Advanced Spatial and Temporal Direct Mode for B Picture Coding

Yue Wang
Graduate University of Chinese Academy of Sciences
Beijing 100086,China
wangyue@jdl.ac.cn

Li Zhang, Siwei Ma and Wen Gao
Institute of Digital Media, School of Electronic Engineering and Computer Science
Peking University, Beijing 100871, China
{lzhang, swma, wgao }@pku.edu.cn

***Abstract*—**The direct mode in H.264/AVC can efficiently improve the coding performance of B pictures, since it exploits the spatial or temporal correlation by deriving its motion vector from previously encoded information. Therefore, it does not require any additional motion information and could save many bits. Considering the spatial and temporal correlation has not been fully exploited in the current direct mode, in this paper, we propose an advanced Spatial and Temporal Direct Mode (STDM) for B picture coding. The motion vector is selected from a set of spatial-temporal neighboring motion vectors, and the selection criterion is to minimize a spatial-temporal cost function. The framework of Decoder-side Motion Vector Derivation (DMVD) is utilized, where encoder and decoder use the same derivation process to obtain motion vectors, thus no index for the chosen motion vectors need to be coded and transmitted. Simulation results show that the proposed method significantly outperforms the current SDM and TDM in H.264/AVC.

I. INTRODUCTION

In the current H.264/AVC video coding standard [1], B pictures can achieve higher compression ratio by more effectively exploiting the temporal correlation between the reference pictures and the current B picture [2]. Therefore, B pictures play an important role in video coding in term of both coding performance and video transmissions system.

One of the most efficient tools for B picture coding is direct mode [3], where motion vectors of Direct Macroblock (MB) or block are derived from previously encoded information. Thus, its motion information can be omitted. H.264/AVC supports two types of direct mode, named Spatial Direct Mode (SDM) and Temporal Direct Mode (TDM).

In SDM, the motion vectors are derived from the motion vectors of spatially adjacent blocks. The spatial predictor is the median of the three previous coded motion vectors from the left (block 'A'), the above (block 'B'), and the above-right blocks (block 'C', or the above-left block (block 'D') when the motion vector of the above-right block unavailable), as shown in Figure. 1. SDM can efficiently exploit the spatial correlation and sometimes the temporal redundancy by considering whether a block is stationary or not according to the motion vector of the co-located block. This method can achieve good performance for smooth and/or scene-changed sequences.

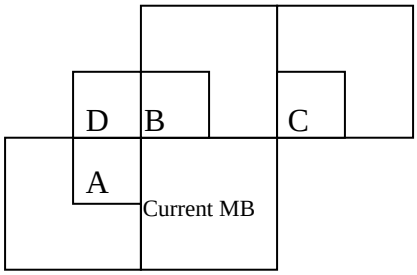

Figure 1. Median prediction of motion vectors for SDM

The other type of direct mode, TDM, derives its forward and backward motion vectors from that of its co-located block in the temporally subsequent reference pictures. As illustrated in Figure. 2, the forward/backward motion vector MV_F /MV_B of direct mode can be calculated as follows:

$$MV_F = \frac{TR_b}{TR_d} \times MV_C \quad (1)$$

$$MV_B = \frac{TR_b - TR_d}{TR_d} \times MV_C \quad (2)$$

where TR_b denotes the temporal distance between the current B picture and the forward reference picture. TR_d denotes the temporal distance between the forward reference picture and the backward reference picture, and MV_C represents the motion vector of the co-located block in the backward reference picture. When the co-located block in the backward reference picture is coded with intra mode, MV_C will be set as zero.

One problem with both SDM and TDM is that the derived motion vectors are not accurate enough in certain cases. For TDM, the current block is not always located in the motion trajectory of its co-located block, particularly when the current block and its co-located block in the backward

