# FAST SPATIAL LAYER MODE DECISION BASED ON TEMPORAL LEVELS IN H.264/AVC SCALABLE EXTENSION

*Yen-Chieh Wang*(*王彥傑*) , *Zong-Yi Chen*(*陳宗毅*) , *Pao-Chi Chang*(*張寶基*)

Dept. of Communication Engineering, National Central University
E-mail:{yjwang, zychen, pcchang}@vaplab.ce.ncu.edu.tw

## ABSTRACT

Scalable Video Coding (SVC) is getting popular but exhibits a problem of high computational complexity compared with H.264/AVC single layer coding. A fast mode decision algorithm that reduces the candidate modes in motion estimation can certainly reduce its computation time. We propose an adaptive temporal level mode decision algorithm in spatial scalable video coding. In spatial scalability, we utilize the spatial candidate modes table constructed based on the statistics of real data to reduce the candidate modes. In temporal scalability, we apply the base layer mode information directly at low temporal levels. The simulation results show that the proposed algorithm can reduce computation time up to 78.28% with less than 0.14 dB video quality degradation compared with JSVM 9.12.

## 1. INTRODUCTION

Multimedia applications on various devices and heterogeneous networks have been becoming popular. However, these applications usually exist different requirements, such as bandwidth constraints, CPU processing capabilities, and display resolutions, to be satisfied simultaneously. Scalable coding is one of good solutions. SVC is finalized as an extension of H.264/AVC video coding standard in 2007 [1]. A coded SVC bit-stream is composed of a base layer (BL) and several enhancement layers (ELs)[1]. The BL contains a lower resolution or quality version of each coded frame which can be adapted to some resource constrained devices, such as portable phones, PDAs, and laptops. The ELs provide a higher quality service such as various frame rates and display resolutions.

The function of variable block size in H.264/AVC increases the candidate modes during motion estimation (ME) process and results in better rate-distortion (R-D) performance. However, it also increases large amounts of computation complexity accordingly. Similar to H.264/AVC, SVC calculates the R-D cost of each candidate mode and selects the one with minimum R-D cost as the best mode. Additionally, inter-layer prediction in SVC increases the prediction times of all modes. And the inherent hierarchical B Picture (HBP) structure executes ME in forward, backward, and bi-prediction directions. Therefore, the computation complexity of SVC is much higher than H.264/AVC. Consequently, a method which can reduce the complexity with just a little R-D loss is desirable.

Recently, many kinds of fast mode decision schemes have been developed for H.264/AVC. Even though these algorithms are efficient in reducing the computational complexity with negligible quality degradation in H.264/AVC, but inter-layer information is not utilized when they apply to SVC.

The spatial and temporal scalabilities in SVC are widely used in scalable coding, but there are few works which focus on the fast algorithm of these two scalabilities both. Li proposed a fast mode decision algorithm for spatial scalability in SVC [3]. In this scheme, they used the mode distribution relationship between base layer and enhancement layers. In [4], a fast mode decision algorithm had been proposed for inter-frame coding supporting spatial scalability by Li. In [5], Lim proposed a fast algorithm which use the mode history map (MHM) and motion vector predictor (MVP) to reduce and modify the candidate modes. In [6], Lin focused on the CGS and temporal scalabilities, and a fast algorithm which utilizes the features between CGS layers to reduce the computation complexity was proposed. In [7], Kim proposed a fast algorithm which uses the neighboring macroblock (MB) modes to reduce the candidate modes of current MB.

In this work, we develop a fast algorithm which uses spatial features in different temporal levels. The proposed algorithm can save the encoding time significantly while maintaining the R-D performance both in spatial and temporal scalabilities.

The rest of the paper is organized as follows. Section 2 introduces the inter-layer prediction, complexity analysis, and the observations of spatial scalability in different temporal levels. In Section 3, we describe the proposed fast mode decision algorithm. Section 4 shows the experimental results. And we conclude this work in Section 5.

