layer to perform motion compensation during encoding. Furthermore, we set the Lagrangian multipliers for motion and mode decision at the base layer based on the QPs of the highest layer and the current layer, respectively, to provide a good balance between the R-D performances of different layers.
- *Early Skip/Direct Decision (ESD):* We also implement a fixed-threshold-based early Skip/Direct decision scheme to further improve the SVC encoder throughput with negligible coding efficiency loss. At the base layer, we perform 1-MEMD only if a block does not satisfy the ESD threshold. At an upper layer, we perform 1-MEMD only if a block does not satisfy the ESD threshold and the lower layer is coded using the ESD mode. In all other cases, the current layer inherits the motion and mode decision of the lower layer. This way, 1-MEMD is performed at most once over all layers. In addition, we only examine the inter 16×16 modes at the upper layer, because we have found that 16×16 modes are usually chosen when the lower layer is coded using the ESD mode.

Simulation results demonstrate that our algorithm significantly reduces the encoder complexity, while achieving slightly higher coding efficiency on average over a set of test sequences, for both CGS and MGS structures, under different resolutions. For instance, given a three-layer encoding structure, almost 5× complexity reduction and more than 10% BD-Rate improvement has been reported for Akiyo sequence coded using MGS, while almost 2× complexity reduction and over 1% BD-Rate gain for Football sequence using CGS. Akiyo and Football are typical video contents with stationary and motion intensive characteristics, respectively.

Preliminary results of this work was presented in [19]. It is extended with the consideration of the Lagrangian multiplier refinement for coding efficiency improvement while keeping the same magnitude of encoder complexity reduction. Meanwhile, MGS coding structure is also evaluated with the different strategy for *key* and *non-key* frames, which is not studied in [19]. Only CIF resolution videos were evaluated in [19], while we include the evaluation results for other HD videos as well in this version. Moreover, two sets of simulations are performed with different SVC encoder settings, i.e., one is the low-complexity SVC encoder without using multiple reference frames, small coding blocks, adaptive inter-layer motion prediction, etc, targeting for the practical real-time applications; while the other is the default SVC encoder with all tools enabled for high-efficiency compression demonstration. Extensive experiments further evident that our proposed method is generally applicable to different video contents and various encoder configurations.

This paper is organized as follows: Sec. II briefly reviews the relevant R-D optimized mode decision algorithm in the conventional multilayer SVC encoder. The proposed one-pass motion estimation and mode decision algorithm is presented in Sec. III, followed by the performance evaluation and discussions in Sec. IV. Finally, Sec. V discusses future research direction.

II. Mode Decision of Conventional SVC Encoder

The R-D optimized mode decision algorithm [20] is applied at each layer of SVC for every MB to search the best mode m^* from all available candidate modes ms with the best trade-off between the bitrate and distortion, i.e.,

$$m^* = \arg\min_{m} J\left(m; f, \hat{f}, \lambda\right), \tag{1}$$

with R-D cost function J defined as

$$J\left(m; f, \hat{f}, \lambda\right) = D\left(f, \bar{f}\left(m, \hat{f}\right)\right) + \lambda R\left(m, f - f_p\left(m, \hat{f}\right)\right), \tag{2}$$

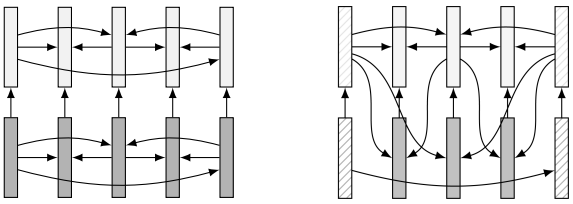
where f is the original signal, $\bar{f}(m, \hat{f})$ and $f_p(m, \hat{f})$ are the reconstructed and predicted signal using reference frame $\hat{f}$ and mode candidate m, respectively, and R is the rate to code the mode m and the corresponding residual $(f - f_p(m, \hat{f}))$. The Lagrangian multiplier λ is a function of the quantization parameter (QP) at the current layer [21]. In the JSVM encoder implementation [22], it is given by $\lambda\,(\mathrm{QP}) = 0.85 \times 2^{\frac{\mathrm{QP}-12}{3}}$.

For the Inter-mode with a given partition, one has to determine the best motion vector (MV) for each sub-block from a list of candidates. Similar to the mode decision process, a cost function J_{MV} is defined to select the best MV for each partition through R-D optimization, i.e.,

$$v^* = \arg\min_{v} J_{\mathrm{MV}}\left(v; f, \hat{f}, \lambda_{\mathrm{MV}}\right), \tag{3}$$

with J_{MV} equal to

$$J_{\mathrm{MV}}\left(v; f, \hat{f}, \lambda_{\mathrm{MV}}\right) = D_{\mathrm{MV}}\left(f, \hat{f}\,(v)\right) + \lambda_{\mathrm{MV}} R\,(v), \tag{4}$$



(a) CGS (b) MGS

Fig. 1. CGS and MGS coding structures for two quality layers.

where $\hat{f}(v)$ is the compensated block using the MV candidate v with $\hat{f}$ as the reference signal. $D_{\mathrm{MV}}(f, \hat{f}(v))$ and $R(v)$ are the motion compensation error and rate to encode v, respectively. The Lagrangian multiplier λ_{MV} depends on the distortion criterion used by D_{MV}. In the case that D is measured in terms of sum of squared error (SSE) $\lambda_{\mathrm{MV}} = \lambda$, while $\lambda_{\mathrm{MV}} = \sqrt{\lambda}$ if D_{MV} is measured by sum of absolute difference (SAD).

For the Intra-mode, there are also multiple predictions from the spatial neighbors with various angular and non-angular directions, associated with 16×16, 8×8 and 4×4 prediction partitions. The best Intra-Direction (DIR) a^* is determined similarly, by minimizing J_{DIR}, i.e.,

$$a^* = \arg\min_{\tilde{a}} J_{\mathrm{DIR}}\left(a; f, \tilde{f}, \lambda_{\mathrm{DIR}}\right), \tag{5}$$

with

$$J_{\mathrm{DIR}}\left(a; f, \tilde{f}, \lambda_{\mathrm{DIR}}\right) = D_{\mathrm{DIR}}\left(f, \bar{f}\left(a, \tilde{f}\right)\right) + \lambda_{\mathrm{DIR}} R\left(a, f - f_p(a, \tilde{f})\right), \tag{6}$$

where $\tilde{f}$ stands for the previously reconstructed signal in neighbor blocks which is used as the reference to derive the predicted block $f_p(a, \tilde{f})$ with intra direction candidate a, $\bar{f}(a, \tilde{f})$ is the reconstructed signal, and λ_{DIR} is determined in the same fashion as λ_{MV}.

Optimal MV v^* and DIR a^* derived from Eqs. (3) and (5) are used to reach the optimal mode m^* via (1) for final encoding. It is worth to note that optimal mode, MV or DIR heavily relies on the reference signal fidelity and Lagrangian multiplier. CGS and MGS use different reconstructed reference frame, yielding quite different coding performance.

A. Conventional Mode Decision Method for CGS

As illustrated in Fig. 1(a), CGS has separate encoding flows at each layer. In the conventional SVC encoder such as JSVM, the R-D optimized motion estimation and mode decision algorithm is applied in all layers [22], using the previously decoded frame at the current layer as the reference frame. It also uses the Lagrangian multiplier determined from the QP used in the current layer. Specifically, the best mode m_i^*, motion vector v_i^* and intra direction a_i^* at i-th layer are determined by (1), (3), and (5) with

$$\hat{f} = \hat{f}_i, \quad \lambda = \lambda\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{MV}} = \lambda_{\mathrm{MV}}\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{DIR}} = \lambda_{\mathrm{DIR}}\,(\mathrm{QP}_i), \tag{7}$$

where QP_i denotes the QP used in i-th layer.

B. Conventional Mode Decision Method for MGS

Compared with CGS, MGS adopts a quite different coding structure, as shown in Fig. 1(b). Since the MGS uses the reference frame from the higher enhancement layer (typically the highest layer is chosen) with better reconstructed quality, MGS can noticeably improve the coding efficiency. However, potential error drift might be introduced and propagated due to the packet/slice loss at higher enhancement layer. To control the error drift due to the possible data missing in the higher layers, key frames are used to code I- or P-frames at the group-of-pictures (GOP) boundaries. With referencing only from the BL, key frames are immune to the loss in higher layers, thus the error drift is confined within the erroneous GOP.

