Fast Implementation of the Scalable Video Coding Extension of the H.264/AVC Standard*

Xin Lu and Graham R. Martin

Department of Computer Science
University of Warwick, Coventry
CV4 7AL, United Kingdom
{xin, grm}@dcs.warwick.ac.uk

Abstract

In order to improve coding efficiency in the scalable extension of H.264/AVC, an inter-layer prediction mechanism is incorporated. This exploits as much lower layer information as possible to inform the process of coding the enhancement layer(s). However it also greatly increases the computational complexity. In this paper, a fast mode decision algorithm for efficient implementation of the SVC encoder is described. The proposed algorithm not only considers inter-layer correlation but also fully exploits both spatial and temporal correlation as well as an assessment of macroblock texture. All of these factors are organised within a hierarchical structure in the mode decision process. At each level of the structure, different strategies are implemented to eliminate inappropriate candidate modes. Simulation results show that the proposed algorithm reduces encoding time by up to 85% compared with the JSVM 9.18 implementation. This is achieved without any noticeable degradation in rate distortion.

1998 ACM Subject Classification I.4.2 Compression (Coding), E.4 Coding and Information Theory

Keywords and phrases Fast mode selection, Inter-layer prediction, Scalable Video Coding (SVC), SVC extension of H.264/AVC.

Digital Object Identifier 10.4230/OASIcs.ICCSW.2013.65

1 Introduction

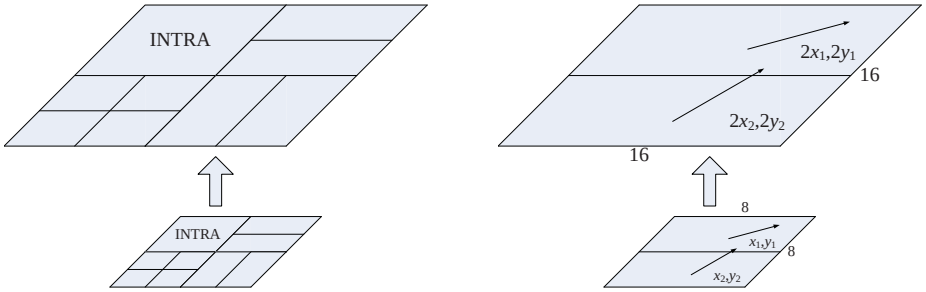
The Scalable Video Coding (SVC) extension of H.264, also known as MPEG-4 part 10 Advanced Video Coding Amendment 3 [2], is the result of joint standardisation work by ISO/IEC JTC1 SC29/WG11 (MPEG) and ITU-T. SVC enables a video sequence to be decoded fully or partially with variable quality, resolution and frame rate depending on the available network bandwidth or application requirements [10].

The coding mode decision process in the enhancement layer(s) requires an extremely large amount of computation. It is observed that this process dominates the encoding time in H.264/SVC. This is due to the application of many time consuming encoding tools. For instance, rate distortion optimisation and inter-layer prediction are involved in the mode decision process. Evaluation results show that the mode decision process in the enhancement layers accounts for 90% of the total computational requirement [3], and the encoding time of the enhancement layer is 10 times more than that of the base layer. In the rate distortion optimised mode decision process, more candidate partition modes are involved in SVC than any previous video coding standard. In SVC, all the encoding tools of H.264/AVC have been

* A longer version of this paper appeared in IEEE T-CSVT [8].

2013 Imperial College Computing Student Workshop (ICCSW'13).
Editors: Andrew V. Jones, Nicholas Ng; pp. 65–72

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



(a) Partition prediction. **(b)** Motion prediction.

