MOTION ALIGNED SPATIAL SCALABLE VIDEO CODING

Debing Liu[1], *Yuwen He*[2], *Shipeng Li*[2], *Debin Zhao*[1], *Wen Gao*[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
Email: {dbliu, dbzhao, wgao}@jdl.ac.cn
[2]Microsoft Research Asia, Beijing 100080, China
Email: heyuwen@tsinghua.org.cn, spli@microsoft.com

ABSTRACT

A motion aligned spatial scalable video coding scheme (MA-SSC) is proposed in this paper. Different from the traditional spatial scalable coding schemes derived from MPEG-2, in the proposed scheme only one set of intra or inter prediction modes are optimally selected by jointly considering the base and enhancement layers. Thus, it saves one set of macroblock (MB) mode and motion vectors. Moreover, the combined motion estimation can reduce the residual coding bits of the base layer. The MA-SSC and traditional spatial scalable coding schemes are both implemented based on H.264 reference software to evaluate their performance. Simulation results show that the enhancement layer coding efficiency of MA-SSC is up to 0.6dB better than that of the traditional scheme, while the base layer coding efficiency of MA-SSC decreases less than 0.3db compared with the single-layer coding.

1. INTRODUCTION

Due to the fast growing of multimedia terminals with different resolutions and different band-width connections, spatial scalability has been recognized as an important functionality of video coding both in industry and academy. MPEG-2 [1] and MPEG-4 [2] have provided spatial scalabilities in their scalability profiles, and MPEG-21 [3] has issued call for the proposals for spatial scalable coding. The spatial scalable coding has two performance bounds. The upper-bound is the single-layer coding case where only the high resolution video is encoded with a non-scalable codec. And the lower-bound is the simulcast case where video in the high and low resolutions are simply encoded independently to provide straightforward spatial scalability. Because the low resolution video is down-sampled from the high resolution video, there is much redundancy between them in the simulcast case. The goal of spatial scalable coding is to reduce the redundancy and make its coding efficiency close to that of single-layer coding. There is an obvious performance gap between the traditional spatial scalable coding and non-scalable coding schemes, so many methods have been proposed to improve its coding performance. Most of them [4] [5] [6] followed the basic framework of MPEG-2 spatial scalable profile. The framework derived from MPEG-2 spatial scalable coding scheme is easy to be implemented with a non-scalable codec and its base layer is fully compatible with single-layer coding. Those make it be used widely. However, the performance loss of the spatial scalable coding is mainly from three kinds of redundancies: the prediction residuals, motion vectors and MB coding modes. The MPEG-2 scheme only partially reduces the residual redundancy while it introduces some additional side information along with the reduction. So its coding efficiency has no significant improvement over the simulcast case. In order to further reduce the redundancy, Benzler [7] proposed a combined subband-DCT scheme. It uses a 4-band analysis filter to decompose the input images to one low-frequency and three high-frequency spatial subband images. The low-frequency subband images are encoded in the base layer and the other three high-frequency subband images are encoded in the enhancement layer. It obtains the motion vectors from the base layer motion estimation and applies the same motion vectors for the enhancement layer motion compensation. Thus, it removes the three kinds of redundancies. But Benzler's scheme requires the quantization steps of the base and enhancement layers to be highly related, so the bit-rate between these two layers cannot be allocated arbitrarily. To solve the problem, it implements additional SNR scalability in the base layer, but it results in further coding efficiency loss and can only control the bitrate distribution coarsely.

In our proposed scheme, the base and enhancement layers use the same motion vectors and intra prediction directions that are obtained by the combined motion estimation and intra prediction, respectively. Moreover, it has been observed that the combined motion estimation can highly reduce the residual coding bits of the base layer. Thus, the mode, motion vectors and residual coding bits of the base layer can be saved. And the proposed scheme has no limitation in bitrate control which is an unavoidable problem in Benzler's 4-band scheme.

This paper is organized as follows. In Section 2 the MA-SSC scheme will be proposed. We will introduce the overall coding framework and the major difference to the traditional framework derived from MPEG-2. In Section 3 experimental results are presented. And we conclude this paper in Section 4.