## 2. COMPLEXITY ANALYSIS AND THE OBSERVATIONS IN SPATIAL SCALABILITY

### 2.1. Inter-layer Prediction

There are three inter-layer predictions in spatial scalability [1]: Inter-layer motion prediction, Inter-layer residual prediction, and Inter-layer intra prediction. In residual prediction of JSVM 9.12, the current encoding MB in EL subtracts the corresponding reconstructed upsampled BL MB residual and executes inter prediction to get smaller EL residual. In inter-layer motion prediction, the current MB uses the upsampled BL motion vector (MV) to be the MVP for prediction. In inter-layer intra prediction, the reconstructed intra signal of corresponding MB in BL is upsampled to be the prediction signal of EL. In this work, we focus on inter-layer motion prediction. In section 2.2, we analyzed the complexity of inter-layer prediction.

In the dyadic spatial scalability with the same quantization parameter (QP) in all layers, the modes size between BL and ELs should also have high probability to be dyadic. For example, when the base layer is QCIF and enhancement layer is CIF resolution, then the best mode in the EL will be 16×16 if the BL mode type is 8×8. Similarly, the best mode in EL will be 8×8 if the BL mode type is 4×4. Even though QP values are different in neighboring layers, we observed that this correlation is still very high.

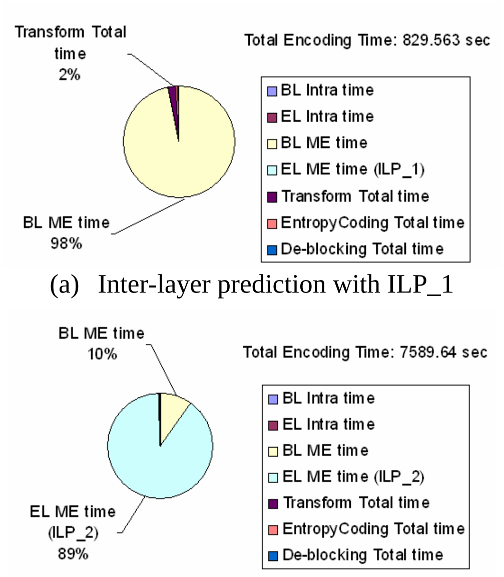
### 2.2. Complexity Analysis

In inter-layer motion prediction, there exist two extreme methods can be used. The first is ILP_1, which copies the upsampled MVs and modes of corresponding MB of BL to be the MVs and modes of EL. The second one is ILP_2, which uses upsampled MVs of corresponding MB of BL to be the predictor and then search all modes. Fig.1 and Fig.2 show the coding complexity and the rate-distortion (R-D) comparison of these two methods.

Because of inter-layer prediction, all modes in EL would be tested more times than BL during ME. In Fig. 1, we can find encoding with ILP_2 increases much more computation complexity than with ILP_1 and ME process in EL occupies almost 90% of total encoding time. However, Fig. 2 shows that the R-D performance improvement encoding with ILP_2 can not be ignored. Hence, encoding with ILP_2 is necessary in general and the speed-up of EL ME is required. In this paper, we concentrate on the fast mode decision in EL. The computation time can be decreased drastically by reducing the candidate modes.

If the modes of EL and BL have strong relationship, we should be able to find methods to achieve the following objectives:

1) Reducing the computation time of ME
2) Performance will be very close to the result of ILP_2



(a) Inter-layer prediction with ILP_1

(b) Inter-layer prediction with ILP_2

Fig. 1: The complexity of ILP_1 and ILP_2

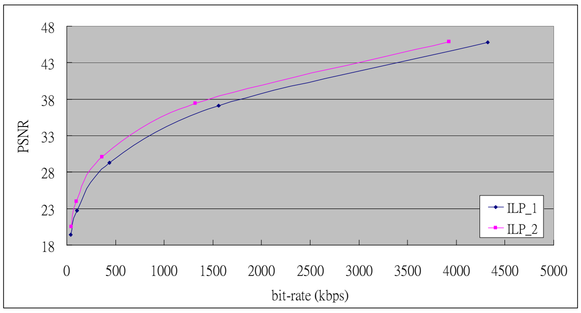


Fig. 2: R-D performance of ILP_1 and ILP_2

### 2.3. Correlations between Base and Enhancement Layers

In this section, we analyze the correlations between the spatial layers in different temporal levels. The statistical analyses are conducted based on the coding of one base layer and one enhancement layer with QCIF and CIF resolution respectively. For simplicity, the QPs in two layers are set to the same values, GOP size is set to 8 which indicates that four temporal layers are provided, and full search is employed with search range 32.