Figure 1 Inter-layer prediction.

inherited [9], and these are supplemented by additional tools to support scalability. SVC uses a layer-based scheme to provide spatial and quality scalability [10], and in order to improve coding efficiency, it employs a mechanism to reuse the coded lower layer data for encoding the corresponding enhancement layer. This is referred to as inter-layer prediction [11]. In this, a new block mode is introduced, a base layer skip (BLSKIP), which applies three types of coding tools including inter-layer texture prediction, inter-layer motion prediction and inter-layer residual prediction, as shown in Figure 1. There are eight available macroblock modes for inter prediction in H.264/AVC, namely MODE_SKIP, MODE_16×16, MODE_16×8, MODE_8×16, MODE_8×8, MODE_8×4, MODE_4×8, MODE_4×4 (Figure 2), and two modes for intra prediction, INTRA_4×4 and INTRA_16×16 [13]. For an enhancement layer in SVC, all the modes concerned with inter prediction, intra prediction and inter layer prediction are evaluated, and the mode with the minimum rate distortion (RD) cost is selected as the best mode for the current macroblock. The rate distortion function for a block ω_k is given by

$$J(\omega_k, \hat{\omega}_k, \text{MODE}|\text{Qp}) = D(\omega_k, \hat{\omega}_k, \text{MODE}|\text{Qp}) + \lambda(\text{Qp}) \times R(\omega_k, \hat{\omega}_k, \text{MODE}|\text{Qp}) \quad (1)$$

where ω_k is an original macroblock at time k, and $\hat{\omega}_k$ is the corresponding constructed block. Qp is the quantisation parameter [12]. R represents the number of bits, and D denotes

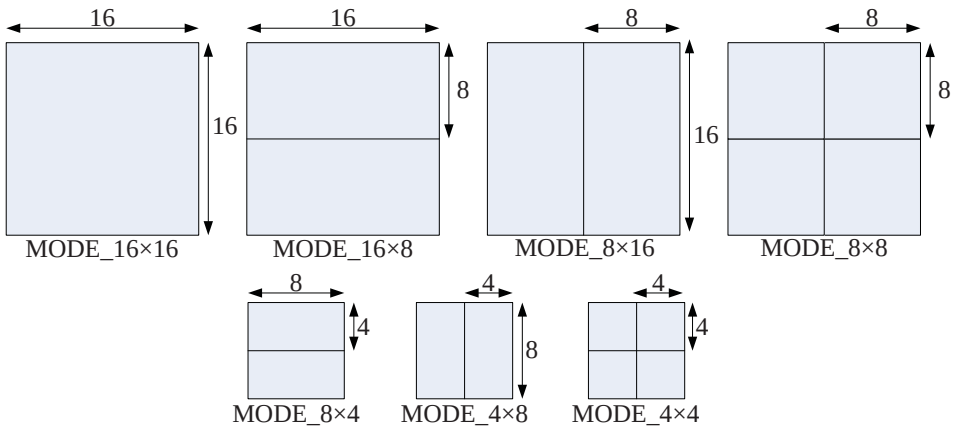


Figure 2 Block modes for inter-frame prediction in H.264/AVC.

a distortion measure. When exhaustive mode selection is used, the rate distortion cost must be evaluated for all modes in order to decide which mode should be employed. A set of time consuming encoding tools are incorporated in SVC as well as in H.264/AVC, for instance, bi-directional motion prediction, quarter-pixel precision motion estimation, and motion compensation using multiple reference blocks. In addition, the inter-layer prediction tools of SVC also incur excessive computational cost. In particular, inter-layer residual prediction doubles the computational complexity of the mode decision process [3]. Inter-layer motion prediction also results in a significant increase in computational complexity.

Several algorithms have been suggested to reduce the computational complexity of SVC. In [5], Kim *et al.* used the RD cost of the base layer skip mode and the information in the base layer to categorise macroblocks into different groups. The algorithm then reduces the number of candidate modes for the enhancement layer. The algorithm's performance is highly dependent on a constant K which determines the time saving and reconstructed picture quality. In [7], depending on the mode distribution relationship between the base layer and enhancement layers, Li *et al.* reduced the number of candidate modes in the enhancement layer according to the best mode in the corresponding position in the base layer. However, this method provides poor results when the correlation between base layer and enhancement layer is weak. Goh *et al.* [4] proposed an algorithm that makes use of the relationship between current macroblocks and their neighbours to decrease the encoding time. Nevertheless, for fast changing video sequences, it results in expected bit-rate degradation. In [16], Zhao *et al.* utilised the encoding mode of the base layer to initialise the candidate mode list of the enhancement layer, thus saving the overall encoding time. However the mode relationship between the macroblock and its neighbours is not investigated, and the time saving can be further improved.