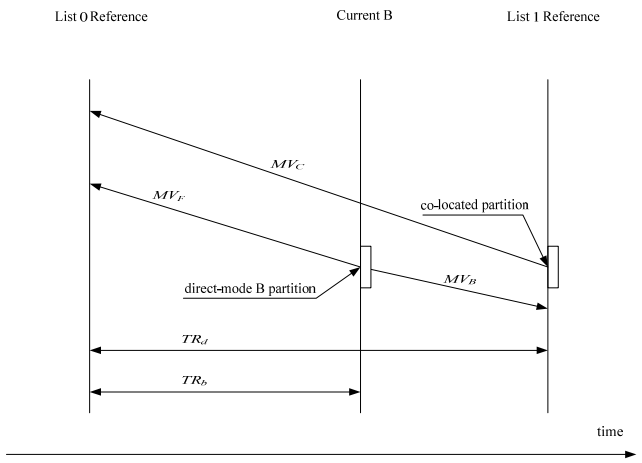

Figure 2. TDM used in H.264/AVC

reference picture belongs to different objects with different motion directions. Similarly, for SDM, the current MB and its spatially neighboring blocks may not share the same motion when they belong to different objects. In order to improve the motion vector accuracy of direct mode block, Ji *et al* [4] proposed a motion vector tracking scheme in TDM which could get more accurate motion vectors than the conventional TDM. A pixel projection technique was proposed in [5] to generate a virtual reference picture for predicting the direct mode blocks. Another enhanced B skip mode is proposed in [6], which performs exhaustive motion search in both encoder side and decoder side to improve the accuracy of the derived motion vectors. All of these techniques haven't combined spatial and temporal information together. Therefore, there is still room for further improving the coding efficiency of direct mode.

In this paper, we propose a new direct mode, named Spatial and Temporal Direct Mode (STDM), which could more effectively exploit both spatial and temporal correlation. The rest of this paper is organized as follows. In Section 2, we will focus on presenting the proposed STDM. Simulation results are presented in Section 3. Finally, Section 4 concludes this paper.

II. SPATIAL AND TEMPORAL DIRECT MODE

Figure. 3 illustrates the coding flow of STDM. Firstly, we will specify several spatial-temporal neighboring motion vectors as candidates for the current direct MB or block, and then its motion vector is derived among them. The selection criterion is to minimize a spatial-temporal cost function. This framework of Decoder-side Motion Vector Derivation (DMVD) was firstly proposed in [7], in which L-shape template matching is used to generate the predicted blocks in P pictures. The detailed coding process of STDM can be described as follows:

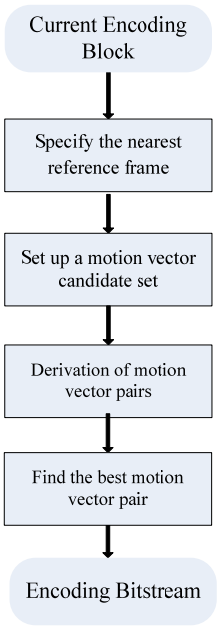

Figure 3. Flowing chart of STDM

Step 1: Specify forward and backward reference frames

For the current MB, we will specify the nearest reference frame in List0 and List1 as the current MB's forward and backward reference frame.

Step 2: Set up a motion vector candidate set

A motion vector candidate set which consists of spatial-temporal neighboring motion vectors is set up. The candidate set only includes forward motion vectors.

As shown in Figure. 1, the spatial candidates are the motion vectors of four adjacent blocks, on the left, top, top-right, and top-left blocks to the current MB. And the temporal candidate is scaled from the co-located motion vector, in the same way as proposed in [8]. If any one of the candidate blocks is intra coded, then its corresponding motion vector will be set to be zero instead.

$$Cadidate\ Set = \{\overrightarrow{mv_A}, \overrightarrow{mv_B}, \overrightarrow{mv_C}, \overrightarrow{mv_D}, Scal(\overrightarrow{mv_{col}})\} \quad (3)$$

Step 3: Derivation of motion vector pairs

Based on the selected candidate motion vectors, we will construct the corresponding forward and backward motion vector pairs. Assuming that the objects follow simple translational motion, the backward motion vector then can be obtained by the following formula:

$$MV_B = MV_F \times \frac{TR_b - TR_d}{TR_b} \quad (4)$$

Step 4: Find the best motion vector pair

After the candidate set is set up, we will check each of the pairs in turn. If one of them could minimize the following spatial-temporal cost function, then it will be selected as the final motion vector of the current direct MB.

$$\begin{aligned} Cost = & SAD(MB_f, MB_b) \\ & + \lambda \times SAD(Template_f, Template_c) \\ & + \lambda \times SAD(Template_b, Template_c) \end{aligned} \quad (5)$$

This cost function consists of 3 parts: sum of absolute differences (SAD) between the forward and backward reference MB, SAD between the forward and current template region, and SAD between the backward and current template region. The template region is an L-shape region with a width of 4 besides current MB or its reference MB (illustrated in Figure. 4). The positions of the referenced MB and template region are offset by the displacement given by the candidate motion vector pair. λ is simply set as 1 in this work.

From formula (5), it can be seen that the first term is a temporal regularization term, which represents that objects are undergoing simple translational motion in the small time period. The second and third terms make use of coded information in current frame considering that the current block should share similar motion with spatially adjacent regions.