#### *2.3.1 Spatial scalability*

Because of the dyadic resolutions, the sampling rate in EL is two times to BL in each direction. The MB mode in EL is strongly related to that in BL. Table 1 shows the conditional probability of MB modes in spatial layers when BL is coded with different block types. Table 1 indicates that under the specific MB mode in BL, the corresponding probabilities of each coded mode in EL. The highlight indicates the highest probability for each case. From Table 1, we can observe the MB mode size in EL tends to be larger than the corresponding MB mode size in BL, i.e. if the best mode is Block_8x8 in BL, the best mode in EL almost ranges from Mode_SKIP to Block_8x8. And when MB

Table 1: Modes correlation in BL and EL

| Enhancement Layer MB Mode | MB Mode in Base Layer | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Average, QP_BL = QP_EL = 25, GOP = 8 | | | | | | | | | | |
| | Mode _SKIP | Mode _16x16 | Mode _16x8 | Mode _8x16 | Blk _8x8 | Blk _8x4 | Blk _4x8 | Blk _4x4 | Blk _skip | Intra _4x4 | Intra _16x16 |
| Mode_SKIP | 0.52 | 0.26 | 0.15 | 0.16 | 0.08 | 0.02 | 0.02 | 0 | 0.27 | 0 | 0.06 |
| Mode_16x16 | 0.4 | 0.54 | 0.45 | 0.46 | 0.42 | 0.07 | 0.09 | 0.04 | 0.42 | 0.03 | 0.03 |
| Mode_16x8 | 0.03 | 0.08 | 0.17 | 0.1 | 0.15 | 0.44 | 0.05 | 0.07 | 0.09 | 0.02 | 0.03 |
| Mode_8x16 | 0.03 | 0.06 | 0.1 | 0.13 | 0.14 | 0.06 | 0.41 | 0.08 | 0.07 | 0.04 | 0 |
| Blk_8x8 | 0.01 | 0.02 | 0.05 | 0.06 | 0.09 | 0.19 | 0.19 | 0.45 | 0.05 | 0 | 0 |
| Blk_8x4 | 0 | 0.01 | 0.02 | 0.02 | 0.03 | 0.08 | 0.06 | 0.12 | 0.02 | 0 | 0 |
| Blk_4x8 | 0 | 0.01 | 0.01 | 0.02 | 0.03 | 0.05 | 0.08 | 0.11 | 0.02 | 0 | 0 |
| Blk_4x4 | 0 | 0 | 0.01 | 0.01 | 0.01 | 0.02 | 0.03 | 0.05 | 0.01 | 0 | 0 |
| Blk_skip | 0.01 | 0.02 | 0.04 | 0.04 | 0.05 | 0.07 | 0.07 | 0.09 | 0.05 | 0 | 0 |
| Intra_4x4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.03 |
| Intra_BL | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.87 | 0.86 |
| Intra_16x16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

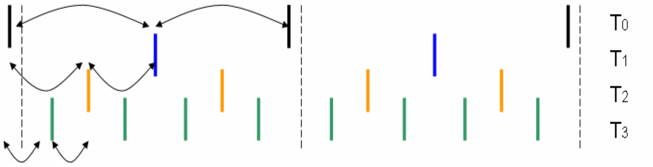


Fig. 3: Hierarchical B Picture structure and temporal distance of the reference frames

in BL is coded as Mode_16x8/Mode_8x16, the best mode in EL is unlikely to be Mode_8x16/ Mode_16x8. When BL MB is coded as Intra mode, EL MB has higher probability to be coded as Intra mode.