2. MOTION ALIGNED SPATIAL SCALABLE VIDEO CODING

2.1 Coding Framework

Figure 1 illustrates the traditional spatial scalable coding framework derived from MPEG-2. Its base layer is fully compatible with the single-layer coding and can be implemented with any non-scalable codecs. The original input images are down-sampled and the down-sampled images will be encoded in the base layer. For the enhancement layer coding, the difference to the non-scalable coding is the motion compensation module. Besides the prediction selected from the intra or inter prediction modes within the enhancement layer itself, there are another two predictions. One is the up-sampled picture from the base layer reconstructed frame and the other is the average combination (C-MCP) of the former two predictions. It is the two additional predictions that partially reduce the residual redundancy between the base and enhancement layers. Obviously it needs some side information

*This work was done when the author was with Microsoft Research Asia.

to indicate which prediction is selected for the enhancement layer motion compensation.

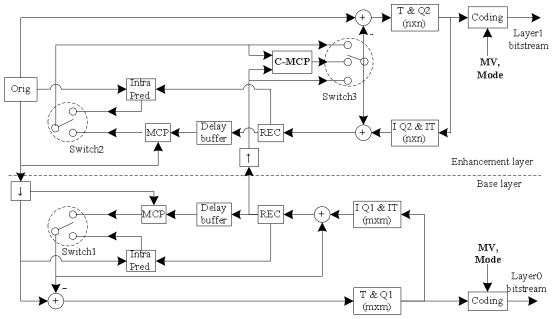
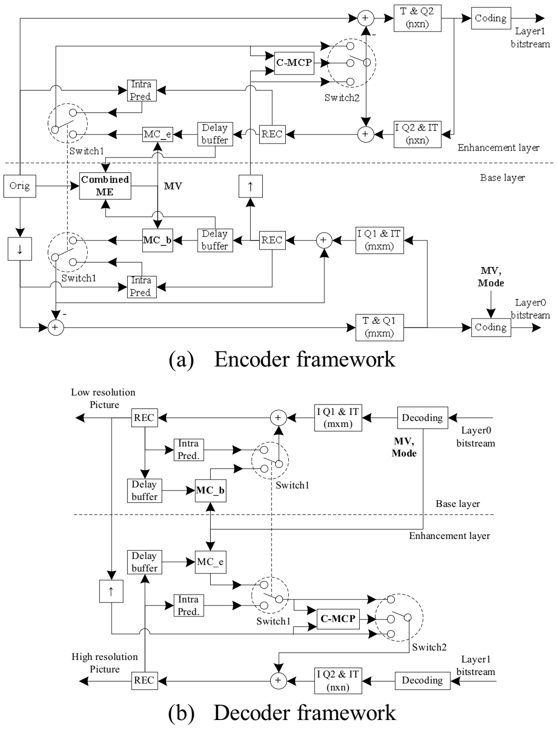


Figure. 1. Traditional coding framework in MPEG-2

Figure 2 illustrates the coding framework of our proposed scheme. For clarity, there are only two layers in the figure. It can be extended to multi-layers if necessary. Compared with the traditional framework in MPEG-2, it has following unique features:

(1) Intra prediction and motion estimation are performed by jointly considering the base layer and the enhancement layer (Combined ME).
(2) The base and enhancement layers share the intra or inter mode selections (Switch1) and only one set of MB's coding modes is needed.
(3) An additional interpolation process is applied in the base layer when it performs motion estimation and motion compensation (MC_b).



(a) Encoder framework

(b) Decoder framework

Figure 2. Coding framework of MA-SSC

From the frameworks of the two schemes, obviously our proposed scheme will save one set of MB coding modes and motion vectors. By implementing the combined ME with the consideration of both base and enhancement layers, we can get the optimal mode and motion vectors for both layers that could further improve the coding efficiency. We will explain it in the following sections.

2.2 Combined Motion Estimation and Mode Selection

The down-sampled image and the original image are highly correlated. From information theory point of view, it is almost true that the down-sampled one is a subset of the original image if there is no strong frequency aliasing due to down-sampling. So it is reasonable to use only one set of motion vectors for the motion compensation of both layers. Now the problem is focused on how to get the motion vectors.

If we obtain the motion vectors just from the base layer motion estimation, then the base layer coding efficiency should be the same as that of the single-layer coding. However since most high frequency information in the original picture is lost during the down-sampling process, when the same motion vectors gotten from the base layer are up-sampled for the enhancement layer coding, it will largely decrease the enhancement layer motion compensation (MC) efficiency and in turn the coding efficiency. On the other hand, if we obtain the motion vectors just from the enhancement layer motion estimation, it cannot guarantee the MC efficiency of the base layer. Furthermore, since the reconstructed base layer frame will be up-sampled as an additional prediction for the enhancement layer coding, the coding efficiency degradation of the base layer will reduce its contribution for the enhancement layer coding. To balance the coding efficiency of the two layers, a combined motion estimation is needed.