For the key frames, the best mode, motion vector and intra direction at i-th layer are determined using

$$\hat{f} = \hat{f}_0, \quad \lambda = \lambda\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{MV}} = \lambda_{\mathrm{MV}}\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{DIR}} = \lambda_{\mathrm{DIR}}\,(\mathrm{QP}_i). \tag{8}$$

For non-key frames, the highest layer is used as reference for both prediction and reconstruction. In a L-layer structure, the best mode, motion vector and intra direction at i-th layer are determined by

$$\hat{f} = \hat{f}_L, \quad \lambda = \lambda\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{MV}} = \lambda_{\mathrm{MV}}\,(\mathrm{QP}_i), \quad \lambda_{\mathrm{DIR}} = \lambda_{\mathrm{DIR}}\,(\mathrm{QP}_i). \tag{9}$$

Different from Inter-mode, Intra-mode decision strictly refers the spatial neighbor blocks at current layer and current frame. Lagrangian multiplier uses the current layer QP to guarantee the best reconstructed quality of Intra-mode coded block. This is kept without change in our proposed algorithm as well.

III. Low-Complexity Multilayer Mode Decision

Our work tries to reduce the encoder complexity while potentially improving the coding efficiency. Generally, there are three factors that are closely related to the SVC coding efficiency and complexity, i.e., reconstructed reference signal fidelity, Lagrangian multiplier and the number of mode candidates. It is expected that coding efficiency could be improved and complexity could be also reduced if we can carefully adjust these three impact factors. To fulfill this purpose, we have proposed the multilayer mode inference, reference signal decoupling and Lagrangian multiplier refinement, and the early skip/direct decision as follows.

A. Multilayer Mode Inference

After the best mode and motion information are determined at the BL (or more generally a lower layer) using the approached described in Sec. III-B, we propose to carry it over to all higher layers. This significantly reduces the complexity by removing the need for motion and mode decision at higher layers, and in the meantime reduces the overhead required for mode and motion information signaling.

Towards this goal, we make use of the macroblock type called *MB_Inferred*, which is defined in the SVC standard for inter-layer mode derivation (with the syntax *base mode flag*

set to 1), to force the mode reuse among layers. With this approach,

- If the lower layer collocated MB is Inter-coded, then the current layer MB is also coded in the Inter-mode (noted as the BLSkip mode), where the MB partition and the corresponding MVs are derived from the lower layer.
- If the lower layer collocated MB is Intra-coded, the current layer MB is coded in the IntraBL mode, which uses the reconstructed lower layer collocated MB as prediction.

Note that when the BL uses an Intra-mode, we do not force the EL to also use the same Intra-prediction mode. Rather, we choose to use the IntraBL mode, because it is simple yet efficient to provide decent coding efficiency.[1] On the other hand, the EL is forced to use the same Inter-prediction when the BL is coded in the Inter-mode.

B. Reference Decoupling and Lagrangian Multiplier Refinement

For a given MB, we propose to conduct motion estimation and mode decision at the base layer only so that the chosen mode and motion (if the chosen mode is Inter) are nearly optimal for all layers. We call this one-pass motion estimation and mode decision (1-MEMD). When a BL block satisfies the ESD threshold, 1-MEMD is conducted at an intermediate layer. In this section, we discuss how to determine motion and mode at a given layer i so that the result is near optimal for the current as well as all upper layers.

Intuitively, motion estimation will be more accurate by using the reconstructed frame with better quality (or less quantization noise). This was the rationale for the development of the MGS approach, which uses the reference pictures at higher EL (or just the highest EL) with less quantization noise to perform motion estimation and compensation. However, such an approach can introduce the error drift problem if the packets are lost at EL. Therefore, *key frame* is set to bound the drift error within a GOP.

Aiming at using the reference frame with better fidelity to improve the motion estimation accuracy, we decide to decouple the reference signal for motion estimation $\hat{f}_{\mathrm{ME}}$ and compensation $\hat{f}_{\mathrm{MCP}}$ in the CGS mode. In other words, we will use the reconstructed frame at the highest EL to do motion search, but use the reconstructed frame at the current layer for motion compensation. In this way, CGS could be supported easily. For MGS, we will follow the standard specification where both estimation and compensation is conducted using the reconstructed frame at the highest EL and error drift is controlled by key frames.

In addition to the reference decoupling, the Lagrangian multipliers used for R-D optimized motion estimation and mode decision are also modified from the conventional approach described in Sec. II, so that the chosen mode and motion well balance the R-D performance among different layers.

Specifically, for an L-layer CGS encoder, for motion estimation, we employ the reconstructed frame at the highest layer as the reference frame and use the Lagrangian multiplier corresponding to QP of the highest layer, for the R-D cost, so that the resulting motion vector is more close to the true motion. However, for mode decision, we use the Lagrangian multiplier associated with the current layer QP for the R-D cost. That is,

$$\hat{f}_{\mathrm{ME},i} = \hat{f}_L, \quad \lambda_{\mathrm{MV},i} = \lambda_{\mathrm{MV}}(\mathrm{QP}_L), \\ \lambda_i = \lambda(\mathrm{QP}_i), \quad \lambda_{\mathrm{DIR},i} = \lambda_{\mathrm{DIR}}(\mathrm{QP}_i), \tag{10}$$

with $\hat{f}_{\mathrm{ME},i}$ representing the reference signal for motion estimation at i-th layer.

For mode decision, we have experimented with using $\lambda(\mathrm{QP}_j), j = i, i+1, \cdots, L$, for evaluating the R-D cost. We have found that using $\lambda(\mathrm{QP}_i)$ would yield a good trade-off among the coding efficiency for different layers. Moreover, after obtaining the optimal motion and mode information via (10), we apply motion compensation using reference signal from current layer, i.e., $\hat{f}_{\mathrm{MCP},i} = \hat{f}_i$, to derive the predicted frame for encoding to avoid error drift at decoder if the packets are lost at ELs.

MGS shares almost the same strategy shown in (10) as described above for CGS. The differences lie in the reference signals used for motion compensation. Specifically, for key frames, we use $\hat{f}_{\mathrm{MCP},i} = \hat{f}_0$, and for non-key frames, we use $\hat{f}_{\mathrm{MCP},i} = \hat{f}_L$.

C. Early Skip/Direct Mode Decision

In spite that various fast motion estimation algorithms have been developed (for example, the TZ-Search in JSVM encoder [5]), motion search still dominates complexity in the encoder. With the multilayer mode inference, EL complexity has been reduced dramatically, however BL still demands computationally intensive motion search. A low-complexity early skip mode decision was introduced by Jeon and Lee [23] and has brought vast interest over the past years, where the mode decision terminates when the skip mode decision meets the certain conditions at early stage. Numerous early skip conditions have been designed for the H.264/AVC. In [24], the motion filed is analyzed, and a statistic model is proposed to guide the mode selection. In [25], the Lagrangian multiplier is modeled to assist the early skip decision. In [26], the temporal correlation between frames is utilized in the early skip threshold derivation. Saha *et al.* [27] present three methods for the early skip decision, using the ρ-domain rate model, the spatial-temporal prediction, and the restricted reference frame. For quality scalability in SVC, the early skip is also studied in [28], where the lower layer information is used to assist the early skip decision at CGS ELs. These approaches rely on either multiple thresholds, or multiple motion compensations for comparison, and some also require the storage of historical data.

We propose a simple yet effective early Skip/Direct (ESD) mode decision scheme by extending the early termination

[1]Note that we in fact compared the two possibilities: One is to let EL inherit the same intra-mode partition and directions used at the BL; another is to use IntraBL. Because SVC does not have an existing mode that allows the inheritance of the intra partition and directions, with the first option, we have to repeat such side information at the EL layer that could harm the coding efficiency. Thus we select the second option.

technique to include the Direct mode. The proposed method uses fixed thresholds for deciding on the Skip and Direct mode. Note that this method is generally applicable for mode decision at a single layer coder (e.g. H.264/AVC) as well as all layers in SVC.

1) ESD Mode: Among all the Inter-modes in H.264/AVC and SVC, Skip and Direct modes are two special cases.

- Skip mode is available in both P- and B-frames (noted as P_SKIP and B_SKIP), using the predictive motion vector (PMV) as MV, and the partition size is always 16×16 and motion compensation residual is not coded..
- Direct mode is similar to Skip mode except that it is available only in B-frames (noted as B_DIRECT), and the residual is quantized and coded [29].

We extend the concept of Direct mode to P-frames by allowing a P_DIRECT mode, coded using the syntax of the Inter 16×16 mode, where its PMV is used as actual MV, and it has non-zero quantized residual. In RD-optimized mode decision, the skip and direct mode are treated in the same way as other modes, skip or direct is chosen only if it provides lowest Lagrangian cost. We propose an early decision approach to determine whether a MB should be coded using the skip or direct mode without going through motion estimation and coding of residuals. The mode chosen at this early stage is marked as early Skip/Direct (ESD) mode internally in the encoder.

