Fast Mode Decision Algorithm for Inter-Frame Coding in Fully Scalable Video Coding

He Li, Z. G. Li, *Senior Member, IEEE*, and Changyun Wen, *Senior Member, IEEE*

***Abstract*—Scalable video coding is an ongoing standard, and the current working draft (WD) is an extension of H.264/AVC. In the WD, an exhaustive search technique is employed to select the best coding mode for each macroblock. This technique achieves the highest possible coding efficiency, but it results in extremely large encoding time which obstructs it from practical use. This paper proposes a fast mode decision algorithm for inter-frame coding for spatial, coarse grain signal-to-noise ratio, and temporal scalability. It makes use of the mode-distribution correlation between the base layer and enhancement layers. Specifically, after the exhaustive search technique is performed at the base layer, the candidate modes for enhancement layers can be reduced to a small number based on the correlation. Experimental results show that the fast mode decision scheme reduces the computational complexity significantly with negligible coding loss and bit-rate increases.**

***Index Terms*—Coarse grain signal-to-noise ratio (CGS), fast mode decision, inter-frame coding, scalable video coding (SVC), spatial, temporal scalability.**

I. Introduction

SCALABLE video coding (SVC) is currently being developed as an extension of H.264/Advanced Video Coding (H.264/AVC) [2]. Compared to the previous video coding standards, SVC is intended to encode the signal once, but enable decoding from partial streams depending on the specific rate and resolution required by a certain application [3]. The basic design idea of SVC is to extend the hybrid video coding approach of H.264/AVC to efficiently incorporate spatial, SNR, and temporal scalability. The spatial and SNR scalability can be realized by a layered approach. The base layer contains a reduced resolution or a reduced quality version of each coded frame. The enhancement layers can be predicted from the base-layer pictures and previously encoded enhancement-layer pictures. Temporal scalability in SVC is achieved by using a structure of hierarchical B pictures [4], and a temporal scalable video coding algorithm allows extraction of video of multiple frame rates from a single coded stream.

Current SVC scheme shows significant achievements in terms of coding efficiency [5]. In this coding system, variable block-size matching motion estimation is used to reduce the temporal redundancy between frames. SVC defines seven macroblock (MB) modes for inter prediction ($\mathrm{MODE\_16 \times 16}$, $\mathrm{MODE\_16 \times 8}$, $\mathrm{MODE\_8 \times 16}$, $\mathrm{MODE\_8 \times 8}$, $\mathrm{MODE\_8 \times 4}$, $\mathrm{MODE\_4 \times 8}$ and $\mathrm{MODE\_4 \times 4}$), nine prediction modes for $\mathrm{INTRA\_4 \times 4}$, and four prediction modes for $\mathrm{INTRA\_16 \times 16}$ and $\mathrm{MODE\_SKIP}$ [2]. For encoding the motion field of an enhancement layer, "Base_layer_mode" and "Qpel_refinement_mode" are added to the modes applicable in the base layer. These two modes indicate that motion and prediction information including the partitioning of the corresponding MB of the base layer is used [2]. In this paper, we use BL_pred to represent these two modes. In order to choose the best coding mode for an MB, SVC calculates the rate distortion cost (RDcost) of every possible mode and selects the one with minimum RDcost as the best mode. Calculation of the RDcost in SVC needs to execute both the forward and backward processes of integer transform, quantization, inverse quantization, inverse integer transform, and entropy coding, and this introduces high computational complexity to the encoder. Therefore, it is desirable to design algorithms to reduce the computational complexity of SVC without compromising the coding efficiency for the implementation of SVC.

Recently, a number of efforts have been made to explore fast algorithms in intra-mode prediction and inter-mode prediction in H.264/AVC video coding. These algorithms achieve significant time savings with negligible loss of coding efficiency. In [6], a fast intra-mode decision algorithm is proposed based on the edge-detection histogram by making use of the Sobel operator. An effort has also been made by Wu *et al.* to use the spatial homogeneity and the temporal stationarity characteristics of video objects to guide the fast inter-prediction process [7]. Moreover, Yu *et al.* proposed fast mode decision algorithms by making use of the spatial complexity of the MB's content and the mode knowledge of the previously encoded frames [8], [9]. All of these methods are efficient in reducing the computational complexity with acceptable quality degradation in H.264/AVC encoder. However, these methods are not applicable to the enhancement layers of an SVC encoder. Fast mode decision for inter-frame coding in SVC is a new topic. Very few works exist so far, even though it plays a very important role in reducing the overall complexity of SVC.

We have observed that the mode distribution between the base layer and its enhancement layers has a certain correlation. In spatial scalability, for each MB at the base layer, the corresponding up-sampled MBs at enhancement layers tend to have the same mode partition. For coarse grain signal-to-noise ratio (CGS) scalability, each enhancement-layer MB tends to have a finer mode partition than the corresponding MB at the base layer. In the case of temporal scalability, the mode partition of MBs in the current frame is most similar to the mode partition of MBs in its reference frames. Motivated by these observations, we propose an effective fast mode decision for spatial, CGS, and temporal scalable video coding. With the proposal, a good mode partition prediction can be achieved if we predict the MB mode

Manuscript received December 5, 2005. This paper was recommended by Associate Editor H. Sun.