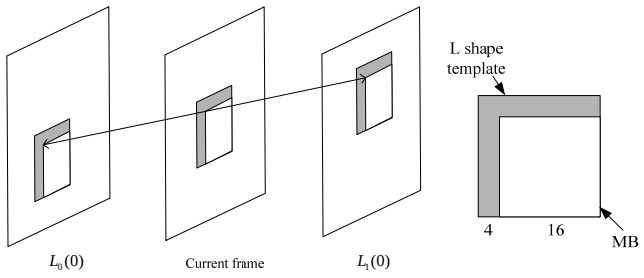

Figure 4. Illustration for the Cost Function

In the following experiments, the above process is applied to B SKIP/DIRECT MB and B_DIRECT_8x8 block. For B_DIRECT_8x8 block, we firstly derive the motion vectors of the MB which the direct block belongs to, then these motion vectors are used for the B_DIRECT_8x8 block directly.

After all these steps finished, the modified reference block is subtracted by the reference block to get the residual and the residual is encoded into the encoding bit stream. B SKIP blocks do not need to code residual

III. Simulation Results

To evaluate the general performance of our proposed Spatial and Temporal Direct Mode compared with conventional direct mode, we integrated this technique into KTA software Version JM11.0 KTA2.4 [9]. Conventional SDM and TDM are all replaced by the proposed STDM. We write a USE_STDM_FLAG into the bitstream. If USE_STDM_FLAG is equal to 1, then the motion vectors of B_SKIP, B_DIRECT_16x16 and B_DIRECT_8x8 are derived via the proposed process. Since STDM totally replace the conventional SDM and TDM, no extra information is needed to be coded in the macroblock level.

Test sequences include *Forman*, *Paris* and *Mobile* in CIF format; *Blowingbubbles*, *BasketballPass* and *BQsquare* in WQVGA format; *Partyscene*, *Racehorse* and *BQMall* in WVGA format; *Cactus*, *Kimono* and *Basketballdrive* in 1080P format. The test conditions are showed in Table 1. Performance under both regular B configuration and hierarchical B configuration are evaluated. To evaluate the average PSNR vs. bit-rate, we employ the method described in [10], which is widely used during H.264/AVC development. The detailed coding gains over SDM and TDM under hierarchical B and regular B configuration are tabulated in Table 2 and Table 3, respectively. From Table 2 we can see that STDM could lead up to 11.83% average bit rate reduction or equivalently 0.483db over the current TDM of H.264 in hierarchical B case.

TABLE I. Test conditions

Number of Reference frames	4
ME method	EPZS FME
Search Range	32
MV resolution	1/4 pel
RD optimization	ON
Entropy coding method	CABAC
GOP structure	IBBPBBP (Regular B)
	IBBBBBBBP (Hierarchical B)
QP	28, 32, 36, 40

Figure. 5 shows the rate-distortion curves of the proposed method in sequence Mobile compared with SDM and TDM under regular B configuration. It can be easily observed that the benefits of the proposed method tend to increase at lower bitrates, which is to be expected considering that motion information at these bitrates tends to take an even larger percentage of the coded information.

Through the regulation of the spatial-temporal cost function, the proposed STDM could find the most accurate motion vectors among the candidate pairs. Since both encoder and decoder share the same derivation process, no extra overhead is introduced. As a result, more blocks are coded with direct mode and better coding performance are achieved.

As for the complexity, for each MB, at most 5 pairs of motion vectors need to be checked, and at most 11 template information (one for the current macroblock and two for each pair of motion vector) and 10 reference macroblock information are required. Complexity and memory requirement are increased a little but still tolerable in most applications. Our future research work will concentrate on reducing the complexity of STDM.

IV. CONCLUSION

This paper proposes an advanced Spatial and Temporal Direct Mode for B picture coding, which could improve the accuracy of derived motion vectors by effectively exploiting both spatial and temporal correlation. Simulation results demonstrate that the proposed method could significantly improve the coding efficiency especially at low bit-rate case compared to the direct mode defined in H.264/AVC, while keeping relatively little increase in complexity.

V. ACKNOWLEDGEMENT

This work was supported in part by National Science Foundation (60833013, 60803068) and National Basic Research Program of China (973 Program, 2009CB320903).

REFERENCES

[1] ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4) AVC: "Advanced Video Coding for Generic Audiovisual Services," March 2005.

[2] M. Flierl and B. Girod. "Generalized B Pictures and the Draft H.264/AVC Video Compression Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 587-597, July 2003.

[3] A. Tourapis, F. Wu and S. Li, "Direct mode coding for bi-predictive pictures in the JVT standard," IEEE International Symposium on Circuits and Systems (ISCAS), pp. 700-703, May 2003.

[4] X. Ji and Y. Lu, "Enhanced direct mode coding for bipredictive pictures," IEEE International Symposium on Circuits and Systems (ISCAS), pp785-788, May 2004.