Intuitively, for a Skip mode coded block, its PMV has to be very accurate, and the prediction error is small and homogeneous. Similarly, Direct mode implies that the PMV is probably accurate, but the prediction error is not negligible. The homogeneity of residual can be guaranteed by examining the 8×8 sub-blocks, which is also shown to be effective in early skip decision [26].

Let D_l and D_c stand for the prediction error in luma and chroma components respectively for an 8×8 sub-block using PMV. If D_l of all the sub-blocks is less than the thresholds $T_{1,\text{Luma}}$ and D_c is less than the $T_{1,\text{Chroma}}$, Skip mode is applied for the current MB. If the Skip mode criterion is not satisfied, but D_l is still below a more relaxed threshold T_2, then the Direct mode is selected. Otherwise, the R-D optimized mode decision is performed. The thresholds are QP dependent, as discussed below. The proposed ESD mode decision is illustrated in Fig. 2, where the *ESD flag* indicates whether the ESD conditions are satisfied. Note that in case that the ESD condition is not satisfied in any one of the 8×8 sub-block, it is unnecessary to check D_l and D_c for the remaining sub-blocks.

2) ESD Threshold Derivation: Intuitively, the ESD thresholds should be similar to the expected quantization error associated with the QP used. If the prediction error D_l and D_c is already less than the quantization error, the error blocks are likely to be quantized to all zeros after going through transform and quantization. Therefore, we propose to use the averaged quantization error among all training blocks coded using non-skip mode at the QP of the current layer as the early skip threshold, i.e., $T_{1,\text{Luma}}(\text{QP}) = \bar{e}_{q(\text{QP})}(\text{QP})$. In H.264/AVC, quantization stepsize q is a monotonic function of QP, i.e., $q = 2^{\frac{\text{QP}-4}{6}}$ [1].

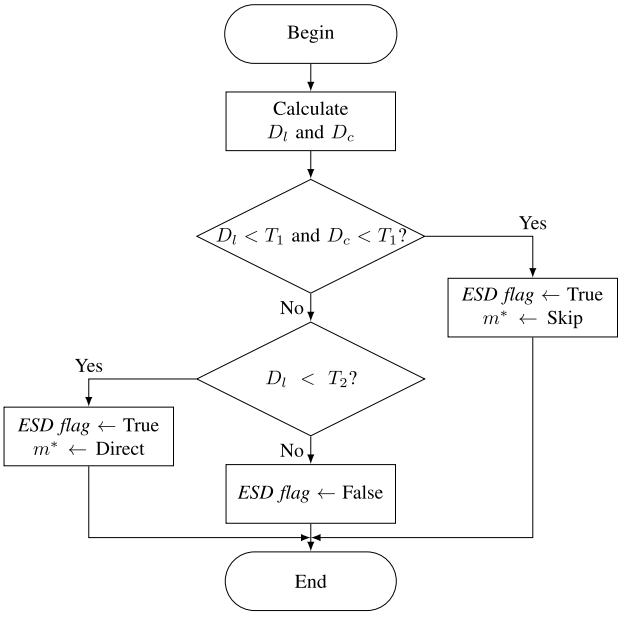


Fig. 2. ESD mode decision, where D_l and D_c are the prediction error in luma and chroma component using PMV, respectively. T_1 and T_2 are dependent on the QP and determined using the method described in Sec. III-C2.

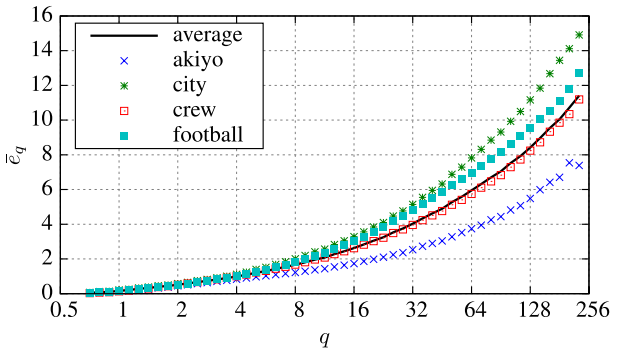


Fig. 3. Quantization error $\bar{e}_q$ in luminance v.s. quantization stepsize q for four test sequences. The quantization error is measured using SAD per pixel, and averaged over all blocks coded with non-Skip modes.

Figure 3 shows the relationship between luma $\bar{e}_q$ and q for four test CIF sequences coded by JSVM using single layer encoding configuration. All the data points follow the same trend, despite slight variations among the sequences. Although the individual threshold could be chosen for each sequence, our experiments show that averaged $\bar{e}_q$ over the four test sequences serves pretty well not only for these four sequences, but also other CIF and HD sequences. As will be shown in later sections, the fixed threshold could give a good trade-off between the coding efficiency loss and complexity reduction. Figure 3 shows the case using the SAD to measure the prediction error. We have found that using the sum of squared difference (SSD) gives a very similar trend. Therefore we choose SAD as the error metric, as it requires less computation.

By using the default JSVM configurations, only the luma component is used in the motion search, therefore a small prediction error in luma does not necessarily imply a small

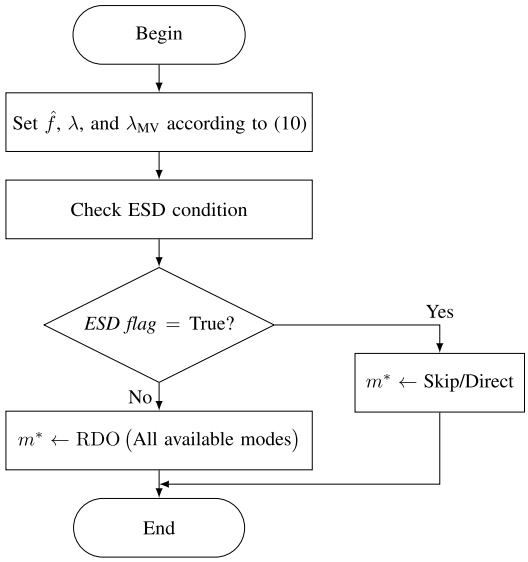


Fig. 4. BL mode decision with ESD enabled.

error in chroma components. The $\bar{e}_q$ for chroma components are also collected in our experiment, and approximately equal to half of the averaged luma quantization error. However, since human eyes are less sensitive in chrominance, we apply $T_{1,\text{Luma}}$ for the chroma components as well, and denote it as T_1. For Direct mode, we only check luminance error against a threshold T_2 to ensure the accuracy of PMV. Our experiment shows that choosing $T_2 = 1.2T_1$ yields less than 0.5% BD-Rate [30] increase compared with the case where there is no early Direct mode decision, but with noticeable encoder complexity reduction (i.e., more than 5%).

3) Multilayer Mode Inference With ESD Mode: Figure 4 shows the BL mode decision with ESD enabled. The ESD condition is checked first, and the conventional modes are examined only if the ESD condition is not satisfied.

When a block at the BL is coded using ESD mode, its MV (reused from PMV) may not reflect the actual motion, because ME is completely bypassed. If such MV is carried to the higher layers, there is no guarantee that this MV remains near-optimal for ELs. To resolve this issue, we perform the motion search at the EL, but with the following constraints:

- ME is conducted only if the current layer MB does not satisfy the ESD theshold, and if ME has not been performed in lower layers (i.e., the lower layer has a ESD flag).
- ME is conducted at the 16×16 block size only. This is because, when the lower layer satisfies the ESD condition, the entire 16×16 block can usually be predicted well with a single MV.

With this approach, MEMD does not necessarily take place at the BL. However, MEMD is conducted at most one time, either at BL or some EL. Motion estimation and mode decision, as well as the motion compensation are performed following the approach discussed in Sec. III-B. Figure 5 illustrates the mode decision at EL when ESD is enabled.

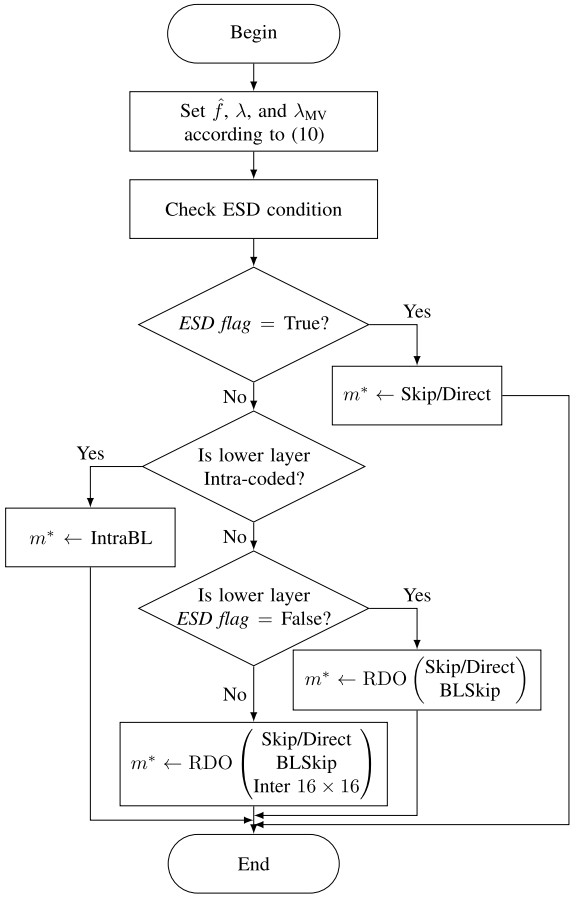


Fig. 5. EL mode decision with ESD mode inference.