H. Li and C. Wen are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ecywen@ntu.edu.sg).

Z. G. Li is with the Media Division, Institute for Infocomm Research, Singapore 119613.

Digital Object Identifier 10.1109/TCSVT.2006.877404

at an enhancement layer from that at the base layer. Therefore, the presented algorithm reduces the number of candidate modes for an MB at enhancement layers by using the mode distribution at the base layer and, hence, the computational complexity significantly. Simulation results illustrate that our algorithm can achieve up to 61% of encoding time saving with negligible peak signal-to-noise (PSNR) loss and bit-rate increases.

The remainder of this paper is organized as follows. Section II presents an overview of inter-frame coding in spatial, CGS, and temporal scalability in SVC. Section III presents in detail the fast mode decision algorithm based on mode-distribution correlation among layers. Experimental results are presented in Section IV, and conclusions are given in Section V.

II. Overview of Inter-Frame Coding in SVC

Here, we begin by briefly reviewing the rate-distortion optimization (RDO) in inter-frame coding. Then, we study the characteristics of different scalabilities in SVC.

A. RDO

Similar to H.264/AVC, the motion estimation and mode decision process in SVC is performed by minimizing the rate-distortion cost function

$$J(= D(\mathrm{MODE}|QP) + \lambda_{\mathrm{SSD}} R(\mathrm{MODE}|QP)).$$

Here, D is the average of the forward and backward sum of absolute difference (SAD) or sum of square difference (SSD) between the current MB and the motion-compensated matching blocks, R denotes the bit cost for encoding the motion vectors, the MB header, and all of the residual information, and λ is a weight parameter to control the contribution of the motion bits in the total cost function. For each possible MB partition, the prediction method together with the associated reference indices r_0 and r_1 and motion vectors mv_0 and mv_1 is determined by minimizing $(D_{\mathrm{SAD}} + \lambda_{\mathrm{SAD}}(R(r_i) + R(mv_i)))$ $(i = 0{,}1)$.

The relationships among quantization parameter (QP), λ_{SSD} and λ_{SAD} are $\lambda_{\mathrm{SSD}} = 0.85 \times 2^{QP/3-4}$ and $\lambda_{\mathrm{SAD}} = \sqrt{\lambda_{\mathrm{SSD}}}$. Clearly, a large quantization step size results in a large value of λ and thus a low bit-rate range and a large amount of distortion [10]. On the other hand, a small quantization step size results in a small value of λ, and, therefore, a high bit-rate range and a small amount of distortion. Consequently, in CGS scalability, the mode partition of each MB at enhancement layers is finer than that of the corresponding MB at the base layer.

We tested two sequences *FOREMAN* and *FOOTBALL* with a JSVM 2.0 encoder as a statistical analysis. All of the test sequences are 100 frames long and the GOP size is 8. In the experiment, two CGS layers are evaluated. The QP values for the enhancement layer and the base layer are set to 10 and 40, respectively. Statistical results for inter MB mode distribution in two CGS layers are shown in Table I. From Table I, we can find that the percentage of fine partitioned MB increases as the quantization step size decreases. This shows that correlation exists between the base layer and its enhancement layers in CGS scalability.

TABLE I
Statistical Analysis of Inter-Mode Distributions

Mode	FOREMAN		FOOTBALL	
	Base	Enhancement	Base	Enhancement
Skip	4300	833	2585	249
16×16	2503	757	2302	753
16×8	637	616	1126	613
8×16	674	834	1218	823
8×8	174	4425	796	3073
BL_pred	0	552	0	1586
Total	8288	8017	8027	7097

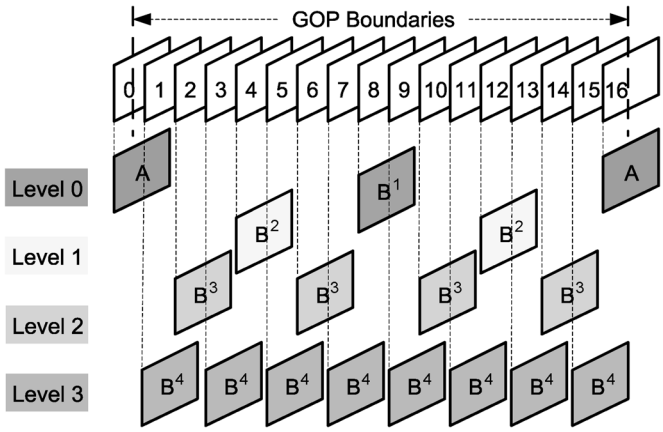

Fig. 1. Temporal scalability with a GOP size of 16.

B. Temporal Scalability in SVC

Temporal scalability in SVC is achieved by using a representation with hierarchical B pictures. Now take Fig. 1, which shows the temporal decomposition of a group of 16 pictures using four decomposition stages as an example. The first picture is independently coded as an instantaneous decoding refresh (IDR) picture, and all remaining pictures are coded in "B...BI" groups of pictures using the concept of hierarchical B pictures [4].

If only pictures B^1 and anchor frames A are transmitted, the reconstructed sequence at the decoder side has 1/8 of the temporal resolution of the input sequence. By additionally transmitting pictures B^2, the decoder can reconstruct an approximation of the picture sequence that has one quarter of the temporal resolution of the input sequence. Finally, if the remaining B pictures are transmitted, a reconstructed version of the original input sequence with the full temporal resolution is obtained.