In this paper we suggest a fast mode decision approach to alleviate the computational complexity of the SVC encoder without any significant loss of compression performance and coding efficiency. The remainder of this paper is organised as follows. Section 2 discusses the formulation of the proposed algorithm, and the overall structure is described in Section 3. Extensive experimental results are presented in Section 4 and conclusions are given in Section 5.

2 Observations and Analysis

The proposed algorithm is formulated as a hierarchical application of knowledge gained from previously processed information and the inherent content of the video being encoded.

2.1 Mode Correlation between Base Layer and Enhancement Layer

In order to support spatial and quality scalability, SVC adopts a multi-layer coding approach. To improve coding efficiency, a new prediction mechanism referred to as inter-layer prediction was incorporated into the standard. The purpose of using inter-layer prediction tools is to exploit as much lower layer information as possible for improving the coding efficiency of the enhancement layer.

In general, the base layer is first encoded independently, followed by the enhancement layer. As the input video of the different layers is generated from the same original video source but at different spatial resolution, the picture content is highly correlated [15]. Specifically, the prediction mode and motion vectors of the enhancement layer are strongly correlated with those of the base layer. Consequently, by exploiting the mode information of the

Table 1 % Mode Correlation between Base Layer and Corresponding Enhancement Layer.

Sequence	Qp				
	24	28	32	36	40
Bus	40.09	47.30	53.16	58.61	65.71
Foreman	42.90	53.97	61.04	69.14	79.94
Mobile	49.92	57.70	63.57	65.44	70.35
Mother-Daughter	78.81	85.40	90.37	95.03	97.64

corresponding macroblock in the base layer, the computational complexity of the mode decision process in the enhancement layer can be reduced significantly.

To illustrate the mode relationship between the base layer and the enhancement layers and to justify our proposed algorithm, we analysed the probability of macroblocks in the enhancement layers being encoded as MODE_SKIP when the mode of the co-located macroblock in the base layer is also MODE_SKIP. Table 1 shows the mode correlation between the base layer and the enhancement layer as defined in equation (2). We examined four video sequences with different degrees of activity and detail. The statistics were collected from the first 90 frames of each video sequence.

$$MC_{IL} = \frac{MB_{B\&E\_SKIP}}{MB_{E\_SKIP}} \times 100\% \tag{2}$$

where $MB_{B\&E\_SKIP}$ is the number of macroblocks predicted as MODE_SKIP in the base layer when the corresponding macroblock in the enhancement layer is MODE_SKIP too. MB_{E_SKIP} is the number of MODE_SKIP macroblocks in the enhancement layer.

From Table 1, we deduce that if the best mode for the macroblock in the base layer is MODE_SKIP, the corresponding macroblock mode in the enhancement layer is very likely to be MODE_SKIP as well. This is true regardless of video sequence.

2.2 Mode Correlation between Macroblock and its Neighbours

Besides the correlation between layers, there is also a significant dependency of neighbouring macroblocks in the enhancement layer. For a majority of video sequences, MODE_SKIP macroblocks tend to occur in clusters, such as a patch of static background. Consequently there is a high possibility that the best mode for the current macroblock is similar to that of its neighbours, that is, coded macroblocks which are located to the above and to the left of the current macroblock.

In order to reveal the mode dependency between macroblocks, we measured the probability that the current macroblock is coded as MODE_SKIP, if one or both of the neighbouring macroblocks are also coded MODE_SKIP.