IV. Experiments and Discussion

A. Simulation Configurations

Seven test sequences with CIF resolution at 30 fps (frame per second), three with 720p resolution at 50 fps and three with 720p resolution at 60 fps are encoded with three CGS and MGS layers respectively, using JSVM 9.19.15 software [5] (and its modification) implemented with the proposed algorithm. The hierarchical B-structure with GOP length of 16 for CGS and 8 for MGS is used with Intra-picture period of 64. The QP difference between adjacent layers is fixed to be 6. For each sequence, the base-layer QPs are chosen to cover a wide range while providing reasonable perceptual quality (mostly with resulting Peak Signal-to-Noise Ratio [PSNR] from 30 dB to 40 dB approximately), as detailed in Table I. For both encoders, SAD is used as the error measurement metric for RD-optimized motion estimation and mode decision, and CABAC is used for entropy coding.

The simulations are conducted with two encoding configurations, i.e., low-complexity settings targeting at practical real-time applications as our primary focus, and the default SVC test conditions with all SVC coding tools enabled as reference. Low-complexity encoder configuration is first introduced in the following paragraphs.

For the motion search, we use only one reference frame for both encoders, as enabling multiple reference frame almost increases the encoder complexity multiple times, and therefore

TABLE I
QP CONFIGURATION FOR DIFFERENT CONTENT

Resolution	Sequence	Frames	Layer	QP			
CIF@30fps	akiyo	289	0	36	40	44	48
			1	30	34	38	42
	ice	225	2	24	28	32	36
	city, crew, foreman	289	0	30	34	38	42
			1	24	28	32	36
			2	18	22	26	30
	football, waterfall	257	0	30	34	38	42
			1	24	28	32	36
			2	18	22	26	30
720p@50fps	mobcal, parkrun	497	0	30	34	38	42
			1	24	28	32	36
	shields	289	2	18	22	26	30
720p@60fps	FourPeople, Johnny, KrishtenAndSara	600	0	36	40	44	48
			1	30	34	38	42
			2	24	28	32	36

is seldom applied in low-complexity encoders for real-time applications.

Even though H.264/AVC and SVC support the block partition size for Inter-mode from 16×16 down to 4×4, according to our experiments, we have noticed that coding efficiency is degraded less than 1% (in terms of BD-Rate) by disabling block size less than 8×8 in Inter-modes, but with quite significant 25% encoder complexity reduction compared with the default JSVM encoding. This is also confirmed during the High-efficiency video coding (HEVC) standardization that smaller block size (less than 8×8) does not provide significant coding efficiency improvement for Inter-frames but with dramatic overhead for memory access and computing. Hence, 4×4 block based motion compensation is not used in HEVC [31]. In the low-complexity settings, we also do not consider the block partitions smaller than 8×8 in Inter-modes, in both original JSVM and the one implemented with the proposed algorithm.

For the inter-layer prediction tools in SVC, adaptive residual prediction is enabled for both encoders. The adaptive inter-layer motion prediction (ILMP) is enabled in a fully R-D optimized encoder; however in the low-complexity encoder profile, we have enforced ILMP for both encoders, i.e., the PMV at the EL is always from the lower layer. This is because from our experiments, we have noticed that the coding efficiency gain brought by adaptive ILMP is marginal, which is also consistent with the results reported by Li *et. al* [32].

With these modifications to the default JSVM targeting for the practical real-time SVC implementation, we note this anchor reference as low-complexity JSVM (i.e., LC-JSVM). Its simulation results are presented and discussed in Section IV-B. Additional performance evaluation using default JSVM with multiple reference pictures, small blocks, and adaptive ILMP, is shown in Section IV-C.

The experiments are conducted on a Linux computer server equipped with Intel Xeon (E5405@2.00GHz) processor and 8GB memory, running Ubuntu 12.04 server edition. Each individual encoding process is executed exclusively, without interfering with other running programs. The relative reduction of total encoding time ΔT (for all layers)[2] is defined as

$$\Delta T = \frac{T_{\text{JSVM}} - T_{\text{Prop}}}{T_{\text{JSVM}}} \times 100\%, \tag{11}$$

averaged over all the QPs, where T_{JSVM} and T_{Prop} are the total encoding time for the default JSVM and the proposed low-complexity algorithm, respectively, and measured using the timing function provided by the operating system. ΔT_m is the reduction of time in mode decision (including the motion search) at each layer, which is derived in the similar manner. For each block, the time consumed in mode decision could be lower than the minimum precision provided by the system timing function, thus we measure the CPU cycle count, and convert it back to time using 2.00 GHz frequency (the CPU frequency is fixed to 2.00 GHz when running the simulations).

B. Performance Evaluation Using Low-Complexity SVC Encoder

Figure 6 shows the performance evaluation (i.e., R-D curve and complexity) for two typical CIF test sequences (Akiyo with stationary scene and Football with intensive motion) using CGS structure, with and without enabling ESD mode. As expected, with our proposed method, the mode decision time at the EL is reduced significantly and remains almost constant among different QPs. At the BL, the complexity reduction from ESD mode is quite noticeable, and the amount of time reduction is sequence dependent. It is noticed that in the default JSVM, the EL takes less time to encode than the BL. This is due to the forced ILMP together with the fast motion search, where the MV from the lower layer is used as PMV, resulting the motion search engine terminates at an early stage. More complexity saving is expected if adaptive ILMP is enabled or the fast motion search is disabled.

The complete simulation results for the CIF test sequences are detailed in Table II. Note that negative BD-Rate means the percentage of reduced rate for the same PSNR, compared to default JSVM. As the experiment results show, our low-complexity multilayer mode decision algorithm, even with ESD mode disabled, achieves an averaged 43.9% total time reduction for encoding all three layers, with averaged 0.7%,[3] 79.9%, and 83.5% time reduction for mode decision for the BL, EL #1 and #2, respectively. With ESD mode enabled in all the layers, the average total time reduction increases to 57.5%, with average 33.0%, 84.5%, and 87.0% mode decision time reduction in different layers.

Recall that the motion estimation is mainly conducted at the BL using the finest reconstruction as reference instead of from the current layer. This leads to averaged 0.8% and 1.6% BD-Rate loss at the BL with ESD disabled and

[2]Note that relative time measurement is a popular metric used by video coding industry (esp. HEVC standardization committee) to illustrate the encoder complexity variations. Many technical proposals are adopted by evaluating its coding efficiency, relative encoder/decoder running time, and other considerations for practical implementation.

[3]Note that due to the reference decoupling , the reference at BL has higher quality than in JSVM, therefore even when ESD is disabled, the fast motion estimation algorithm (e.g., TZ-search used by JSVM) could terminate at an early stage from our experiments and potentially reduce the time for motion estimation.