For inter-frame coding, the MBs are classified into coarse-partitioned MBs (e.g., $\mathrm{MODE\_}SKIP$ and $\mathrm{MODE\_}16\times 16$) and fine-partitioned MBs (e.g., $\mathrm{MODE\_}8 \times 8$ and $\mathrm{INTRA\_}4 \times 4$). The number of fine-partitioned MBs depends on the temporal distance of the current frame and reference frames [11]. Suppose that the temporal distance between certain motion compensation pair is d_i and the corresponding mean for the percentage of fine-partitioned MBs is η_i. The relationship between η_i and d_i is given by

$$\eta_i = f(d_i) \tag{1}$$

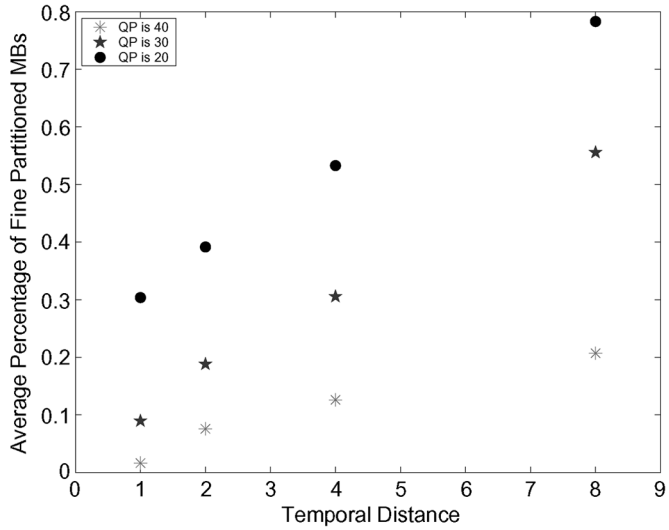

Fig. 2. Average percentage of fine-partitioned MBs for the MOBILE sequence.

Experiments on various video sequences are designed to investigate the statistical relationships between temporal distances and MB mode distributions. Fig. 2 shows the experimental results for *MOBILE* sequence with a GOP size of 16. It can be seen that f is an increasing function of d_i. When QP is fixed, at a low temporal level, the hierarchical B frames are generated with large temporal distance. The temporal correlation between the current frame and its referencing frames is low, consequently, the percentage of fine-partitioned MBs is high. Therefore, in temporal scalability, the partition of each MB in the low temporal level frames is finer than that of the corresponding MBs in the high temporal level frames, and this shows that correlation also exists between the base layer and its enhancement layers in temporal scalability.

C. Spatial Scalability in SVC

Here, spatial scalable coding of video is considered at multiple resolutions (e.g., QCIF, CIF, and 4CIF) with a factor of two in horizontal and vertical resolution. An oversampled pyramid representation is used for spatial scalability, where for each spatial resolution a separate refinement of motion and texture information is deployed [2].

When the base layer represents a layer with half the spatial resolution, according to the inter-layer prediction technique, the motion vector field including the MB partitioning is scaled. Therefore, the intra- and inter-MBs can be predicted using the corresponding signals of previous layers. Moreover, the motion description of each layer can be used for a prediction of the motion description for the following enhancement layers. In addition, for most cases, the up-sampling MBs at the enhancement layers tend to have the same mode partition. Therefore, in our proposed scheme, the base-layer MB mode is used to predict the corresponding enhancement-layer MB mode.

III. Proposed Fast Mode Decision Algorithm

It is observed that there are correlations between the base layer and its enhancement layers for spatial, CGS, and temporal scalability. Therefore, a good prediction could be achieved if we predict each MB mode partition at the enhancement layer from the corresponding MB at the base layer. Since temporal scalability can be achieved by a representation with hierarchical B pictures, it will be described separately from other scalabilities.

A. Spatial and CGS Scalability

Based on the considerations in Section II, three methods are proposed for spatial and CGS scalability.

1) Selective Intra-Mode Prediction: In SVC, if there is a significant change between the reference and current frames (for example, a scene change), it may be more efficient to encode the MB by intra mode. Therefore, in inter-frame coding in SVC, the encoder has to compute RDcost of all intra modes ($\mathrm{INTRA\_4}\times 4$ and $\mathrm{INTRA\_16}\times 16$) that involve testing all intra-prediction directions for all of the MBs. This process is very complex and the number of computing RDcost values is about five times higher than the case of inter modes [12]. However, as statistical data of intra mode indicates, the probability for an MB to have an intra mode in B slice is at most 7% and 4% on average, although the exact figure depends on specific input video characteristics. Such a small probability suggests that we should distinguish the intra-coded MBs in B slices at enhancement layers and only compute the RDcosts of intra modes for those MBs [13], [14].

As discussed in Section II-A, an enhancement layer is a motion and residual information refinement of its base layer. Therefore, for intra blocks in spatial and CGS enhancement layer, BL_pred and $\mathrm{INTRA\_4}\times 4$ are frequently selected modes in most cases. Without $\mathrm{INTRA\_16}\times 16$, BL_pred and $\mathrm{INTRA\_4}\times 4$ can preserve the accuracy of intra prediction well [15]. Therefore, $\mathrm{INTRA\_16}\times 16$ are removed to reduce the high computational load of intra coding while keeping the performance.