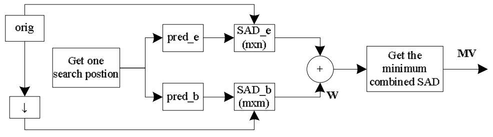


Figure 3. Combined motion estimation

Figure 3 illustrates the combined motion estimation process, in which "pred_b" and "pred_e" are the prediction from the base and enhancement layer reference frames, respectively. And "SAD_b", "SAD_e" are the sum of absolute (prediction) difference (SAD) of the base and enhancement layers, respectively. After all positions have been checked, the motion vector with the minimum cost is selected. Due to the different resolutions in the base and enhancement layer reference frames, "SAD_b" and "SAD_e" are calculated in different block size. For every search position, the cost function is as follows:

$$Cost_{\text{inter}} = SAD\_e + w \times SAD\_b + C \times \lambda \times bits(MV) \quad (1)$$

where *w* is a scale factor for the base layer. For better performance, *w* should vary along with the quality difference between the base layer and the enhancement layer (In our simulations, *w* is set to 1.). And *λ×bits(MV)* means the cost of the motion vector. Its calculation is the same as the single-layer coding. We add the coefficient *C* to control the motion vector coding bits (In our simulations, *C* is set to 1.3).

It's common sense that in single-layer coding, if the value of *C* is decreased, it tends to obtain more accurate motion vectors, which makes the motion vector coding bits increased and the residual coding bits reduced; on the contrary, if the value of *C* is increased, the motion vector coding bits will be reduced and the residual coding bits will be increased. It won't influence the coding efficiency very much if we change the value of *C* in a limited range. Obviously it's no use in single-layer coding. However in our proposed scheme, since the motion vectors are in high resolution, if we increase the value of *C*, the motion vectors are still accurate enough for the base layer, so the base layer residual coding bits won't be increased very much. Since the motion vectors will be encoded in the base layer, if the value of *C* is increased, it will be very useful to improve the base layer coding efficiency while keeping the enhancement layer coding efficiency.

To match with the cost calculation of the combined motion estimation, the intra prediction of MA-SSC also changes. Its cost function is as follows:

$$Cost_{intra} = SAD\_e + w \times SAD\_b, \tag{2}$$

In the proposed framework (Figure 2), it should be noted that the base and enhancement layers share the same intra or inter modes (Switch1). In fact, they are determined by the values of $Cost_{inter}$ and $Cost_{intra}$. If $Cost_{inter}$ is less than $Cost_{intra}$, then the MB coding modes are inter mode, otherwise, the intra prediction modes are selected.

2.3 Motion Compensation in Base Layer

Through the combined ME, a set of motion vectors in high resolution will be obtained and coded in the base layer. Compared with single-layer coding, the overhead of motion vectors for base layer will be increased. The increasing is from two aspects. One is the number of the motion vectors and the other is the precision of every MV.

For the first, the combined motion estimation can decrease the number of motion vectors much(see section 2.2), more over, the increased number of motion vectors makes the block size of inter prediction smaller for the base layer, so the inter prediction is more accurate and the corresponding base layer residual coding bits will be reduced. Thus the base layer coding efficiency degradation is not much because of the increased number of motion vectors.

For the latter, an additional sub-pixel interpolation is performed in the base layer to fully take advantage of the more accuracy bits of the motion vectors. In our simulation, the base and enhancement layers are QCIF and CIF format, respectively. That means the 1/4 pixel motion vector in the enhancement layer will have the 1/8 pixel accuracy in the base layer. A bilinear interpolation is performed on the base layer reference frames. The added interpolation can reduce the base layer residual coding bits effectively.

3. EXPERIMENTAL RESULTS

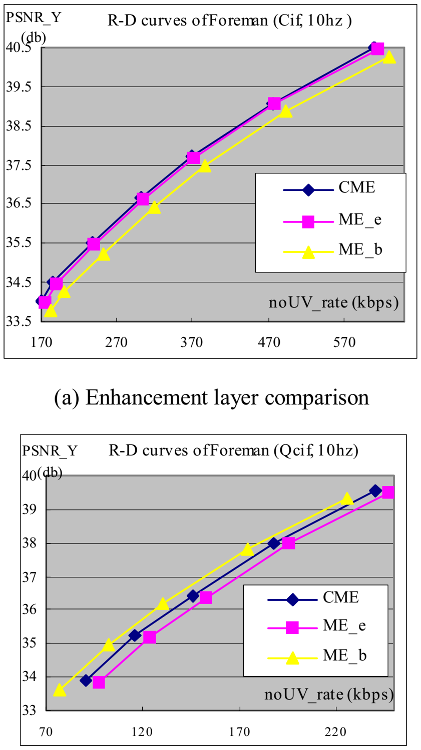
In this section, we will evaluate the performance of CME and compare the coding efficiency of MA-SSC with the traditional scheme derived from MPEG2 (in all the following tables and figures, it will be labeled as MPEG2). Simulations are performed on the JVT software JM 93. We use standard DCT to down-sample the original signals to get the low resolution signals [4]. The base and enhancement layers are in QCIF and CIF formats, respectively. The related coding conditions are as follows:

Table 1. Conditions of the experiments

RDO	Off
Inter search mode	16x16, 16x8, 8x16, 8x8
Search range	8
Transform8x8Mode	Only 8x8 transform
Entropy coding method	UVLC
Frame type	The First is I and the others are P frames.
Reference frame number	1

Test 1. Combined motion estimation

This test evaluates the performance of the combined motion estimation (CME). The method CME will be compared with the case that only use the enhancement layer to do motion estimation (ME_e) and the case that only use the base layer to do motion estimation (ME_b). In figure 4 (a), we fix the base layer QP to 26 and change the enhancement layer QP from 22 to 34 by step 2. It shows that CME and ME_e are much better than ME_b (about 0.4 db). In figure 4 (b), we set the identical QP for the base and enhancement layers and change the QP from 22 to 30 by step 2. It shows that the base layer coding efficiency of CME will be improved about 0.3 db above ME_e while decreased about 0.3db on ME_b. As a result, CME is the best.



(a) Enhancement layer comparison

(b) Base layer comparison

Figure 4. Coding efficiency of combined motion estimation

Test 2. Base layer coding efficiency of MA-SSC

This test compares the base layer coding efficiency of MA-SSC with that of non-scalable coding which is the base layer of MPEG2 scheme. In figure 5, we set the identical QP for the base and enhancement layers and change the QP from 22 to 30 by step 2. It shows that the base layer coding efficiency of MA-SSC will decrease about 0.3 db on the non-scalable coding. That's because the mode and motion vectors are in high resolution and they will be encoded in the base layer.

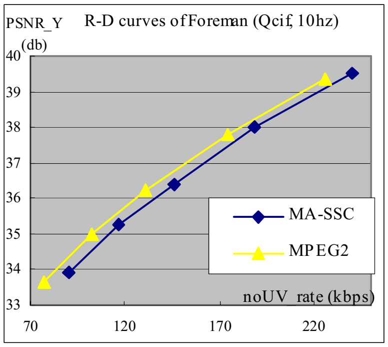


Figure 5. Comparison of the base layer coding efficiency of MA-SSC

Test3. Enhancement layer coding efficiency of MA-SSC

This test compares the enhancement layer coding efficiency of MA-SSC with that of single-layer coding, simulcast and traditional spatial scalable coding scheme derived from MPEG-2. In the test, the base layer QP is fixed to 26 and the enhancement layer QP is changed from 22 to 34 by step 2. Table 2 lists the detail coding information of those methods, in which xx_e and xx_b means xx of enhancement layer and base layer, respectively. Ybit is residual bitrate of Y component, “Mode+MV” is the bitrate of mode and motion vectors, and the “other” includes the header, CBP, delta quant and the additional side information. The base layer PSNRs of MA-SSC and traditional scheme are 36.41dB and 36.35dB, respectively.

Table 2. Detailed comparison of coding efficiency (foreman, CIF, 10hz)

	QP	PSNR_e	Ybit_e (kbps)	Ybit_b (kbps)	Mode+ MV_e (kbps)	Mode+ MV_b (kbps)	Other (kbps)	Total Rate (kbps)
Single-layer	22	40.47	437.38	0	79.26	0	25.65	542.29
	26	37.67	222.43	0	63.56	0	19.19	305.18
	30	35.29	119.55	0	49.85	0	14.62	184.02
MA-SSC	22	40.53	398.6	81.73	0	66.69	63.42	610.44
	26	37.77	176.77	81.75	0	50.18	60.13	368.83
	30	35.52	65.91	81.12	0	34.16	55.08	236.27
MPEG2	22	40.51	400.8	98.22	78.22	17.7	60.79	655.73
	26	37.73	179	98.22	61.1	17.7	56.25	412.27
	30	35.53	72.05	98.22	45.76	17.7	52.03	285.76
Simulcast	22	40.47	437.38	98.22	79.26	17.7	31.96	664.52
	26	37.67	222.43	98.22	63.56	17.7	25.5	427.41
	30	35.29	119.55	98.22	49.85	17.7	20.93	306.25

From table 2, it can be seen that the bits saving of MA-SSC is mainly from the mode, motion vector and base layer residual bits when compared with the traditional spatial scalable coding scheme.