Equation (3) denotes the mode correlation between the current macroblock and its neighbours.

$$MC_{SN} = \frac{MB_{C\&N\_SKIP}}{MB_{C\_SKIP}} \times 100\% \tag{3}$$

where MC_{SN} is the mode correlation between a macroblock and its spatial neighbours. $MB_{C\&N\_SKIP}$ is the number of the macroblocks which are predicted MODE_SKIP when the macroblock's neighbours in the enhancement layer are MODE_SKIP as well. MB_{C_SKIP} is the number of MODE_SKIP macroblocks in the enhancement layer. Table 2 shows the mode correlation between a macroblock and its spatial neighbours as defined in equation (3).

From the observations above, we infer that if the best mode for the co-located macroblock in the base layer or the neighbouring macroblocks in the enhancement layer is MODE_SKIP, there is a high probability that the current macroblock in the enhancement layer is also MODE_SKIP, because of the strong dependency that exists.

2.3 DCT Coefficients and Picture Content

In a smooth region of an image that has been transformed using the discrete cosine transform (DCT), the DCT energy generally tends to be concentrated in the low frequency components. Whereas, in a block comprising high detail, the frequency domain energy is concentrated in the AC coefficients. On this basis, the sum of the AC coefficients can be chosen as a coarse measure of homogeneity.

For a 16×16 macroblock, the energy of the AC coefficients E_{AC} is calculated as

$$E_{AC} = \sum_{x=0}^{15}\sum_{y=0}^{15}\left(f\left(x,y\right)\right)^2 - \frac{1}{256}\left(\sum_{x=0}^{15}\sum_{y=0}^{15} f(x,y)\right)^2 \tag{4}$$

Yu *et al* [14] showed that for fast mode selection in H.264/AVC, when the total energy of the AC coefficients in the macroblock is greater than 92735, the macroblock is best categorised as containing high spatial detail, as shown in (5), and for this to be considered when choosing a reduced subset of modes for evaluation.

$$\text{homogeneity} = \begin{cases} high & \text{if } \mathrm{E_{AC}} > 92735 \\ low & \text{otherwise} \end{cases} \tag{5}$$

In the evaluation, an AC energy threshold of 92735 was chosen to categorise the homogeneity of the macroblock content.

2.4 Detection of Motion Activity

As mentioned in [10], not all lower layer up-sampling data is suitable for inter-layer prediction, especially for video sequences with slow motion and high spatial detail. Therefore it is important to identify the amount of motion as well as the spatial detail.

Motion vector difference (MVD) between key frames in each GOP is chosen as the assessment of motion. MVD is the difference between the actual motion vector and the predicted motion vector, defined as

$$|MVD| = |MV_{actual} - MV_p| \tag{6}$$

where MV_{actual} is the actual motion vector for the block and MV_p is the predicted motion vector.

Table 2 % Mode Correlation between Macroblock and its Neighbours.

Sequence	Qp				
	24	28	32	36	40
Bus	45.24	52.18	56.95	62.52	66.88
Foreman	45.00	55.24	61.93	69.26	77.53
Mobile	46.06	52.43	57.78	62.16	66.27
Mother-Daughter	71.73	82.05	87.07	92.86	96.16

Generally, sequences containing little motion tend to have small MVDs and vice versa. The MVD is easy to extract from the coded data, and this satisfies the overall objective of the research, to reduce computational complexity. In the evaluation, a MVD of 1.0 was chosen as the threshold.

3 Proposed Algorithm

The basic motivation is to reduce the number of mode candidates in the enhancement layer by exploiting the information in co-located macroblocks in the base layer and in neighbouring macroblocks. As the coded data of the base layer will be reused either directly or indirectly, its efficacy influences the performance of the encoder. The mode chosen for a macroblock in the base layer is estimated using an exhaustive search evaluation. As to the enhancement layer, our proposed algorithm is described as follows:

1. Check the mode of the co-located macroblock in the base layer. If it is intra coded, it means that a matching macroblock via inter prediction in the reference frames could not be found. Usually, such macroblocks contain fast changing or highly detailed information. For such macroblocks, we compare the RD cost of all the modes (including inter-layer prediction) and select the mode with minimum RD cost as the best mode for the current macroblock. Otherwise, proceed to step 2.
2. Check the co-located macroblock in the base layer and the neighbouring macroblocks. If at least one of these macroblocks is encoded with MODE_SKIP, we evaluate the RD cost of employing either MODE_SKIP or MODE_16×16. If mode MODE_SKIP has the least RD cost, it is chosen as the best mode for the current macroblock. Otherwise, proceed to step 3.
3. Measure the homogeneity of the macroblock content. In the case of low homogeneity, i.e. $E_{AC} \leq 92735$, large block partition sizes (MODE_16×16, MODE_16×8, MODE_8×16) require evaluation. Otherwise, proceed to step 4.
4. Observe the MVD which is easily extracted from the coded bit stream. If MVD<1, the macroblock contains little motion, and motion estimation can be performed with fewer candidates. Otherwise, more candidates are chosen corresponding to a larger search range.

4 Experimental Evaluation

The proposed algorithm was implemented using the JSVM 9.18 reference software [1]. Four standard video test sequences with diverse motion content were utilised with different Qp factors ranging from 24 to 40. The GOP size for the hierarchical B structure was set to 8 and over 90 frames were coded to generate a reliable result. We considered only the two-layer case. The same Qp was used for both base layer and enhancement layer. We chose computation Time Reduction (TR), bit-rate and PSNR as the performance measures. Table 3 shows the coding results of the standard JSVM 9.18 implementation and the proposed algorithm.

Table 3 shows that, in the case of the Bus sequence with fast motion activity and the Mobile sequence with high spatial detail, as the percentage of MODE_SKIP macroblocks in the base layer is small, the time reduction is lower than the other test sequences. Even so, the computation time is reduced by over 61%. For the Mother-Daughter sequence comprising little motion, an average time saving of 78% is achieved. In Table 4, it is apparent that the encoding time reduction increases as the value of Qp increases. The maximum time saved is 85% for the sequence containing slow motion. The proposed algorithm shows minimal degradation in coding efficiency with a bit-rate increment of no more than 1.6%, and a

Table 3 Performance when Encoding QCIF/CIF Sequences.

Sequence	Qp	JSVM		Proposed		TR
		BR(bits/s)	PSNR(dB)	BR(bits/s)	PSNR(dB)	
Bus	40	344.28	26.93	345.56	26.89	67.06
	36	565.78	29.44	570.09	29.39	64.36
	32	942.03	32.15	952.37	32.09	63.23
	28	1579.67	35.07	1593.86	34.98	61.28
	24	2581.37	38.00	2606.26	37.81	61.39
Foreman	40	157.93	30.28	157.76	30.25	73.19
	36	246.51	32.68	247.43	32.63	68.61
	32	390.35	34.96	392.54	34.91	65.07
	28	635.55	37.32	643.00	37.25	62.71
	24	1073.34	39.57	1090.13	39.47	60.53
Mobile	40	469.64	25.86	470.31	25.83	71.27
	36	752.69	28.44	754.84	28.41	69.85
	32	1314.27	31.25	1319.59	31.21	69.04
	28	2342.70	34.48	2359.49	34.42	68.58
	24	3940.10	37.71	3966.08	37.63	67.63
Mother-Daughter	40	66.83	32.24	66.66	32.22	84.59
	36	107.42	34.62	107.52	34.61	81.94
	32	175.36	37.08	175.43	37.07	78.42
	28	284.31	39.59	284.58	39.56	74.98
	24	463.07	41.87	463.36	41.83	69.45

Table 4 Overall Comparison of Proposed Algorithm and JSVM Implementation.