Fig. 3 shows the flowchart of our proposed selective intra-mode prediction method, where MODE_{BL} stands for the optimally selected MB mode at the base layer corresponding to the current MB at the enhancement layer. We divide the modes into two classes: $Class\_inter$ and $Class\_intra$. If MODE_{BL} is $\mathrm{INTRA\_4}\times 4$ or $\mathrm{INTRA\_16}\times 16$, then the corresponding MB at the enhancement layer belongs to $Class\_intra$, and the candidate mode set is reduced to BL_pred and $\mathrm{INTRA\_4}\times 4$.

As discussed in Section II-B, the hierarchical B frames at low temporal level are generated with large temporal distance, and they have more motion and texture information than that at a high temporal level. Therefore, MBs in the low temporal level frames have high probability to be intra coded. As a result, if MODE_{BL} is not intra coded, for high temporal level frames, the corresponding MB at the enhancement layer belongs to $Class\_inter$. In order to decide whether an MB is intra coded in low temporal level frames, we regard $\mathrm{MODE\_8}\times 8$ and $\mathrm{INTRA\_4}\times 4$ as the representative block sizes of $Class\_inter$ and $Class\_intra$. The RDcosts for $\mathrm{MODE\_8}\times 8$ (RDcost8) and $\mathrm{INTRA\_4}\times 4$ (RDcost4) for the MBs at low temporal level frames are estimated. If RDcost4 is less than RDcost8, we assume that the probability of intra coding is high and the best mode is set to $\mathrm{INTRA\_4}\times 4$. On the other hand, if RDcost8 is less than RDcost4, the best mode would belong to $Class\_inter$. Then, the following methods are used in the mode decision

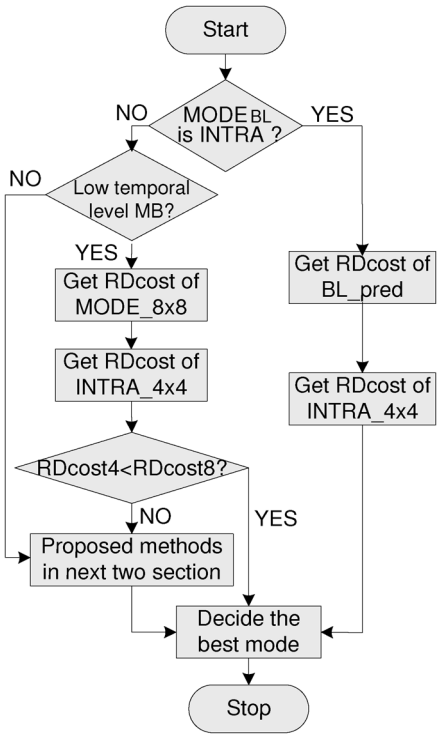

Fig. 3. Flowchart of proposed selective intra-mode prediction.

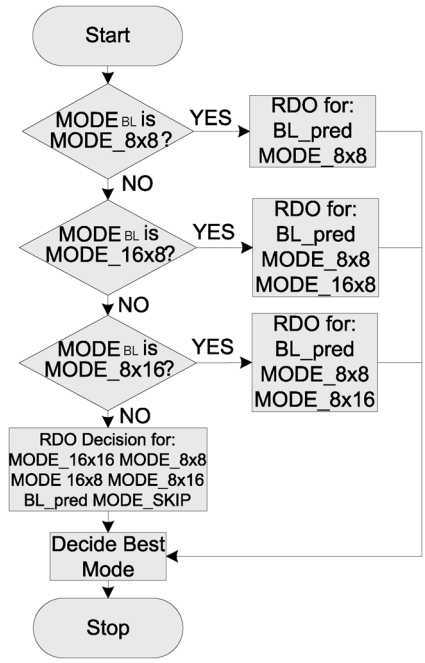

Fig. 4. Flowchart of proposed selective reduction of candidate modes method.

process for the MBs in high and low temporal level frames which belong to $Class\_inter$.

2) Selective Reduction of Candidate Modes: Since enhancement layers have refined motion and residual information of that at the base layer, we can reduce the number of candidate modes for certain MBs. If the MB mode at the base layer is MODE_8 $\times$ 8, then the candidate mode set is reduced to BL_pred and MODE_8 $\times$ 8. If the MB mode at the base layer is MODE_16 $\times$ 8 (or MODE_8 $\times$ 16), then the candidate mode set is reduced to BL_pred, MODE_8 $\times$ 8, and MODE_16 $\times$ 8 (or MODE_8 $\times$ 16). Fig. 4 shows the flowchart of selective reduction of candidate modes method.