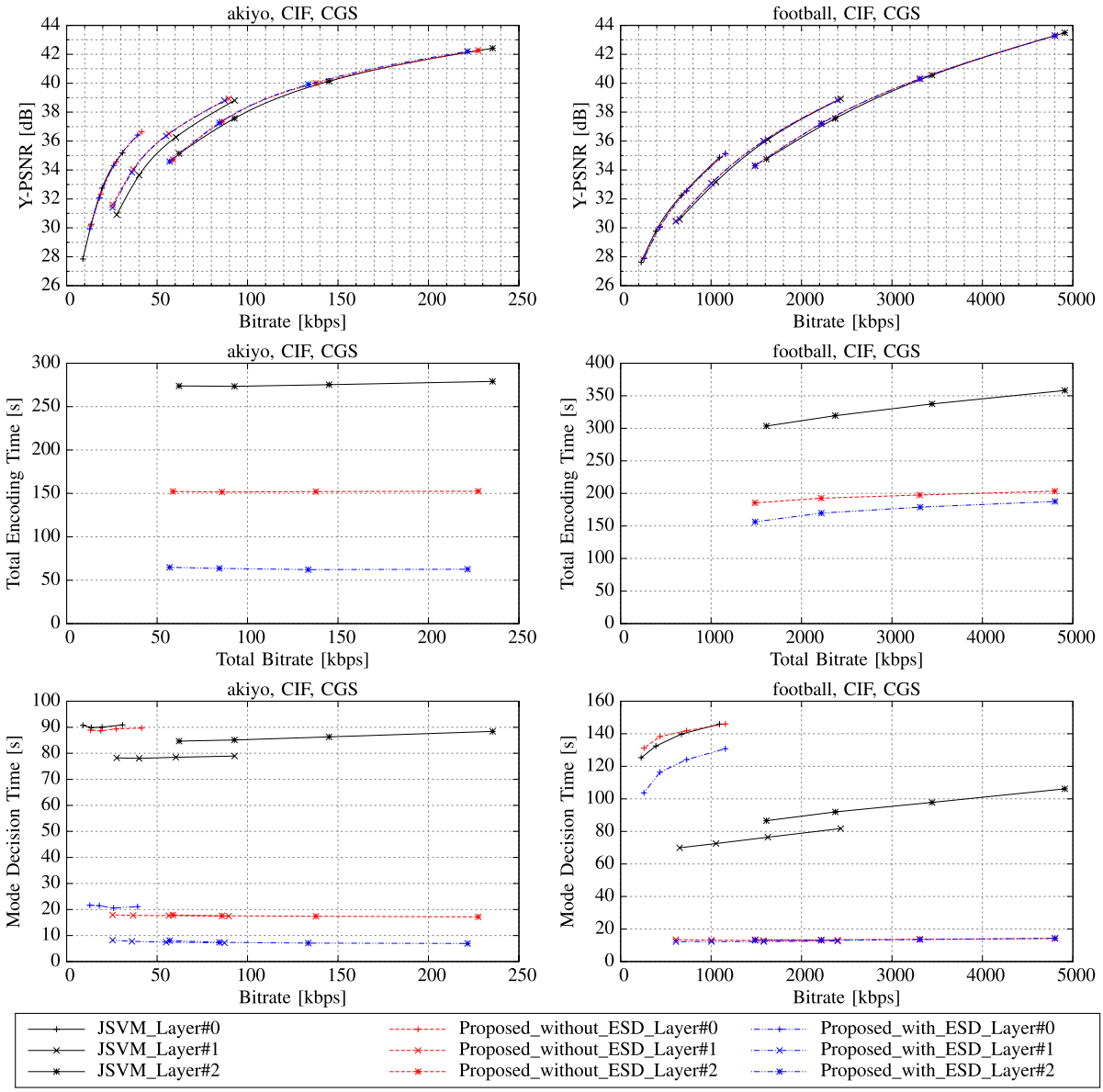


Fig. 6. Performance comparison of proposed algorithm v.s. default JSVM for sequences Akiyo and Football using CGS coding structure. The top row plots the R-D curve for each layer. The middle row shows the total bit rate v.s. the total encoding time of all three layers. The bottom row shows the time consumed in mode decision (including motion estimation) at each layer.

enabled, respectively. However, the higher layers benefit from the mode chosen at the BL and the inheritance of BL modes, resulting in averaged BD-Rate improvement of 5.8% and 5.1% at layer #1, and 1.1% and 0.8% at layer #2, with ESD mode disabled and enabled, respectively.

The performance evaluation for 720p test sequences are listed in Table III. With ESD mode disabled, the BD-Rate gains are -5.4%, -2.6%, and 1.9% averaged for each layer respectively, with average overall encoding time saving of 48.9%.[4] With ESD mode enabled, the coding efficiency drops (but still with gains), but the saving for overall encoding time increases to 68.4% on average.

[4]The tested 720p sequences have frame rate at 50Hz, which results in better motion search accuracy compared with 30Hz CIF test sequences. With reference decoupling, it leads to better performance at BL compared with CIF sequences.

Reported above are the simulations under the CGS coding structure. The simulation results using MGS coding structure for test sequences Akiyo and Football are demonstrated in Fig. 7. It is noticed that the proposed algorithm has significant coding gain over the LC-JSVM software for sequence Akiyo. Our investigation shows that this gain comes from our cross-layer mode decision in the MGS key frames. Since the sequences Akiyo has a stationary background, most of the bits are consumed in the key frames. While the JSVM encoder mode selection is optimized only for the current layer, our proposed cross-layer mode decision algorithm achieves near global optimality and provides much higher coding efficiency at the ELs.

Table IV and V list the complete performance evaluation for the CIF and 720p test sequences encoded using MGS structure. Since MGS already benefits from using the highest

TABLE II
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR CIF USING CGS ON TOP OF LC-JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
akiyo	0	1.2%		1.4%	0.6%		76.5%
	1	−11.6%	44.8%	77.5%	−11.0%	77.0%	90.2%
	2	−2.4%		79.6%	−3.4%		91.4%
city	0	1.0%		0.3%	1.4%		18.7%
	1	−4.4%	45.5%	79.9%	−4.2%	52.4%	81.2%
	2	0.1%		84.6%	0.2%		85.1%
crew	0	4.1%		2.4%	5.6%		22.5%
	1	−2.9%	42.2%	81.5%	−2.2%	51.4%	83.8%
	2	−0.7%		85.1%	−0.2%		86.0%
football	0	2.7%		−2.6%	3.6%		12.7%
	1	−2.5%	40.8%	82.4%	−2.1%	47.5%	83.5%
	2	−1.3%		85.6%	−1.2%		85.9%
foreman	0	1.5%		2.3%	2.9%		33.9%
	1	−5.1%	44.3%	80.1%	−4.2%	57.6%	84.1%
	2	−1.0%		84.3%	−0.8%		85.9%
ice	0	0.0%		1.9%	1.7%		51.3%
	1	−8.6%	42.1%	78.7%	−6.6%	66.2%	89.2%
	2	−2.9%		81.2%	−1.1%		90.2%
waterfall	0	−5.0%		−0.5%	−4.6%		15.1%
	1	−5.7%	47.8%	79.2%	−5.4%	52.8%	79.7%
	2	0.7%		84.0%	1.0%		85.2%
Average	0	0.8%		0.7%	1.6%		33.0%
	1	−5.8%	43.9%	79.9%	−5.1%	57.7%	84.5%
	2	−1.1%		83.5%	−0.8%		87.0%

TABLE III
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR 720p USING CGS ON TOP OF LC-JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
mobcal	0	−9.0%		0.3%	0.0%		63.7%
	1	−7.9%	46.8%	78.6%	−3.6%	70.8%	83.4%
	2	2.2%		83.4%	6.2%		87.8%
parkrun	0	−5.1%		−0.3%	−4.5%		27.5%
	1	3.9%	53.1%	81.4%	3.9%	61.4%	81.6%
	2	2.2%		86.6%	2.2%		87.0%
shields	0	−2.1%		0.3%	−0.2%		71.3%
	1	−3.9%	46.8%	78.9%	−4.5%	72.8%	81.4%
	2	1.4%		84.0%	1.4%		87.6%
Average	0	−5.4%		0.1%	−1.6%		54.2%
	1	−2.6%	48.9%	79.6%	−1.4%	68.4%	82.1%
	2	1.9%		84.7%	3.3%		87.5%

layer as reference, the coding efficiency of proposed method is expected to have less gain over the conventional R-D optimized method compared to that using CGS structure. Since the mode decision is tuned toward the EL, despite that the BL may suffer from coding efficiency loss, the EL still has gains in the coding efficiency for the CIF sequences. The complexity reduction for proposed method without ESD is similar to that in the CGS case. For the CIF sequences, with ESD mode supported, the BL benefits from more than 50% average mode decision time reduction. With the additional complexity reduction in the EL mode decision, the total encoding time for all three layers is reduced by 66.5% on average. For the 720p sequences, slight coding efficiency degradation is observed,[5] however the complexity saving follows the same trend as in the CGS case.

[5]This is due to reason that Lagrangian multiplier is not tuned for the middle layer, thus the selected mode is suboptimal. The MGS structure also affects the Skip mode decision, thus T_1 and T_2 derived from single layer coded CIF sequences may not fit very well.