Figure 6 shows the enhancement layer R-D curves of Crew, Foreman, Soccer and Coastguard. It shows that the enhancement layer coding efficiency of MA-SSC is totally higher (about 0.6db) than that of the traditional scheme derived from MPEG2 and it is about 0.5db less than single-layer coding which is the upper-bound of the spatial scalable coding. It should be noted that the improvement is much larger in the low bitrate. That’s because the bits saving of MA-SSC is mainly from the coding mode, motion vector and residual bits in the base layer and those are more significant in low bitrate.

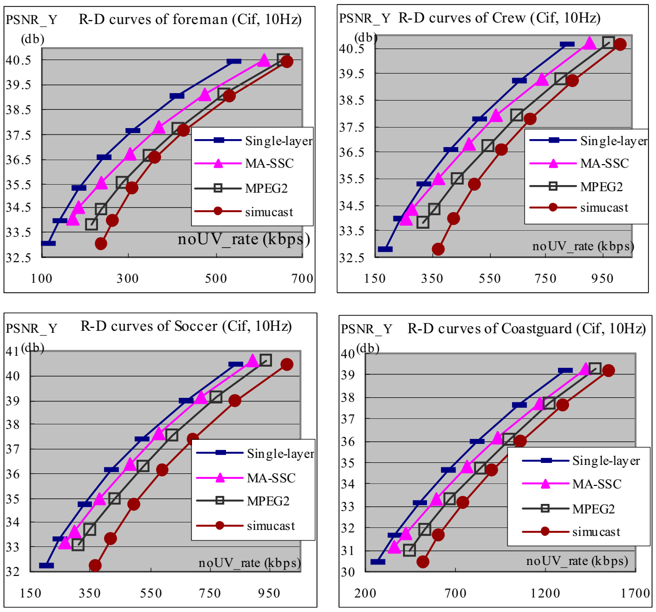


Figure 6. Comparison of enhancement layer coding efficiency

4. CONCLUSIONS

A motion aligned spatial scalable coding scheme (MA-SSC) is proposed in this paper. Compared with the traditional scheme derived from MPEG-2, it saves the coding of one set of mode and motion vectors. Moreover, the mode and motion vectors in high resolution can reduce the residual coding bits of the base layer. Those make the enhancement layer coding efficiency of MA-SSC achieves about 0.6db gains than that of the traditional scheme derived from MPEG2. As for the base layer, although the mode and motion vectors in high resolution increase the overhead, the base layer coding efficiency only decreases about 0.3db than that of non-scalable coding with the help of the combined motion estimation and the additional interpolation in the base layer reference frame.

5. REFERENCES

[1] Revised Text for ITU-T recommendation H.262-ISO/IEC 13818-2: Information Technology-Generic Coding of Moving Pictures and Associated Audio Information: Video, MPEG-2: ISO/IEC JTC1/SC29/WG11, Mar. 1995.

[2] ISO/IEC 14496-2: Information Technology-Coding of Audio-Visual Objects-Part2: Visual, MPEG-4: ISO/IEC JTC1/SC29/WG11, Dec. 1999.

[3] ISO/IEC N6025, “Requirements and Applications for Scalable Video Coding”, ISO/IEC JTC1/SC29/WG11, Oct. 2003.

[4] R. Dugad, N. Ahuja, “A Scheme for Spatial Scalability Using Nonscalable Encoders”, IEEE Transactions on Circuits and Systems for Video Technology, Vol 13, NO. 10. Oct. 2003, pp. 993-999.

[5] Ł. Błaszak, M. Domański, S. Maćkowiak, “Spatio-Temporal Scalability in AVC codecs”, ISO/IEC JTC1/SC29/WG11, MPEG2003/M9469, Mar. 2003.

[6] R. Lange, Ł. Błaszak, M. Domański, “Simple AVC-based codecs with spatial scalability”, 2004 International Conference on Image Processing.

[7] U. Benzler, “Spatial Scalable Video Coding Using a Combined Subband-DCT Approach”, IEEE Transactions on Circuits and Systems for Video Technology, VOL. 10, No.7, Oct. 2000, pp. 1080-1087.