3) Selective Residual Prediction at Enhancement Layers: This algorithm is used to examine whether the MBs at enhancement layers need residual prediction. For a pixel $p_{i,j}$, we use $R_Y(i,j)$, $R_{Cb}(i,j)$, and $R_{Cr}(i,j)$ to denote the coded previous layer residual for luma and chroma information in an MB. The sum of coded residual in the previous layer (SPR) is given as

$$\mathrm{SPR} = \sum_{i=0}^{15}\sum_{j=0}^{15}|R_Y(i,j)| + \sum_{i=0}^{7}\sum_{j=0}^{7}|R_{Cb}(i,j)| + \sum_{i=0}^{7}\sum_{j=0}^{7}|R_{Cr}(i,j)|. \quad (2)$$

Residual prediction is preformed at the enhancement layer if SPR is greater than a threshold θ. Otherwise, there is no residual prediction. The choice of θ provides a tradeoff between coding speed and quality. For most of video sequences, there is a high probability that the value of SPR for each MB is in two ranges: one is from 0 to 5, and the the other is greater than 100. This is shown by the experimental results given in Table II. In the experiment, two CGS layers are used. The QP for the base layer (denoted as BLQP) ranges from 40 to 20 and the QP for enhancement layer is set to 10. Since the coding quality will be degraded when θ is greater than 100, θ is selected as less than or equal to 5.

TABLE II
STATISTICAL ANALYSIS OF THE SPR VALUE

Sequence	BLQP	SPR<=5	5<SPR<100	SPR>=100
FOREMAN	40	96.46%	0	3.540%
	30	83.55%	0	16.45%
	20	56.35%	8.590%	35.06%
FOOTBALL	40	68.98%	0	31.02%
	30	38.96%	0	61.04%
	20	30.23%	2.309%	67.46%

B. Temporal Scalability

According to our experiments, the best prediction mode of each MB in the current frame is most similar to the optimal mode of the corresponding MBs in its reference frames [16]. For the frames in temporal level 0, only the anchor intra-coded frames can be used for motion-compensated prediction. Therefore, the original exhaustively block matching method is used in our scheme to search for the best mode for each MB in the frames at temporal level 0. In order to illustrate our idea, again we take Fig. 1 as an example. Frame 8 is estimated by using the block matching method without any fast mode decision algorithm. Frames 4 and 12 are at temporal level 1, and the best mode in frame 8 is the candidate mode for frames 4 and 12. Similarly, at temporal level 2, the best modes in frames 4 and 8 are the candidate modes for frame 6. Therefore, each frame has one or two candidate modes that are generated from the backward and/or forward reference frames.

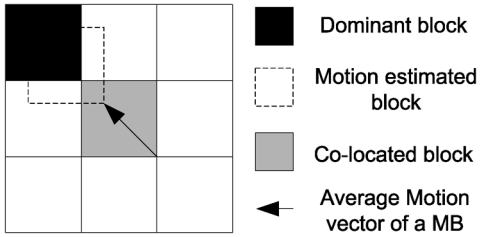

Fig. 5. Definition of the dominant block from the motion-estimated block.

Now, we are in the position to present our proposed scheme for temporal scalability as follows.

1) Determination of Low- and High-Motion MBs: We divide the MB into two classes, low-motion MB and high-motion MB. In natural video sequences, many MBs, especially the MBs in the background area, exhibit similar motion even if not still and are thus considered low-motion MBs. In this section, we propose a method to distinguish the low- and high-motion MBs by estimating the motion energy for each MB. After exhaustively search for the best mode for each MBs in the frames at temporal level 0, each possible MB partition (i.e., 16×16, 16×8, 8×16, 8×8, 8×4, 4×8, or 4×4) has independent motion vectors (mv_{x0}, mv_{y0}) and (mv_{x1}, mv_{y1}) for bi-direction prediction. The motion energy for each possible partition, denoted as MEb_0 and MEb_1, can be calculated as follows:

$$MEb_i = |mv_{xi}| + |mv_{yi}|, \quad i = 0, 1. \tag{3}$$

Note that the energy computed from (3) is the l_1_norm of vectors and it is equivalent to that defined in [18] where l_2_norm is used. We now use MEB_0 and MEB_1 to denote the average motion energy of an MB with respect to the backward and forward reference frames. Then

$$MEB_i = \frac{1}{N} \sum_{n=1}^{N} MEb_{ni}, \quad i = 0, 1 \tag{4}$$

where N represents total number of motion vectors of concerned MB. In our scheme, a threshold τ is set to distinguish high- and low-motion MBs as follows:

$$\begin{cases} \text{High-Motion MB}, & \text{if } MEB_0 > \tau \text{ or } MEB_1 > \tau \\ \text{Low-Motion MB}, & \text{otherwise.} \end{cases} \tag{5}$$

In our experiments, motion vector resolution is 1/4 pel and the MB size is 16×16. The threshold should be set greater than 8 pixels, which results in 32 in motion vector value. Based on the experimental results on all of the test sequences, we found that setting the threshold τ to 40 achieves good and consistent results for all of the test sequences.

2) Candidate Modes Assignment for High- and Low-Motion MBs: As shown in Fig. 5, every square represents a 16×16 MB in the reference frame [17]. For each MB in the current frame, the co-located MB locates at the same position in the reference frame. The arrow represents an average motion vector of all the partitions in the current MB. The dotted MB is the motion compensated one which most closely matches the current block. The dominant MB has the largest overlapping with the motion-compensated one.