*2.3.2 Temporal scalability*

Fig. 3 shows the HBP structure, $T_0$ and $T_3$ denote the lowest and the highest temporal level (TL) in this structure. The temporal distance of reference frames increase with the decrement of TLs, i.e. the distance between reference frames and the current frames in $T_1$ is four, and that is one in $T_3$. Compared with higher TLs, the lower TLs have stronger spatial layer dependency than temporal layer dependency. Thus we can utilize the spatial dependency directly in low TLs. That is to say we can save the computation complexity by separating temporal levels into low and high parts. However, according to the encoding order, the distortion of low TLs will propagate to higher TLs, so TL1 is more important than TL2 and TL3 by considering the ME references. If we want to speed up the computation time in low TLs, we must reduce the distortion of TL1 to maintain the overall performance by any means.

## 3. THE PROPOSED FAST MODE DECISION ALGORITHM

By the analysis in Section 2, we propose a combined spatial and temporal fast mode decision algorithm. Fig. 4 shows the flow chart of the proposed algorithm, and Table 2 is the spatial candidate modes table.

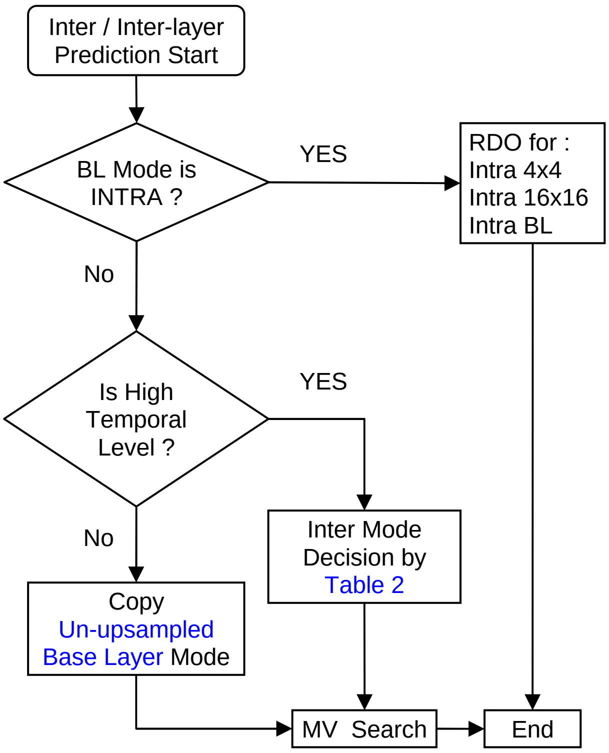


Fig. 4: Flow chart of the proposed algorithm

Table 2: Spatial candidate modes table

| EL MB Mode | Base Layer MB Mode | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MODE_SKIP | MODE_16x16 | MODE_16x8 | MODE_8x16 | BLK_8x8 | BLK_8x4 | BLK_4x8 | BLK_4x4 | BLK_SKIP |
| MODE_SKIP | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| MODE_16x16 | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ | ★ |
| MODE_16x8 | | | ★ | | ★ | ★ | ★ | ★ | ★ |
| MODE_8x16 | | | | ★ | ★ | ★ | ★ | ★ | ★ |
| BLK_8x8 | | | | | ★ | ★ | ★ | ★ | ★ |
| BLK_8x4 | | | | | | ★ | | ★ | |
| BLK_4x8 | | | | | | | ★ | ★ | |
| BLK_4x4 | | | | | | | | ★ | |
| BLK_SKIP | | | | | ★ | ★ | ★ | ★ | ★ |

In Table 2, based on the collocated MB mode in BL, EL MB only executes ME with the modes labeled by star.

In our proposed algorithm, when the corresponding BL MB mode is intra coded, the candidates in EL are reduced to only three intra modes. Otherwise, we separate the TLs into low and high first. If the current encoding frames belong to low TL, we copy the BL mode as the mode in EL directly. If the TL is high, the spatial candidate modes table is used to reduce the candidate modes.