TABLE IV
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR CIF USING MGS ON TOP OF LC-JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
akiyo	0	12.9%		−1.0%	13.3%		89.8%
	1	−26.7%	44.1%	75.7%	−22.9%	81.5%	92.2%
	2	−16.9%		78.0%	−11.8%		89.3%
city	0	14.8%		−0.9%	16.9%		66.2%
	1	−1.6%	43.8%	77.5%	4.5%	69.5%	86.1%
	2	2.1%		81.2%	3.7%		81.2%
crew	0	−4.0%		−1.8%	−1.7%		30.6%
	1	0.5%	40.5%	79.9%	4.2%	54.9%	83.9%
	2	3.4%		83.1%	4.5%		83.9%
football	0	−3.3%		−1.6%	−1.7%		25.0%
	1	1.1%	40.6%	80.7%	3.8%	52.5%	84.7%
	2	2.6%		83.5%	3.1%		83.6%
foreman	0	6.3%		−1.3%	8.9%		53.7%
	1	−1.0%	42.5%	78.0%	4.9%	65.1%	86.0%
	2	2.3%		81.7%	4.2%		83.1%
ice	0	−4.8%		−1.4%	−2.7%		57.2%
	1	−6.9%	41.2%	77.1%	−2.7%	68.5%	89.1%
	2	−2.4%		79.5%	1.2%		88.8%
waterfall	0	9.5%		−0.7%	10.4%		77.2%
	1	−7.4%	46.2%	77.2%	−3.1%	73.6%	88.1%
	2	−1.0%		81.0%	0.4%		80.2%
Average	0	4.5%		−1.3%	6.2%		57.1%
	1	−6.0%	42.7%	78.0%	−1.6%	66.5%	87.2%
	2	−1.4%		81.2%	0.8%		84.3%

TABLE V
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR 720p USING MGS ON TOP OF LC-JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
mobcal	0	0.0%		−0.1%	5.9%		82.2%
	1	4.7%	45.7%	76.2%	13.4%	76.7%	83.9%
	2	2.6%		80.4%	6.2%		85.3%
parkrun	0	0.0%		−0.4%	0.5%		61.6%
	1	12.5%	51.3%	78.9%	14.9%	70.0%	80.1%
	2	2.3%		83.5%	2.7%		84.2%
shields	0	0.1%		−0.5%	3.7%		83.2%
	1	5.0%	45.6%	76.2%	10.2%	76.8%	83.1%
	2	1.0%		80.5%	2.7%		85.0%
Average	0	0.0%		−0.3%	3.4%		75.6%
	1	7.4%	47.5%	77.1%	12.8%	74.5%	82.4%
	2	2.0%		81.5%	3.9%		84.8%

C. Performance Evaluation Using Default SVC Encoder

We have demonstrated the efficiency of proposed algorithm using the low-complexity SVC encoder configuration targeting for the practical implementation, where only one reference frame is used, block size less than 8×8 is disabled for motion estimation and compensation, and ILMP is enforced at EL. In this section, we provide another set of experiments where anchor reference uses the default SVC encoder (i.e., multiple reference pictures, complete block size supported in motion estimation and compensation as well as the adaptive inter-layer motion compensation).

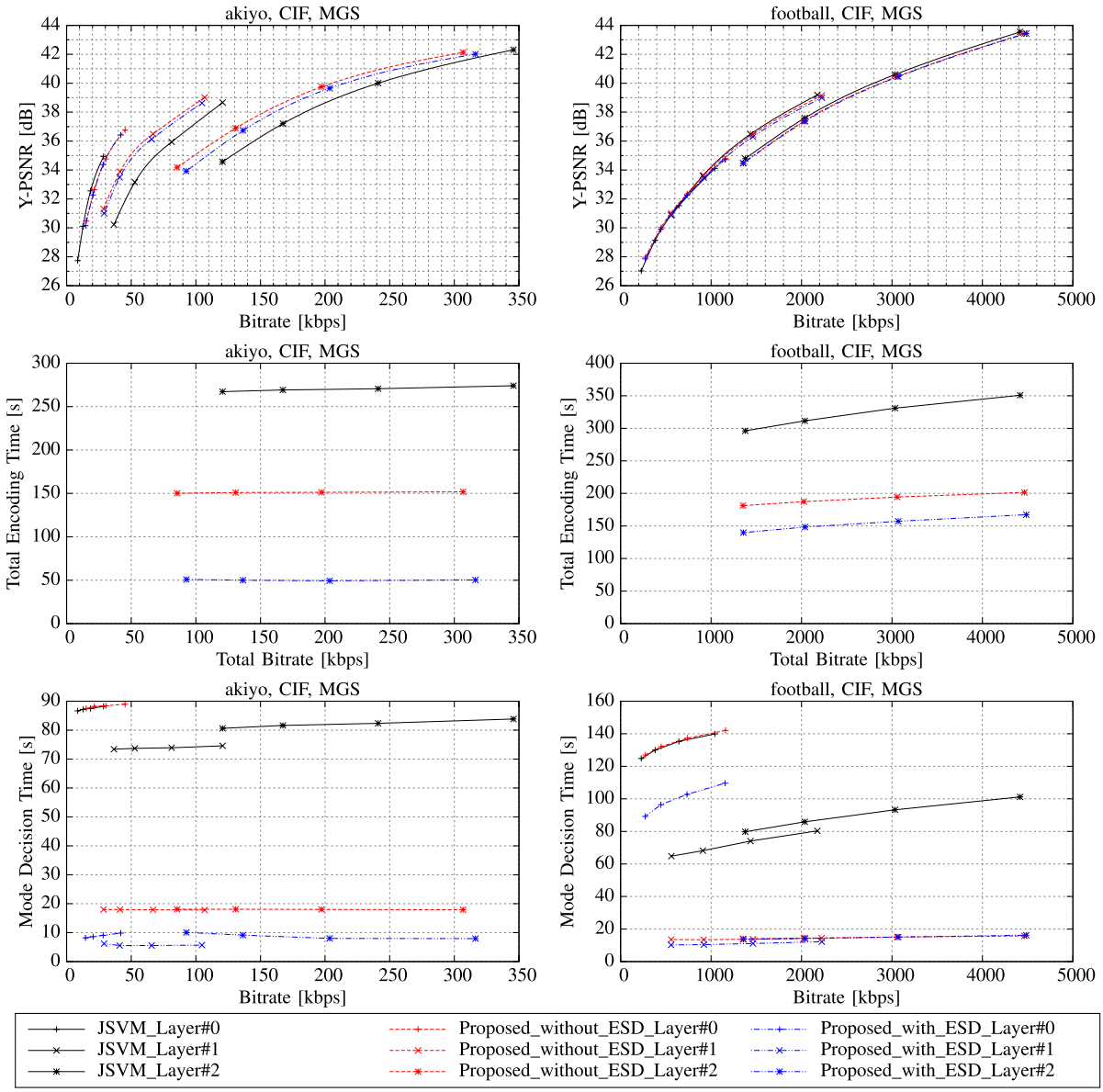


Fig. 7. Performance comparison of proposed algorithm v.s. default JSVM for sequences Akiyo and Football using MGS coding structure. The top row plots the R-D curve for each layer. The middle row shows the total bit rate v.s. the total encoding time of all three layers. The bottom row shows the time consumed in mode decision (including motion estimation) at each layer.

Except the anchor reference software, we keep other parameters the same without change, i.e., QPs, ESD threshold, etc. Meanwhile, we have included 720p 60Hz Class E sequences from the HEVC common test content to further verify the efficiency of our proposed scheme. Simulations are presented in Table VI, VII, VIII and IX. As we can see, similar performance is observed to not only speed up the multi-layer SVC encoding but also improve the coding efficiency for typical video contents (such as Akiyo, FourPeople, etc.). It is also noted that even for the newly introduced HEVC test sequences, our algorithm still demonstrates its efficiency. For instance, it reports averaged 85.9% time saving and −4.0% BD-Rate improvement for MGS encoder with ESD enabled. But we also observe the BD-Rate loss for other sequences, such as Crew and Football with complex texture and motion. This is due the reason that these sequences generally use different modes among layers (particular when enabling all features), but we enforce the mode inference between successive layers. Coding efficiency could be improved if we relax this condition by including other potential mode candidates, but of course the complexity reduction would be decreased. In general, our proposed algorithm is applicable to various contents and encoder settings according to the experiments performed in Section IV-B and IV-C.

