Fast macroblock mode selection algorithm for multiview depth video coding

Zongju Peng (彭宗举)[1,2,3], Mei Yu (郁 梅)[2], Gangyi Jiang (蒋刚毅)[1,2*], Feng Shao (邵 枫)[2], Yun Zhang (张 云)[1,3], and You Yang (杨 铀)[1,3]

[1] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

[2] *Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China*

[3] *Graduate University of Chinese Academy of Sciences, Beijing 100049, China*

**E-mail: jianggangyi@126.com*

Received April 17, 2009

Huge computational complexity of multiview video plus depth (MVD) coding is an obstacle for putting MVD into applications. A fast macroblock mode selection algorithm is proposed to reduce the computational complexity of multiview depth video coding. The proposed algorithm, implementing on a joint coding scheme, combines an effective prediction mechanism and an object boundary discriminating method. The prediction mechanism which is designed based on the macroblock mode similarities reduces the number of macroblock mode candidates in depth video coding. The object boundary discriminating method extracts the regions, which are with discontinuous depth values and important for virtual view rendering, by using macroblock deviation factor. Experimental results show that the proposed algorithm can significantly promote the coding speed of depth video by 2.00–3.40 times, while maintaining high rate distortion (RD) performance in comparison with the full search algorithm.

OCIS codes: 110.0110, 100.6890, 330.1690.

doi: 10.3788/COL20100802.0151.

With the fast development in the areas of integrated optics with sensors and network infrastructures, three-dimensional (3D) video systems will soon be used in a great number of applications. Integral imaging technology, one of the most promising methods for 3D scenes representation, attracts a lot of research interests[1,2]. Multiview video plus depth (MVD)[3] is an alternative to integral imaging for representing 3D scenes. MVD signals include multiple texture videos and associated depth videos of the same scene. MVD signals are first captured at different sparse viewpoints and compressed, then transmitted to client. The MVD bit streams are decoded and utilized to synthesize the virtual views with depth-image-based rendering (DIBR) technique.

To efficiently compress MVD signals, Park *et al.* proposed the view-temporal prediction structures that can be adjusted to various characteristics of general multiview video[4]. In Ref. [5], an effective algorithm was proposed to eliminate the color inconsistency between multiview videos for better coding and rendering performances. Yang *et al.* proposed an image region partition and regional disparity estimation algorithm for multiview video coding[6]. For standardizing encoding of MVD, the joint multiview video model (JMVM) was developed, based on the video coding standard H.264/AVC. In JMVM, an exquisite view-temporal prediction structure based on hierarchical B pictures (HBP) is used to exploit not only the temporal correlations within a single view, but also the inter-view correlations among different views[7].

The JMVM has nine macroblock modes, including SKIP, Inter 16×16, Inter 16 × 8, Inter 8 × 16, Inter 8 × 8, Inter 8×8 Frext, Intra 16×16, Intra 8×8, and Intra 4×4. These modes are probed by the full search algorithm to determine the optimal macroblock mode for the best rate distortion (RD) performance. The mode with the minimal RD cost is then selected as the best mode for Inter frame coding. Unfortunately, the full search algorithm is time consuming. The computational complexity of MVD coding can be approximately expressed as O $(\eta\times\alpha\times\beta\times\theta)$, where η, α, β, and θ denote the number of videos in each view, views, average reference frames, and macroblock modes, respectively. It is an obstacle for putting MVD into applications. To reduce the complexity of MVD coding, the fast macroblock mode selection algorithms were proposed to accelerate the coding speed of multiview texture video[8,9]. However, the fast algorithms for multiview depth video so far are marginal.

This letter focuses on reducing the computational complexity of multiview depth video coding. Firstly, a joint coding scheme is proposed based on macroblock mode similarity between the texture videos and the associated depth videos. Then, a fast depth video coding algorithm is presented by combining an effective prediction mechanism and an object boundary discriminating method. Finally, the fast algorithm is implemented and evaluated.

Figure 1 shows an MVD-based 3D video system. In texture video and associated depth video, boundaries of objects in the scene coincide and directions of object movements are also very similar. Therefore, the macroblock mode distributions of the texture image and its associated depth image will be similar. Figures 2(a) and (b) show the mode distributions of the texture image and the associated depth image of a frame in Ballet test sequence. The blocks with red, green, and blue borders denote the macroblocks encoded with SKIP, Inter, and Intra modes. It can be found that the macroblock modes are similar between these two images. The similarity can be utilized to speed up the coding process. Based on the analyses above, a joint MVD coding scheme is proposed and