By the observation in Section 2.3.2, in order to improve the video quality of low TLs in our proposed algorithm, we copy the un-upsampled mode of BL mode in EL directly. Table 3 shows the R-D performance of using upsampled and un-upsampled mode respectively in low TLs without using spatial candidate modes table and intra modes reduction. The distortion of TL1 using un-upsampled mode is less than using upsampled mode. Therefore, we use the un-upsampled mode of BL mode directly to be the EL mode. By experiments, we find that the performance is better when we regard the temporal distance farther than three frames as low TLs.

Table 3: Performance comparison of BL mode copy (Sequence: BUS)

| | UP | | | Un-up | | |
|---|---|---|---|---|---|---|
| QP | PSNR (dB) | Bit-rate (Kbps) | Time (sec) | PSNR (dB) | Bit-rate (Kbps) | Time (sec) |
| 15 | 41.839 | 2431.678 | 2571.8 | 41.842 | 2426.386 | 2643.56 |
| 25 | 34.705 | 817.526 | 1916.4 | 34.715 | 815.046 | 1899.3 |
| 35 | 28.131 | 244.582 | 1578.5 | 28.140 | 244.485 | 1576.53 |

## 4. EXPERIMENTAL RESULTS

For the experiments, the proposed algorithm is implemented with JSVM 9.12 and tested using five video sequences in two scenarios. With two spatial layers, the HBP structure is employed with GOP size equaling to 8 and 16. More encoder configurations are stated in Table 4.

Table 4: Test Environments

| | Scenario #1 | | Scenario #2 | |
|---|---|---|---|---|
| Layer | BL | EL | BL | EL |
| QP | 15, 20, 25, 30, 35, 40 | | 15, 20, 25, 30, 35, 40 | |
| Resolution | QCIF | CIF | CIF | 4CIF |
| Frame rate | 15 | 15 | 30 | 30 |
| Frames | 150 | | 300 | |
| Sequence | BUS, MOBILE, SOCCER | | CITY, CREW, SOCCER | |
| ◎ JSVM 9.12 | | | ◎ GOP : 8, 16 | |
| ◎ Full Search with Search Range = 32 | | | ◎ Reference Frame : 1 | |

In Table 5, the proposed scheme is compared with JSVM 9.12 in terms of *d_PSNR*, *d_Bitrate*, *TS*, and $TS_e$, which are defined as follows.

$$d\_PSNR = PSNR_{proposed} - PSNR_{JSVM}$$

$$d\_Bitrate = \frac{bitrate_{proposed} - bitrate_{JSVM}}{bitrate_{JSVM}} \times 100\%$$

$$TS = \frac{run\_time_{proposed} - run\_time_{JSVM}}{run\_time_{JSVM}} \times 100\%$$

$$TS_e = \frac{EL\_ME\_run\_time_{proposed} - EL\_ME\_run\_time_{JSVM}}{EL\_ME\_run\_time_{JSVM}} \times 100\%$$

In Table 5, only the results of three QPs and GOP = 8 are shown because of the limited space. The proposed scheme averagely provides 71% and 75% overall time saving with the 80% and 85% EL ME time saving in scenario one and two respectively. Moreover, the average bit-rate increase is no more than 1.26% and 0.26% with negligible PSNR degradation in scenario one and two. The results of proposed algorithm compared with ILP_1 are shown in Table 6. With the significant improvement in R-D performance, the encoding time of our proposed algorithm is close to ILP_1. And the results with GOP = 16 and other QPs have similar performance both in rate-distortion and time saving mentioned above.