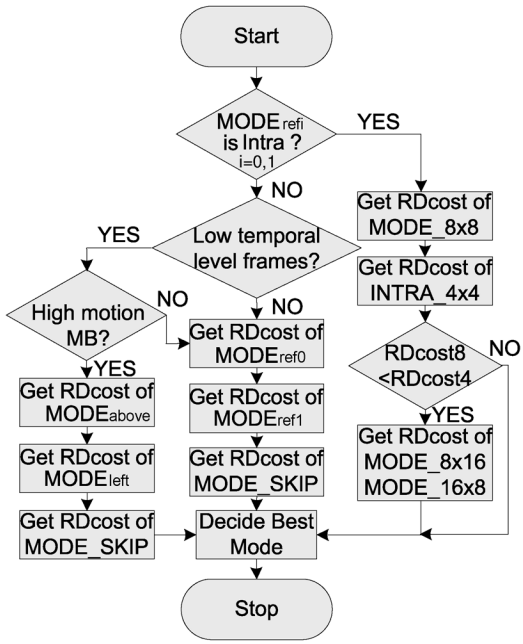

Fig. 6. Flowchart of proposed scheme for temporal scalability.

It is believed that motion is usually continuous, i.e., a directional feature of the current MB is similar to that of the motion compensated MB. Therefore, in our scheme, we need to examine whether the dominant MB is at the same position of the co-located one. For low-motion MBs, the dominant MB tends to have the same position of the co-located one. As a result, the MB mode of the co-located MB in the reference frame will be the candidate mode for the corresponding MB in the current frame. On the other hand, for high-motion MBs, the dominant one tends to locate at the neighborhoods of the co-located MB. As a result, the best modes of the co-located MB as well as its neighboring ones are composed of the candidate mode set for the corresponding MB in the current frames.

3) Overall Algorithm in Temporal Scalability: The flowchart of our scheme for temporal scalability is shown in Fig. 6. Similar to Section III-A1, if the MB mode at forward and/or backward reference frames (denoted by $\mathrm{MODE}_{\mathrm{ref}(i)}$) are intra coded, the rate distortion cost for $\mathrm{MODE\_8{\times}8}$ (RDcost8) and $\mathrm{INTRA\_4{\times}4}$ (RDcost4) are estimated. If RDcost4 is less then RDcost8, we assume that the ability of intra coding is high and the best mode is set to be $\mathrm{INTRA\_4 \times 4}$. On the other hand, if RDcost8 is less than RDcost4, we can regard that the best mode would be inter coded. Then,$\mathrm{MODE\_16 \times 8}$ and $\mathrm{MODE\_8 \times 16}$ will be the members of candidate mode set. On the other hand, if the MB mode at forward and/or backward reference frames are not intra coded, we need to examine whether the MB is in low temporal level frames. As discussed previously, low-temporal-level frames tend to have finer mode partition size than high-temporal-level frames. Moreover, the generated large distortion in the low-temporal-level frames coding will propagate and affect the coding efficiency of high-temporal-level frames.

TABLE III
SIMULATION CONDITIONS

		All Tested Video Sequences
QP Setting	Base	40
	Enhancement	10 to 30
Resolution	Base	QCIF
	Enhancement	CIF
Frame Rate	Base	7.5Hz
	Enhancement	15Hz
Coding Option Used		MV search range is ±32 pels. Reference frame number is 1. MV resolution is 1/4 pel.
Codec		JSVM 2.0 encoder

Therefore, it is important to increase the motion estimation accuracy in low-temporal-level frames. For high-motion MBs in low-temporal-level frames, it is difficult to find a temporal correlation from its reference frames. Instead, we need to consider the spatial correlation among MBs in the current frame. This is due to the fact that there is usually a high correlation between pixels that close to each other. Therefore, for a high-motion MB at low-temporal-level frames, the best modes of above and left MBs, denoted by $\mathrm{MODE}_{\mathrm{above}}$ and $\mathrm{MODE}_{\mathrm{left}}$, respectively, are considered as the candidate modes for current MB. On the other hand, for MBs at high-temporal-level frames or low-motion MB at low-temporal-level frames, the best modes of MBs in the forward and backward reference frames, denoted by $\mathrm{MODE}_{\mathrm{ref0}}$ and $\mathrm{MODE}_{\mathrm{ref1}}$, respectively, are considered as the candidate modes for current MB.

IV. EXPERIMENTAL RESULTS

The performance of our proposed fast mode decision algorithm for inter-frame coding in SVC is evaluated through simulation studies. Our scheme is implemented on a JSVM 2.0 encoder [2]. The test platform used is Intel Pentium IV, 1.83-GHz CPU, 256-M RAM with Windows XP professional operating system. The test condition is shown in Table III. In our experiments, six standard test sequences including *FOREMAN*, *FOOTBALL*, *BUS*, *HARBOUR*, *CITY*, and *CREW* have been tested.

The testing parameters in our experiments include the average time saving (TS), Bjontegaard delta PSNR (BDPSNR), and Bjontegaard delta bit rate (BDBR) [19]. BDPSNR and BDBR are used to represent the average PSNR and bit-rate differences between the RD curves derived from JSVM encoder and the proposed fast algorithm, respectively.