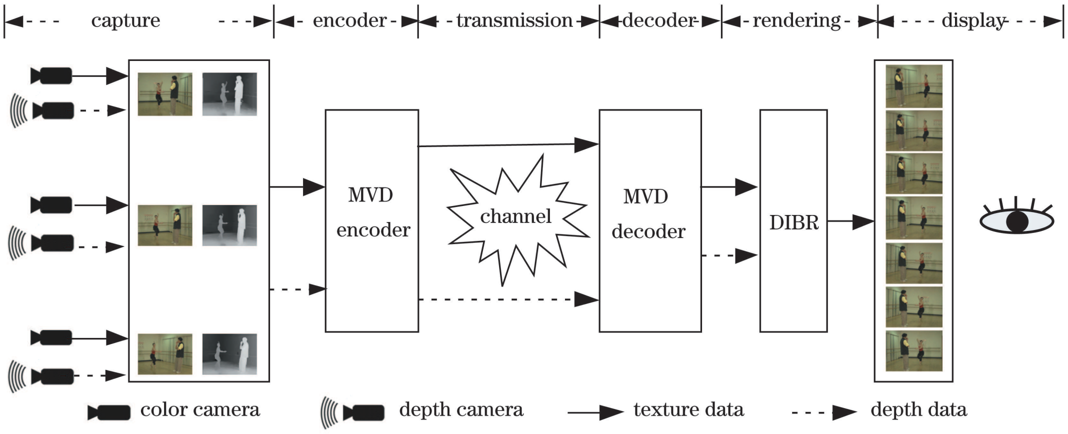

Fig. 1. A MVD-based 3D video system.

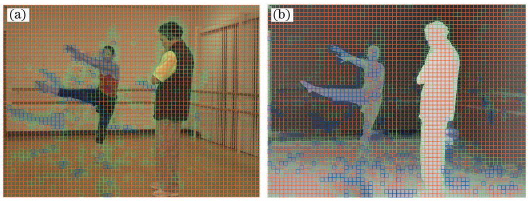

Fig. 2. Optimal modes of (a) texture image and (b) depth image at S_0T_8 of Ballet.

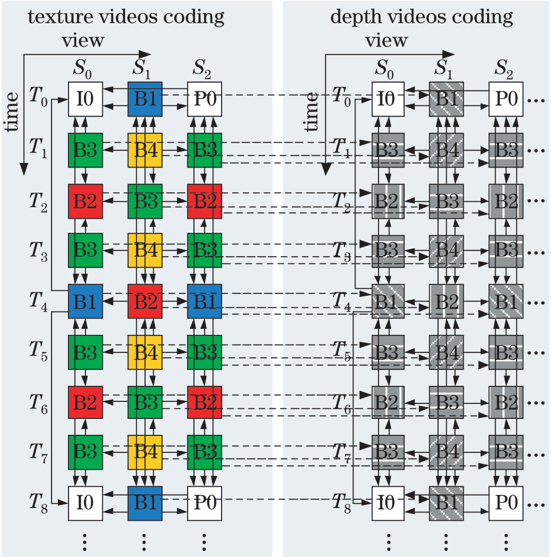

Fig. 3. Joint MVD coding scheme.

illustrated in Fig. 3, where S_m denotes the individual view ($m = 0, 1, 2$) and T_n is the consecutive time instant ($n = 0, 1, 2,\cdots,8$). For example, S_0T_6 represents the frame locating at the 6th time instant in the view 0. I, B, and P are Intra encoded, bi-directional Inter prediction, and Inter prediction frames. The left part and the right part of Fig. 3 are coding prediction structures of the texture video and the depth video, respectively. The dashed lines, linking a texture frame and the corresponding depth frame at the same viewpoint and the same time instant, denote the mode prediction relationships. Since frame B occupy most of encoding time, mode prediction of frames I and P are not taken into consideration. In the proposed scheme, the coding process of MVD signals includes two steps. The texture videos are encoded firstly, and their macroblock modes are recorded for mode predictions of the associated depth videos. Then, the depth videos are encoded based on an effective prediction mechanism and an object boundary discriminating method.