D. Performance Comparison With Other SVC Mode Decision Algorithms

Compared with the existing low-complexity mode decision algorithms for SVC [11]–[17], our method achieved significantly greater complexity reduction. The time saving factors reported in these prior studies ranged from 30.5% [13] to 77.8% [12], whereas our method ranges from 47.5% (Football CGS) to 81.5% (Akiyo MGS). Note that most of these works reported the savings at EL only while our work

TABLE VI
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR CIF USING CGS ON TOP OF DEFAULT JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
akiyo	0	0.3%		−5.2%	1.7%		65.8%
	1	−10.4%	49.7%	78.1%	−9.3%	78.9%	92.3%
	2	−1.2%		79.4%	−3.0%		94.0%
city	0	1.4%		−8.2%	1.6%		2.7%
	1	−1.2%	56.4%	82.8%	−1.2%	60.3%	84.6%
	2	3.5%		88.7%	3.7%		89.6%
crew	0	11.0%		−12.2%	11.5%		12.0%
	1	3.3%	62.8%	88.5%	3.8%	70.0%	91.1%
	2	4.7%		93.9%	5.1%		94.9%
football	0	7.5%		−17.2%	7.7%		2.1%
	1	1.2%	63.6%	90.3%	1.3%	68.9%	91.9%
	2	1.9%		94.3%	2.0%		94.9%
foreman	0	6.6%		−8.7%	7.4%		23.7%
	1	−0.4%	57.7%	84.4%	0.1%	68.8%	89.3%
	2	3.5%		90.1%	4.1%		92.3%
ice	0	11.9%		−11.7%	12.5%		41.7%
	1	−0.7%	54.1%	82.8%	−0.2%	75.4%	93.4%
	2	3.5%		86.7%	4.3%		95.6%
waterfall	0	3.8%		−5.2%	3.8%		−2.1%
	1	−0.2%	55.7%	81.6%	0.0%	56.8%	81.9%
	2	2.9%		87.4%	3.0%		87.7%
Average	0	6.1%		−9.8%	6.6%		20.8%
	1	−1.2%	57.1%	84.1%	−0.8%	68.4%	89.2%
	2	2.7%		88.6%	2.7%		92.7%

TABLE VII
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR 720p USING CGS ON TOP OF DEFAULT JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
FourPeople	0	2.3%		8.6%	4.5%		56.7%
	1	−10.5%	49.7%	77.5%	−9.5%	76.3%	91.0%
	2	−1.7%		79.6%	−3.7%		92.8%
Johnny	0	−0.7%		−14.9%	1.3%		71.6%
	1	−15.5%	46.7%	75.6%	−14.6%	80.5%	92.3%
	2	−1.1%		77.5%	−5.2%		92.8%
KrishtenAndSara	0	2.4%		−14.4%	3.2%		65.0%
	1	−10.3%	47.9%	76.6%	−10.2%	80.2%	92.3%
	2	3.9%		79.3%	−0.3%		93.6%
Average	0	1.3%		−12.7%	3.0%		64.4%
	1	−12.2%	48.1%	76.5%	−11.5%	79.0%	91.9%
	2	0.4%		78.8%	−3.1%		93.1%

TABLE VIII
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR CIF USING MGS ON TOP OF DEFAULT JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
akiyo	0	1.2%		−11.0%	9.7%		86.6%
	1	0.1%	50.0%	71.1%	2.3%	81.0%	94.1%
	2	−2.9%		79.1%	−3.1%		90.0%
city	0	0.8%		−3.2%	1.8%		44.0%
	1	−0.1%	54.3%	79.7%	4.5%	62.7%	85.5%
	2	3.2%		83.0%	1.7%		74.4%
crew	0	2.7%		−12.7%	2.1%		−25.8%
	1	−1.1%	56.7%	85.1%	0.8%	50.4%	84.5%
	2	5.1%		88.9%	5.2%		86.6%
football	0	1.7%		−11.0%	2.0%		−34.2%
	1	−1.0%	58.2%	87.0%	1.2%	49.8%	88.3%
	2	2.2%		89.5%	2.3%		86.7%
foreman	0	2.2%		−8.5%	2.3%		21.9%
	1	−0.6%	53.7%	81.0%	1.2%	59.6%	86.0%
	2	3.8%		84.7%	1.3%		81.4%
ice	0	4.6%		−14.2%	6.1%		22.0%
	1	−2.9%	51.3%	80.5%	−0.3%	63.3%	89.7%
	2	2.8%		83.2%	4.0%		89.8%
waterfall	0	1.3%		−0.9%	0.2%		70.1%
	1	2.5%	55.0%	78.9%	6.2%	71.0%	89.4%
	2	1.6%		82.3%	0.0%		77.3%
Average	0	2.1%		−8.8%	3.5%		26.4%
	1	−0.5%	54.2%	81.3%	2.2%	62.5%	88.2%
	2	2.3%		84.4%	1.6%		83.7%

TABLE IX
PERFORMANCE EVALUATION OF PROPOSED ALGORITHM FOR 720p USING MGS ON TOP OF DEFAULT JSVM

Sequence	Layer	without ESD			with ESD		
		BD-Rate	ΔT	ΔT_m	BD-Rate	ΔT	ΔT_m
FourPeople	0	3.2%		−13.1%	7.3%		87.5%
	1	1.1%	48.2%	75.9%	1.7%	85.1%	95.3%
	2	−1.3%		78.9%	−3.1%		91.6%
Johnny	0	3.2%		−17.2%	8.9%		92.7%
	1	2.2%	46.5%	74.7%	2.3%	86.8%	95.9%
	2	−2.2%		77.8%	−5.6%		92.1%
KrishtenAndSara	0	4.7%		−23.9%	7.8%		89.1%
	1	2.1%	45.4%	74.6%	1.6%	85.7%	95.4%
	2	0.0%		78.2%	−3.3%		92.0%
Average	0	3.7%		−18.0%	8.0%		89.8%
	1	1.8%	46.7%	75.0%	1.9%	85.9%	95.5%
	2	−1.2%		78.3%	−4.0%		91.9%

presents the overall saving for all layers. As indicated in Tables II, III, IV, and IV, our encoder can consistently reduce the computation time by more than 80% for mode decision in each enhancement layer. Furthermore, all these prior works were tested on only two quality layers, and had BD-Rate loss ranging from 0.3% to 0.9% (except for [17], which considers intra mode only), whereas our method achieved slight gain in the coding efficiency for most test cases.

The reason that we can achieve significantly higher complexity reduction is because we perform MEMD only once, whereas the prior methods all perform MEMD at every layer. These prior works focused on how to reduce the complexity for performing MEMD at a particular layer, some by making use of the correlation between modes of adjacent layers.

For the Intra-mode decision at the EL, the proposed scheme only performs IntraBL mode, whereas [14] needs to perform nine Intra-predictions under 4×4 block size, plus the IntraBL mode. Other methods [11], [12], [16], [18] require even more computational resource to further check 16×16 block based predictions in addition to 4×4 block based prediction and IntraBL.

Compared with the well-known multilayer mode decision of SVC which achieves coding efficiency improvement with significant increase in complexity [7], [8], our method could significantly reduce the encoder complexity while slightly improving the coding efficiency. Instead of forcing all layers using the same mode, the work in [7] determines the modes for all layers simultaneously to improve the overall

R-D performance of enhancement layers with sacrifice at the base layer. Even though this can lead to more significant coding efficiency gain than our method, it requires significantly more computation time, thus cannot be practically extended to three-layer coding structure. A simplified method [8] requires roughly 7% of encoding time increase on average (with a two-layer structure) of the reference software.

Our solution is quite orthogonal with the scheme proposed in [9] and [18]. Those simplified R-D decision and improved λ derivation methods could be possibly used in our work to further improve the performance.

V. Concluding Remarks

In this work, we propose a novel low-complexity mode decision algorithm for multilayer quality scalable video coding. The core idea behind our proposed scheme is to perform motion estimation and mode decision only once at the base layer, and let the higher layers inherit the decisions made at the BL. Although this can significantly reduce the encoding complexity, it can lead to coding efficiency loss at higher layers if the base layer decision is made to optimize the coding efficiency of the base layer only. In order for the decision made at the BL to be nearly optimal for all layers, we use the highest layer reconstructed frame as the reference frame for temporal prediction and set the Lagrangian multiplier according to the QP of the current and higher layers. We also propose a simple early Skip/Direct decision method to further boost the encoding speed. Significant complexity reduction can be achieved because motion estimation is done at most once, and mode candidates are significantly reduced at enhancement layers. By forcing the EL to inherit the motion and mode information from the lower layer, we also reduce the signaling overhead for such information, which in turn lead to slight gain in the coding efficiency. The proposed scheme is quite different from prior works where ME is required at every layer.

Experiments have shown an average of more than $2\times$ (up-to $5\times$) speedup for a three-layer encoder against conventional rate-distortion optimized reference software JSVM for all test sequences, and targeting for the practical real-time applications. More specifically, average of $2\times$ and $3\times$ for both CIF and HD sequences coded using CGS, and $3\times$ and $4\times$ for MGS coded CIF and HD, respectively. Slight BD-Rate improvement could be also obtained in several typical sequences in the meantime, but the improvement is content dependent. As an example, more than $4\times$ complexity reduction and more than 3% BD-Rate improvement has been reported for Akiyo sequence coded using CGS and over $5\times$ complexity reduction with 12% BD-Rate gain for MGS, while almost $2\times$ complexity reduction and more than 1% BD-Rate gain for Football sequence using CGS, and $2\times$ complexity reduction with 3% BD-Rate loss for MGS. HD test sequences have more complexity reduction with a little worse coding efficiency than the CIF sequences.[6]

One of our future research focus is to extend the current algorithm to the scalable extension of HEVC (a.k.a. SHVC [3]). Since SHVC has quite substantial changes from the SVC, such as multi-loop decoding, recursive coding tree block, increased number of intra prediction candidates, merge mode, etc, the extension is not straight forward. But the principle behind the current algorithm is still applicable, for instance, leveraging the high mode correlation between successive layers to reduce the mode candidates at enhancement layer, re-using the motion information from lower layer to (at least) skip motion estimation extensively, etc.