In Fig. 5, we compare the total encoding time of three methods:

1) JSVM 9.12
2) the scheme use the spatial candidate modes table only
3) the proposed algorithm

Obviously, the saving time does not vary acutely with the change of QP, and the executing time of method 2 is close to the time of method 3 in higher QP. There are two reasons for this situation. First, the spatial base layer is a downsampled version of the original picture and lots of residual signals tend to be quantized to zero under high QP. The detail in the picture, which is the high frequency part, is filtered because of the downsampling process and the residual signal has no significant difference between using large and small block sizes, which implies that the bitrate plays a more important role than distortion does in the mode decision cost function. Consequently the best MB mode in BL is likely to be encoded with larger MB modes. Second, the spatial candidate mode in EL is just the un-upsampled mode in low TL in the proposed algorithm. Therefore, the encoding time of method 2 and 3 becomes closer in higher QP.

## 5. CONCLUSIONS

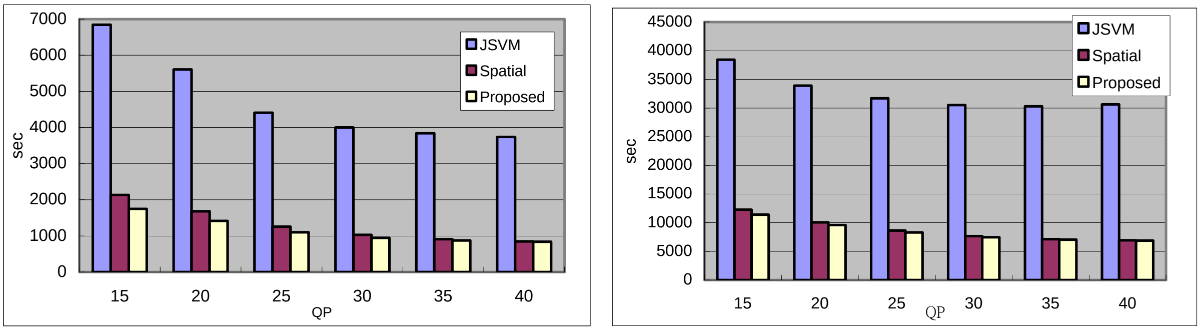
In this paper, we propose an adaptive temporal level mode decision algorithm in spatial scalable video coding. In spatial scalability, the spatial candidate modes table is used to reduce the candidate modes. In temporal scalability, we separate TLs into low and high to decide whether the mode of BL to be copied directly or not. The proposed algorithm can save up to 78.28% computation complexity while maintaining very good rate-distortion performance at the same time.

## ACKNOWLEDGEMENT

This work was supported in part by the NSC, Taiwan under Grant 96-2221-E-008-016-MY2.

Table 5: Performance comparisons with GOP size is 8

| GOP = 8 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SCENARIO #1 | | | | | | SCENARIO #2 | | | | | |
| Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) | TS$_e$(%) | Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) | TS$_e$(%) |
| BUS | 20 | -0.04 | 2.11 | -66.42 | -73.55 | CITY | 20 | -0.02 | 0.69 | -75.11 | -85.55 |
| | 30 | -0.1 | 2.22 | -72.07 | -82.2 | | 30 | -0.04 | 0.45 | -77.01 | -88.5 |
| | 40 | -0.14 | 0.49 | -76.09 | -87.82 | | 40 | -0.03 | -0.22 | -77.97 | -89.19 |
| MOBILE | 20 | -0.05 | 1.67 | -72.13 | -78.92 | CREW | 20 | -0.04 | 0.52 | -72.9 | -81.23 |
| | 30 | -0.1 | 1.69 | -75.32 | -85.55 | | 30 | -0.06 | 0.08 | -76.23 | -87.19 |
| | 40 | -0.1 | -0.05 | -77.32 | -88.71 | | 40 | -0.04 | -0.14 | -78.28 | -89.09 |
| SOCCER | 20 | -0.08 | 1.44 | -65.34 | -73.53 | SOCCER | 20 | -0.04 | 0.63 | -71.75 | -81.14 |
| | 30 | -0.09 | 0.78 | -72.4 | -82.72 | | 30 | -0.06 | 0.36 | -75.58 | -86.59 |
| | 40 | -0.1 | -0.49 | -76.72 | -87.82 | | 40 | -0.05 | -0.27 | -77.51 | -88.71 |