Mode prediction accuracy should be investigated before designing a feasible mode prediction mechanism. A mode prediction mechanism **P** can be set as

$$\begin{aligned}&\mathbf{P}=\bigcup\nolimits_{i=1}^{C}\{<\mathbf{A}_i,\mathbf{B}_i>\},\\&\mathbf{A}_i\bigcap\mathbf{A}_j=\varnothing(0<i<C,0<j<C,i\neq j),\\&\bigcup\nolimits_{i=1}^{C}\mathbf{A}_i=\mathbf{\Omega},\end{aligned}\qquad(1)$$

where C is the number of prediction rules, $\mathbf{A}_i$ and $\mathbf{B}_i$ are the macroblock mode sets of texture video and depth video in the ith prediction condition rule, respectively. $\mathbf{\Omega}$ is the set of all macroblock modes searched with the full search algorithm. $<\mathbf{A}_i,\mathbf{B}_i>$ denotes that if the optimal mode of a macroblock in texture video belongs to $\mathbf{A}_i$, all macroblock modes in $\mathbf{B}_i$ will be searched to obtain the encoding macroblock mode of corresponding macroblock in depth video. Regarding to the mode prediction mechanism **P**, the prediction accuracy can be defined as

$$k(\mathbf{P})=\sum\nolimits_{i=1}^{C}\sum\nolimits_{a\in\mathbf{A}_i}\sum\nolimits_{b\in\mathbf{B}_i}s(a,b),\qquad(2)$$

where $s(a,\ b)$ denotes the probability of macroblocks in depth video or texture video, and these macroblocks in texture video are with the optimal mode a while their corresponding macroblocks in depth video are with the optimal mode b.

The simplest prediction mechanism, a special case of **P** which is called direct mapping (DM), is that the macroblock mode of the corresponding macroblock in the texture video is used as the final macroblock mode of current macroblock in the depth video. The DM prediction mechanism $\mathbf{P}_{\mathrm{DM}}$ can be represented as

$$\begin{aligned}&\mathbf{P}_{\mathrm{DM}}=\bigcup\nolimits_{i=1}^{C}\{<\{\mathrm{m}_i\},\{\mathrm{m}_i\}>\},\\&\bigcup\nolimits_{i=1}^{C}\{\mathrm{m}_i\}=\ \mathbf{\Omega},\end{aligned}\qquad(3)$$

where m_i denotes macroblock mode. The coding speed will be greatly improved with $\mathbf{P}_{\mathrm{DM}}$. However, DM prediction method may seriously deteriorate the RD performance of depth video coding because of the inaccurate mode prediction. Figure 4 shows the prediction accuracy obtained by statistically analyzing the optimal macroblock modes of Ballet test sequence. $k(\mathbf{P}_{\mathrm{DM}})$ is 62.19% on average. In order to improve the RD performance of $\mathbf{P}_{\mathrm{DM}}$, an improved prediction mechanism $\mathbf{P}_{\mathrm{PRO}}$ is proposed and represented as

$$\begin{aligned} &\mathbf{P}_{\mathrm{PRO}} = \{<\mathbf{E},\mathbf{F}>\}\bigcup\{<\overline{\mathbf{E}},\boldsymbol{\Omega}>\},\\ &\mathbf{E} = \{\mathrm{SKIP},\mathrm{Intra}\},\\ &\mathbf{F} = \{\mathrm{SKIP},\mathrm{Inter16}\times 16,\mathrm{Intra}\}, \end{aligned} \qquad (4)$$

where $\overline{\mathbf{E}}$ is the complementary set of $\mathbf{E}$. $\mathbf{P}_{\mathrm{PRO}}$ consists of two prediction rules. If the macroblock mode of corresponding texture macroblock belongs to $\mathbf{E}$, the first prediction rule $<\mathbf{E}$, $\mathbf{F}>$ will be adopted, which means that SKIP, Inter16×16, and Intra will be searched to determine the coding mode of the depth macroblock. Otherwise, the second prediction rule $<\overline{\mathbf{E}},\boldsymbol{\Omega}>$ is adopted and the full search method is utilized to obtain the optimal mode of the depth macroblock. As shown in Fig. 4, the prediction accuracy of $\mathbf{P}_{\mathrm{PRO}}$ is higher than that of $\mathbf{P}_{\mathrm{DM}}$, and rises up to 94.47%. However, in $\mathbf{P}_{\mathrm{PRO}}$ the remaining 5.53% macroblocks which cannot be predicted correctly will still decrease the RD performance of the depth video. The experimental results show that most of such macroblocks locate at object boundaries, and these macroblocks are important to the quality of rendered virtual view image. Therefore, the RD performance in regions with discontinuous depth value should be kept. To extract the regions with discontinuous depth, a macroblock deviation factor (MDF) $f(x,y)$ is defined to represent the variation degree of depth values in a macroblock and calculated by

$$f(x,y) = \left(\frac{N}{16}\right)^2 \sum_{i=1}^{16/N}\sum_{j=1}^{16/N} |\, r(i\times N, j\times N) - g(x,y)\,|, \quad (5)$$

where $(x,\ y)$ is the coordinate of the current macroblock in the macroblock-size of units, $r(i,\ j)$ is the depth value at relative coordinate $(i,\ j)$ in the current macroblock, N is a scale of down sampling, and can be set as 1, 2, 4, 8, or 16. $g(x,\ y)$ is the mean of the depth values of current macroblock, and calculated by

$$g(x,y) = \left(\frac{N}{16}\right)^2 \sum_{i=1}^{16/N}\sum_{j=1}^{16/N} r(i\times N, j\times N). \qquad (6)$$