Sequence	Performance	Qp					Average
		40	36	32	28	24	
Bus	ΔPSNR(dB)	-0.04	-0.05	-0.06	-0.09	-0.19	-0.09
	ΔBR(%)	0.37	0.76	1.10	0.90	0.96	0.82
	TR(%)	67.06	64.36	63.23	61.28	61.39	63.46
Foreman	ΔPSNR(dB)	-0.03	-0.05	-0.05	-0.07	-0.1	-0.06
	ΔBR(%)	-0.11	0.37	0.56	1.17	1.56	0.71
	TR(%)	73.19	68.61	65.07	62.71	60.53	66.02
Mobile	ΔPSNR(dB)	-0.03	-0.03	-0.04	-0.06	-0.08	-0.05
	ΔBR(%)	0.14	0.29	0.40	0.72	0.66	0.44
	TR(%)	71.27	69.85	69.04	68.58	67.63	69.27
Mother-Daughter	ΔPSNR(dB)	-0.02	-0.01	-0.01	-0.03	-0.04	-0.02
	ΔBR(%)	-0.25	0.09	0.04	0.09	0.06	0.01
	TR(%)	84.59	81.94	78.42	74.98	69.45	77.88

decrease in PSNR of no more than 0.2%. Compared with Lee's algorithm [6], our proposed scheme achieves a greater time reduction for sequences with varying motion activity and spatial detail. The method also outperforms that of both Zhao [16] and Kim [5].

5 Conclusion

A fast hierarchical mode selection algorithm for the SVC extension of H.264/AVC has been described. The scheme reduces the number of mode candidates that need to be evaluated by exploiting the base layer information. Simulation results show that the algorithm achieves a reduction in encoding time of up to 85% with negligible reduction in coding efficiency and reconstructed video quality.

References

1 JSVM (Joint Scalable Video Model) reference software for SVC. Online. Available: CVS server garcon.ient.rwth-aachen.de.

2 H.264: Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 Advanced Video Coding, Mar. 2010.

3 Z. Y. Chen, J. W. Syu, and P. C. Chang. Fast inter-layer motion estimation algorithm on spatial scalability in H.264/AVC scalable extension. In *Proc. IEEE ICME*, pages 442–446, 2010.

4 G. Goh, J. Kang, M. Cho, and K. Chung. Fast mode decision for scalable video coding based on neighboring macroblock analysis. In *Proc. ACM Appl. Computing*, pages 1845–1846, 2009.

5 S. T. Kim, K. Reddy Konda, P. S. Mah, and S. J. Ko. Adaptive mode decision algorithm for inter layer coding in scalable video coding. In *IEEE Trans. Circuits Syst. Video Technol.*, pages 1297–1300, 2010.

6 B. Lee, M. Kim, S. Hahm, C. Park, and K. Park. A fast mode selection scheme in inter-layer prediction of H.264 scalable extension coding. In *Proc. IEEE BMSB*, pages 1–5, 2008.

7 H. Li, Z. G. Li, and C. Wen. Fast mode decision for coarse grain SNR scalable video coding. In *Proc. IEEE ICASSP*, volume 2, pages II–II, 2006.

8 X. Lu and G. R. Martin. Fast mode decision algorithm for the H.264/AVC scalable video coding extension. *IEEE Trans. Circuits Syst. Video Technol.*, 23(5):846–855, 2013.

9 H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable H.264/MPEG4-AVC extension. In *Proc. IEEE ICIP*, pages 161–164, 2006.

10 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.*, 17(9):1103–1120, 2007.

11 C. A. Segall and G. J. Sullivan. Spatial scalability within the H.264/AVC scalable video coding extension. *IEEE Trans. Circuits Syst. Video Technol.*, 17(9):1121–1135, 2007.

12 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.*, 13(7):688–703, 2003.

13 T. Wiegand, G. J. Sullivan, G. Bjϕntegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.*, 13(7):560–576, 2003.

14 A. C. W. Yu, G. R. Martin, and H. Park. Fast inter-mode selection in the H.264/AVC standard using a hierarchical decision process. *IEEE Trans. Circuits Syst. Video Technol.*, 18(2):186–195, 2008.

15 R. Zhang and M. L. Comer. Efficient inter-layer motion compensation for spatially scalable video coding. *IEEE Trans. Circuits Syst. Video Technol.*, 18(10):1325–1334, 2008.

16 T. Zhao, H. Wang, and S. Kwong. Fast inter-layer mode decision in scalable video coding. In *Proc. IEEE ICIP*, pages 4221–4224, 2010.