A. Spatial and CGS Scalability

In this experiment, the total number of frames is 50 for each sequence, and the group of picture size is 16. The experimental results are given in Table IV and Figs. 7 and 8. Note that, in the table, positive values mean increments, and negative values mean decrements. It can be seen that our scheme achieves consistent time saving over a large bit-rate range with negligible losses in PSNR and increments in bit rate. By comparing Figs. 7 and 8, the difference between two RD curves at a high bit rate is larger for the *FOOTBALL* sequence. This is because *FOOTBALL* represents a sequence with high motion and fine details. The motion correlation between the base layer and enhancement layer is lower compared with low-motion sequence.

TABLE IV
SIMULATION RESULTS IN SPATIAL AND CGS SCALABILITY

Sequence	BDPSNR [dB]	BDBR [%]	AVTS [%]
FOREMAN	-0.037	0.698	51.56
FOOTBALL	-0.128	0.921	61.02
BUS	-0.035	0.049	53.07
HARBOUR	-0.032	0.209	51.46
CITY	-0.039	0.384	49.27
CREW	-0.065	1.102	52.14
Average	-0.056	0.561	53.09

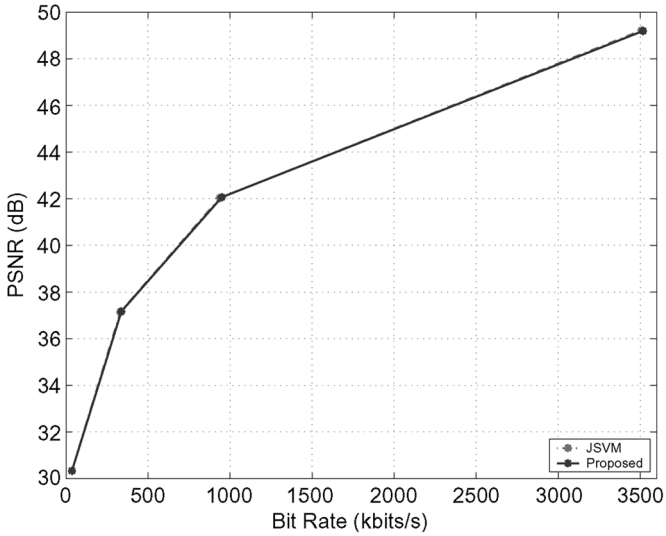

Fig. 7. Rate-distortion curve for FOREMAN in spatial and CGS scalability.

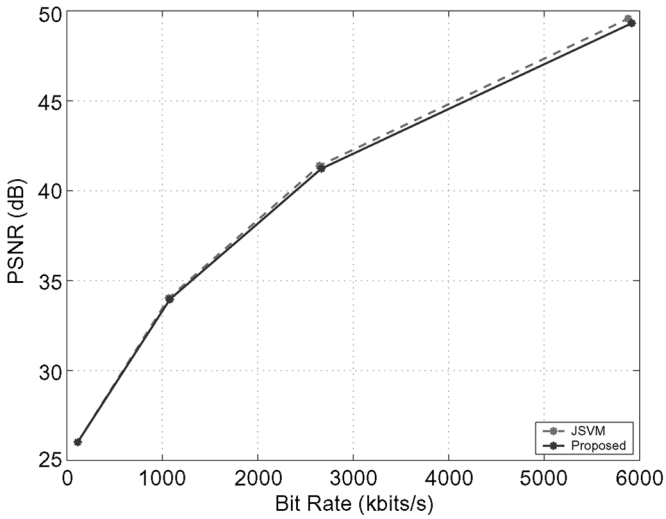

Fig. 8. Rate-distortion curve for FOOTBALL in spatial and CGS scalability.

B. Temporal Scalability

In this experiment, the total number of frames is 100 for each sequence, and the group of picture size is 16. Enhancement-layer frames are the output frames that have full temporal resolution of input frames. Base-layer frames are the output frames that have half-temporal resolution of input frames. In our

TABLE V
SIMULATION RESULTS IN TEMPORAL SCALABILITY

Sequence	BDPSNR [dB]	BDBR [%]	AVTS [%]
FOREMAN	-0.209	3.392	37.86
FOOTBALL	-0.133	1.665	30.51
BUS	-0.151	2.047	36.50
HARBOUR	-0.072	1.140	42.65
CITY	-0.115	2.117	43.79
CREW	-0.156	2.551	35.31
Average	-0.139	2.152	37.77

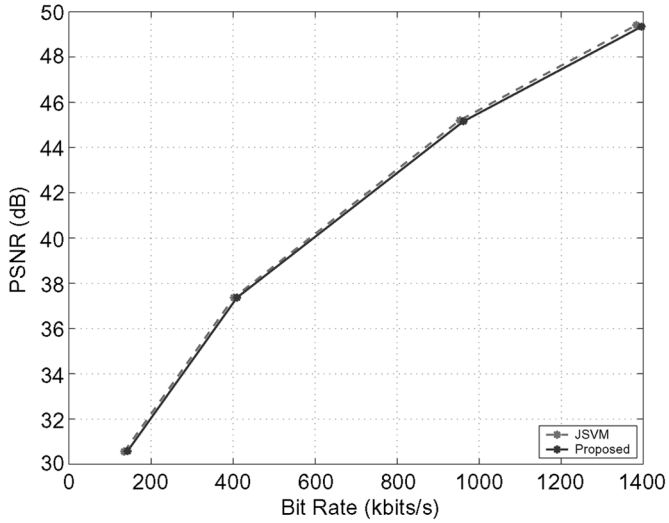

Fig. 9. Rate-distortion curve for FOOTBALL in spatial and CGS scalability.

case, the frame rate of enhancement-layer frames and base-layer frames are 15 and 7.5 frames/s, respectively.