[6]It is observed that HD sequences under MGS coding structure has some coding efficiency loss, which is because the Lagrangian multiplier and the ESD thresholds are not well tuned for this scenario.

Acknowledgment

The authors would like to thank anonymous reviewers for their comments to improve this manuscript.

References

[1] *Advanced Video Coding for Generic Audiovisual Services*, document Rec. ITU-T H.264, Apr. 2013.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[3] J. Chen, J. Boyce, Y. Ye, M. Hannuksela, G. J. Sullivan, and Y.-K. Wang, *HEVC Scalable Extensions (SHVC) Draft Text 7 (Separated Text)*, JCT-VC document R1008, Jul. 2014.

[4] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[5] *ITU-T ISO/IEC JTC 1, Reference Software for Scalable Video Coding*, ITU-T document Rec. H.264.2, 2009.

[6] D. Alfonso, M. Gherardi, A. Vitali, and F. Rovati, "Performance analysis of the scalable video coding standard," in *Proc. Packet Video*, Nov. 2007, pp. 243–252.

[7] H. Schwarz and T. Wiegand, "R-D optimized multi-layer encoder control for SVC," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2. Sep./Oct. 2007, pp. II-281–II-284.

[8] X. Li, P. Amon, A. Hutter, and A. Kaup, "One-pass multi-layer rate-distortion optimization for quality scalable video coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 637–640.

[9] L. Zhao, Y. Zhou, F. Wu, and M. Ai, "Inter-layer correlation considered R-D models and Lagrange multiplier for SVC MGS coding," *Signal, Image Video Process.*, vol. 8, no. 8, pp. 1581–1589, Dec. 2012.

[10] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[11] H. Li, Z. G. Li, and C. Wen, "Fast mode decision algorithm for inter-frame coding in fully scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 889–895, Jul. 2006.

[12] H.-C. Lin, W.-H. Peng, H.-M. Hang, and W.-J. Ho, "Layer-adaptive mode decision and motion search for scalable video coding with combined coarse granular scalability (CGS) and temporal scalability," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2. Sep./Oct. 2007, pp. II-289–II-292.

[13] B. Lee, M. Kim, S. Hahm, I.-J. Cho, and C. Park, "A low complexity encoding scheme for coarse grain scalable video coding," in *Proc. IET Int. Conf. Vis. Inf. Eng.*, Jul./Aug. 2008, pp. 753–758.

[14] C.-S. Park, B.-K. Dan, H. Choi, and S.-J. Ko, "A statistical approach for fast mode decision in scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1915–1920, Dec. 2009.

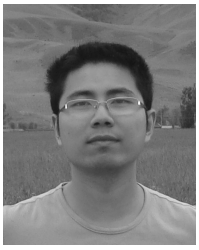
[15] C.-H. Yeh, K.-J. Fan, and G.-L. Li, "Fast mode decision algorithm for scalable video coding using Bayesian theorem detection and Markov process," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 563–574, Apr. 2010.

[16] Z. Deng, X. Cai, and Y. Cui, "Fast mode decision algorithm for inter-layer intra prediction in SVC," in *Proc. IEEE Broadband Netw. Multimedia Technol.*, Oct. 2011, pp. 212–216.

[17] B.-G. Kim, G.-S. Hong, and K.-W. Rim, "Fast mode decision algorithm for inter-frame coding in H. 264/AVC extended scalable video coding," *Int. J. Soft Comput.*, vol. 6, no. 4, pp. 102–110, 2011.

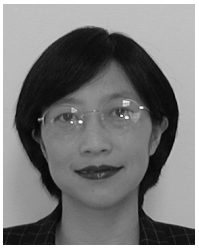
[18] S.-T. Kim, K. Konda, C.-S. Park, C.-S. Cho, and S.-J. Ko, "Fast mode decision algorithm for inter-layer coding in scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 55, no. 3, pp. 1572–1580, Aug. 2009.

[19] M. Xu and Y. Wang, "One-pass mode decision for low-complexity and high-efficiency encoding of quality scalable video," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2013, pp. 1–5.
[20] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.
[21] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.
[22] J. Reichel, H. Schwarz, and M. Wien, *Joint Scalable Video Model 11 (JSVM 11)*, document JVT-X202, Jul. 2007.
[23] B. Jeon and J. Lee, *Fast Mode Decision for H.264*, document JVT-J033, Dec. 2003.
[24] M. Bystrom, I. Richardson, and Y. Zhao, "Efficient mode selection for H.264 complexity reduction in a Bayesian framework," *Signal Process., Image Commun.*, vol. 23, no. 2, pp. 71–86, Feb. 2008.
[25] C. S. Kannangara *et al.*, "Low-complexity skip prediction for H.264 through Lagrangian cost estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 202–208, Feb. 2006.
[26] Y. V. Ivanov and C. J. Bleakley, "Skip prediction and early termination for fast mode decision in H.264/AVC," in *Proc. IEEE Int. Conf. Digit. Telecommun.*, Aug. 2006, pp. 1–7.
[27] A. Saha, K. Mallick, J. Mukherjee, and S. Sural, "SKIP prediction for fast rate distortion optimization in H.264," *IEEE Trans. Consum. Electron.*, vol. 53, no. 3, pp. 1153–1160, Aug. 2007.
[28] L. Shen, Y. Sun, Z. Liu, and Z. Zhang, "Efficient SKIP mode detection for coarse grain quality scalable video coding," *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 887–890, Oct. 2010.
[29] A. M. Tourapis, F. Wu, and S. Li, "Direct mode coding for bipredictive slices in the H.264 standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 119–126, Jan. 2005.
[30] T. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves (VCEG-M33)*, document ITU-T SG16 Q.6, Apr. 2001.
[31] *High Efficiency Video Coding*, document Rec. ITU-T H.265, 2013.
[32] X. Li, P. Amon, A. Hutter, and A. Kaup, "Performance analysis of inter-layer prediction in scalable video coding extension of H.264/AVC," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 66–74, Mar. 2011.



Meng Xu received the B.S. degree in physics from Nanjing University, China, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic School of Engineering, New York University, in 2009 and 2014, respectively. During the Ph.D. studies, he was an Intern with the Dialogic Media Laboratory, NJ, in 2010, Samsung Telecommunications America, TX, in 2013, and Huawei Technologies USA, CA, in 2013. He was with Huawei Technologies USA, Santa Clara, CA, from 2014 to 2015, as a Video Coding Researcher, and submitted more than 30 contributions during the standardization of HEVC screen content coding. His research interests include video coding and its applications. He has served as the Vice Chair with the JCT-VC Adhoc Group of SCC extensions software development (AHG8) since 2014.

Zhan Ma (S'06–M'11) received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2004 and 2006, respectively, and the Ph.D. degree from the Polytechnic School of Engineering, New York University, New York, in 2011. From 2011 to 2014, he has been with Samsung Research America, Dallas TX, and Futurewei Technologies, Inc., Santa Clara, CA. He is currently with the School of Electronic Science and Engineering, Nanjing University, Jiangsu, China, as a Faculty Member. His current research focuses on the next-generation video coding, energy-efficient communication, and multispectral signal compression.



Yao Wang (M'90–SM'98–F'04) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1983 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, in 1990. Since 1990, she has been on the Faculty of Electrical and Computer Engineering, Polytechnic School of Engineering, New York University. She was with Princeton University in 1998, and Thomson Corporate Research, Princeton, from 2004 to 2005. She was a Consultant with AT&T Laboratories–Research from 1992 to 2000. Her research areas include video communications, multimedia signal processing, and medical imaging. She is the leading author of a textbook entitled *Video Processing and Communications*, and has authored over 250 papers in journals and conference proceedings. She has served as an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. She received the New York City Mayor's Award for Excellence in Science and Technology in the Young Investigator Category in 2000. She is a co-winner of the IEEE Communications Society Leonard G. Abraham Prize Paper Award in the field of Communications Systems in 2004, and a co-winner of the IEEE Communications Society Multimedia Communication Technical Committee Best Paper Award in 2011. She also received the Overseas Outstanding Young Investigator Award from the Natural Science Foundation of China in 2005, and was named Yangtze River Lecture Scholar in Tsinghua University by the Ministry of Education, China, in 2007. She was a keynote speaker at the 2010 International Packet Video Workshop.