(a) MOBILE in Scenario #1

(b) SOCCER in Scenario #2

Fig. 5: The encoding time histogram of three methods

Table 6: Performance comparison of ILP_1 and the proposed algorithm

| ILP_1 | | | | | Proposed Algorithm | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SCENARIO #1 | | | | | | | | | |
| Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) | Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) |
| BUS | 15 | -0.27 | 12.71 | -92 | BUS | 15 | -0.02 | 1.76 | -65.54 |
| | 25 | -0.75 | 15.1 | -89.42 | | 25 | -0.07 | 2.13 | -68.87 |
| | 35 | -1.21 | 8.82 | -87.26 | | 35 | -0.14 | 1.61 | -74.47 |
| MOBILE | 15 | -0.21 | 10.34 | -93.29 | MOBILE | 15 | -0.03 | 1.44 | -71.18 |
| | 25 | -0.59 | 13.72 | -89.63 | | 25 | -0.07 | 1.73 | -73.22 |
| | 35 | -1.43 | 6.06 | -88.74 | | 35 | -0.13 | 1.08 | -76.7 |
| SOCCER | 15 | -0.56 | 13.68 | -90.42 | SOCCER | 15 | -0.05 | 1.56 | -62.88 |
| | 25 | -1.04 | 7.17 | -88.64 | | 25 | -0.09 | 1.19 | -69.05 |
| | 35 | -0.83 | -1.39 | -87.86 | | 35 | -0.1 | 0.36 | -75.04 |
| SCENARIO #2 | | | | | | | | | |
| Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) | Sequence | QP | d_PSNR(dB) | d_Bitrate(%) | TS(%) |
| CITY | 15 | -0.15 | 8.52 | -90.2 | CITY | 15 | -0.03 | 0.34 | -74.37 |
| | 25 | -0.66 | 10.43 | -87.82 | | 25 | -0.03 | 0.68 | -76.29 |
| | 35 | -1.29 | 1.7 | -87.37 | | 35 | -0.04 | -0.05 | -77.44 |
| CREW | 15 | -0.25 | 3.9 | -91.67 | CREW | 15 | -0.04 | 0.43 | -70.66 |
| | 25 | -0.43 | 2.59 | -88.85 | | 25 | -0.05 | 0.37 | -75.01 |
| | 35 | -0.77 | -3.75 | -87.77 | | 35 | -0.05 | -0.16 | -77.21 |
| SOCCER | 15 | -0.23 | 6.68 | -90.08 | SOCCER | 15 | -0.03 | 0.4 | -70.29 |
| | 25 | -0.63 | 4.21 | -88.11 | | 25 | -0.05 | 0.57 | -73.84 |
| | 35 | -0.78 | -3.54 | -87.62 | | 35 | -0.06 | 0.05 | -76.73 |

## REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, Sept. 2007.

[2] A. Segall and G. J. Sullivan “Spatial Scalability Within the H.264/AVC Scalable Video Coding Extension,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1121–1135, Sep. 2007.

[3] H. Li, Z. G. Li, and C. Wen, “Fast Mode Decision Algorithm for Inter-Frame Coding in Fully Scalable Video Coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, No. 7, pp. 889 -895, July 2006.

[4] H. Li, Z. G. Li, C. Wen, and L. P. Chau, “Fast mode decision for spatial scalable video coding,” *IEEE International Symposium on Circuits and Systems*, pp. 3005-3008, May 2006.

[5] S. Lim, J. Yang, and B. Jeon, “Fast coding mode decision for scalable video coding,” *International Conference on Advanced Communication Technology (ICACT)*, vol.3, pp.1897-1900, Feb. 2008.

[6] 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 the combination of coarse grain scalability(CGS) and temporal Scalability,” IEEE International Conference on Image Processing, pp. 289-292, Sept. 2007.

[7] S.T. Kim, Krishna Reddy Konda and C.-S. Cho, “Fast Mode Decision Algorithm for Spatial and SNR and Scalable Video Coding,” *IEEE International Symposium on Circuits and Systems (ISCAS 2009)*, pp. 872 – 875, May 2009.