[5] D. Liu, D. Zhao, J. Sun and W. Gao, "Direct Mode Coding for B Pictures using Virtual Reference Picture," IEEE International Conference on Multimedia & Expo (ICME), pp. 1363-1366, July 2007.T. Murakami and S. Saito, "Advanced B Skip Mode with Decoder-side Motion Estimation," ITU-T Q.6/SG16 VCEG, VCEG-AK12, April, 2009.

[6] T. Murakami and S. Saito, "Advanced B Skip Mode with Decoder-side Motion Estimation," ITU-T Q.6/SG16 VCEG, VCEG-AK12, April, 2009.

[7] S. Kamp, M. Evertz and M. Wien, "Decoder Side Motion Vector Derivation," MPEG Doc. M14917, October 2007.

[8] Guillaume Laroche, Joel Jung and Beatrice Pesquet-Popescu "RD Optimized Coding for Motion Vector Predictor Selection", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 12, pp. 1681-1691, December 2008.

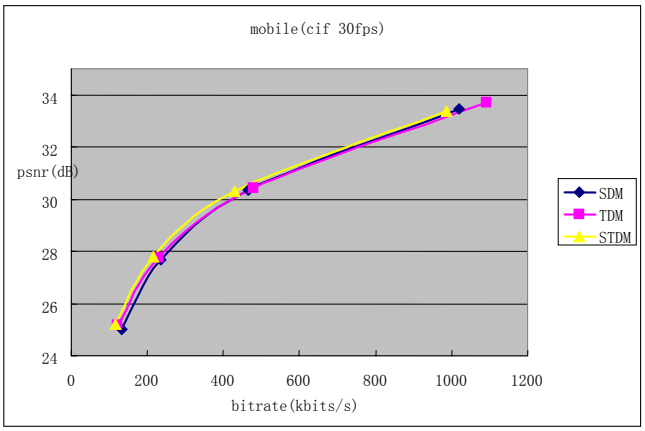

Figure 5. Rate-distortion curves for all pictures in Mobile sequence under regular B configuration

[9] KTA software Ver. JM11.0 KTA2.4.Available: http:// iphome.hhi.de/suehring/tml/download/KTA/jm11.0kta2.4.zip.

[10] G. Bjontegaard, "Calculation of average PSNR differences between RD-Curves," Doc. VCEG-M33, March 2001.

TABLE II. PERFORMANCE COMPARISONS UNDER HIERARCHICAL B CONFIGURATION

SEQUENCE	RESOLUTION	COMPARED WITH SDM		COMPARED WITH TDM	
		-ΔRate (%)	ΔPSNR (dB)	-ΔRate (%)	ΔPSNR (dB)
Foreman	*CIF*	5.51	0.249	14.47	0.657
Paris	*CIF*	5.06	0.251	10.75	0.551
Mobile	*CIF*	12.36	0.503	14.04	0.555
BasketballPass	*WQVGA*	7.82	0.343	12.79	0.563
BlowingBubbles	*WQVGA*	7.47	0.280	11.57	0.436
BQsquare	*WQVGA*	1.32	0.042	11.49	0.440
PartyScene	*WVGA*	3.41	0.128	7.93	0.301
Racehorse	*WVGA*	3.01	0.122	4.85	0.192
BQmall	*WVGA*	6.18	0.277	11.51	0.527
Cactus	*1080p*	9.91	0.360	11.31	0.399
Kimono	*1080p*	12.06	0.509	15.11	0.598
Basketballdrive	*1080p*	9.81	0.363	16.08	0.582
Average		**6.99**	**0.286**	**11.83**	**0.483**

TABLE III. PERFORMANCE COMPARISONS UNDER REGULAR B CONFIGURATION

SEQUENCE	RESOLUTION	COMPARED WITH SDM		COMPARED WITH TDM	
		-ΔRate (%)	ΔPSNR (dB)	-ΔRate (%)	ΔPSNR (dB)
Foreman	*CIF*	5.64	0.241	4.68	0.201
Paris	*CIF*	4.61	0.229	1.70	0.083
Mobile	*CIF*	8.58	0.350	6.34	0.249
BasketballPass	*WQVGA*	6.87	0.285	2.44	0.098
BlowingBubbles	*WQVGA*	6.72	0.264	3.57	0.134
BQsquare	*WQVGA*	3.68	0.149	7.88	0.323
PartyScene	*WVGA*	3.77	0.151	5.46	0.220
Racehorse	*WVGA*	2.03	0.076	2.22	0.086
BQmall	*WVGA*	5.05	0.220	4.27	0.186
Cactus	*1080p*	8.15	0.285	3.73	0.124
Kimono	*1080p*	10.08	0.399	6.27	0.227
Basketballdrive	*1080p*	10.97	0.376	10.58	0.354
Average		**6.35**	**0.252**	**4.93**	**0.190**