The average PSNR and bit-rate differences in terms of BDPSNR and BDBR and the average TS in this experiment are shown in Table V. The results show that the proposed method is also very effective in reducing the encoding time, especially for the sequence with high motion and fine detail. The total encoding time is reduced up to 37.8%. Fig. 9 presents the rate-distortion curves for the output frames that have full temporal resolution for *FOOTBALL*. From this figure, we can conclude that our scheme can achieve consistent TS over a large bit-rate range with negligible loss in PSNR and increments in bit rate.

V. CONCLUSION

In this paper, we present a fast mode decision algorithm for inter-frame coding in SVC by using the mode distribution correlation between the base layer and its enhancement layers. The number of candidate modes for luma and chroma blocks in an MB that takes part in RDO calculation has been reduced significantly at enhancement layers. This fast mode decision algorithm is able to achieve a reduction of 53% encoding time on average, with a negligible average PSNR loss of 0.056 dB and 0.56% bit-rate increase in spatial and CGS scalability. For temporal scalability, our proposed scheme can achieve a reduction of 37.8% encoding time on average, with an acceptable average PSNR loss of 0.139 dB and 2.152% bit-rate increase.

REFERENCES

[1] J. Reichel, H. Schwarz, and M. Wien, Scalable Video Coding-Join Draft 4 , ISO/IEC JTC1/SC29/WG11/JVT-Q201. Nice, France, Oct. 2005.
[2] ——, Joint Scalable Video Model 2.0 Reference Encoding Algorithm Description , ISO/IEC JTC1/SC29/WG11/N7084. Buzan, Korea, Apr. 2005.
[3] J.-R. Ohm, "Advances in scalable video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.
[4] J. Reichel, H. Schwarz, and M. Wien, Joint Scalable Video Model (JSVM) 4.0 Reference Encoding Algorithm Description , ISO/IEC JTC1/SC29/WG11/N7556. Nice, France, Oct. 2005.
[5] Report of the formal verification tests on AVC (ISO/IEC 14496-10 |ITU-T Rec. H. 264) MPEG2003/N6231, Dec. 2003.
[6] F. Pan, X. Lin, R. Susanto, K. P. Lim, Z. G. Li, G. N. Feng, D. J. Wu, and S. Wu, "Fast mode decision algorithm for intraprediction in H.264/AVC video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 813–822, Jul. 2005.
[7] D. Wu, F. Pan, K. P. Lim, S. Wu, Z. G. Li, X. Lin, S. Rahardja, and C. C. Ko, "Fast intermode decision in H.264/AVC video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 953–958, Jul. 2005.
[8] A. C. Yu, "Efficient block-size selection algorithm for inter-frame coding in H.264/MPEG-4 AVC," in *Proc. IEEE ICASSP*, 2004, pp. 169–172.
[9] A. C. Yu and G. R. Martin, "Advanced block size selection algorithm for inter frame coding in H.264/MPEG-4 AVC," in *Proc. IEEE ICIP*, 2004, pp. 95–98.
[10] Z. G. Li, Y. C. Soh, and C. Y. Wen, *Switched and Impulsive Systems: Analysis, Design and Applications*. Berlin, Germany: Springer-Verlag, 2004, pp. 197–219.
[11] S.-J. Choi and J. Woods, "Motion compensated 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb, 1999.
[12] B. Jeon and J. Lee, "," Fast Mode Decision for H.264 , ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q6/J033, Hawaii, Dec. 2003.
[13] H. Li, Z. G. Li, and C. Wen, "Fast mode decision for spatial scalable video coding," in *Proc. ISCAS*, May 2006, pp. 3005–3008.
[14] ——, "Fast mode decision for coarse grain SNR scalable video coding," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 2, pp. 545–548.
[15] L. B. Yang, Y. Chen, J. F. Zhai, and F. Zhang, Low Complexity Intra Prediction for Enhancement Layer , ISO/IEC JTC1/SC29/WG11/Q084. Nice, France, Oct. 2005.
[16] H. Li, Z. G. Li, and C. Wen, "Fast mode decision for temporal scalable video coding," in *Proc. Picture Coding Symp.*, Beijing, China, Apr. 2006.
[17] M. C. Hwang, J. K. Cho, J. H. Kim, and S. J. Ko, "A fast intra prediction mode decision algorithm based on temporal correlation for H.264," in *Proc. ITC-CSCC*, Jeju, Korea, Jul. 2005, vol. 4, pp. 1573–1574.
[18] H. Zhu, C. K. Wu, Y. L. Wang, and Y. Fang, "Fast mode decision for H.264/AVC based on macroblock correlation," in *Proc. 19th Int. Conf. Adv. Inf. Netw. Appl.*, 2005, vol. 1, pp. 775–780.
[19] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," presented at the 13th VCEG-M33 Meeting, Austin, TX, Apr. 2–4, 2001.