Figure 5(a) shows the MDF with N=1 of every macroblock in frame S_0T_6 of Ballet depth video. It is clear that the MDF values of the macroblocks locating at the boundaries are higher because the depth values of such macroblocks change drastically. By contrast, most of the macroblocks in depth image are smooth, and their MDF values approach to 0. A threshold T can be set to determine whether the macroblock at (x,y) locates at boundaries or not. In Fig. 5(b), the macroblocks with green borders are distinguished as boundary regions with $f(x,\ y)>T$ (where T=5).

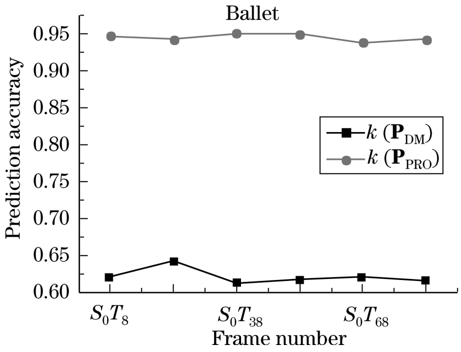

Fig. 4. Prediction accuracy comparison between $k(\mathbf{P}_{\mathrm{DM}})$ and $k(\mathbf{P}_{\mathrm{PRO}})$.

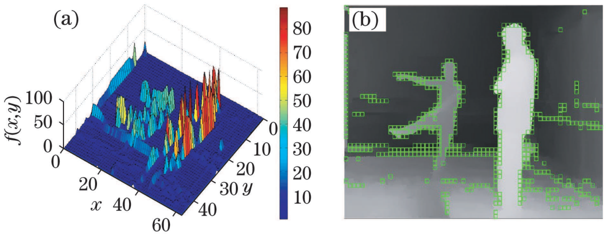

Fig. 5. Features of macroblock in frame S_0T_6 of Ballet depth image for (a) $f(x,y)$ of every macroblock and (b) macroblocks with $f(x,y) > T$ (where T=5).

Except for sum of absolute difference (SAD) in Eq. 5, sum of squared difference (SSD) can also be used to form $f(x,y)$. Based on the exploratory experiments, threshold condition $f(x,y)>T$ built by SSD and SAD can be used to extract nearly the same boundary regions if T is reasonably set. However, computational complexity should also be taken into consideration. Computational burden of SSD is more than that of SAD. So SAD is adopted in this letter.

By combining $\mathbf{P}_{\mathrm{PRO}}$ and MDF, a fast macroblock mode selection algorithm for multiview depth video is put forward. In the proposed algorithm, $f(x,y)$ is calculated firstly when a macroblock at (x,y) is encoded. If $f(x,y)\le T$ and the macroblock mode of the corresponding macroblock in the associated texture video is SKIP or Intra, the macroblock modes, SKIP, Intra, and Inter 16×16, will be searched to find the optimal mode of current macroblock in the depth video. Otherwise, the full search method is adopted to obtain the best mode. The flowchart is illustrated by Fig. 6.

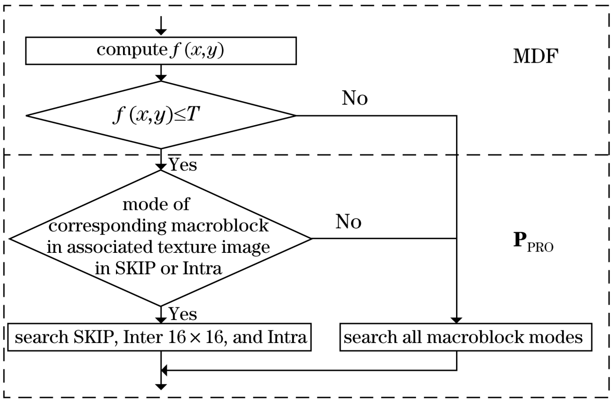
compute f (x,y)
MDF
f $(x,y)\le T$
No
Yes
mode of corresponding macroblock in associated texture image in SKIP or Intra
No
$\mathbf{P}_{\mathrm{PRO}}$
Yes
search SKIP, Inter 16 × 16, and Intra
search all macroblock modes

Fig. 6. Flowchart of proposed macroblock mode selection algorithm in depth video coding.

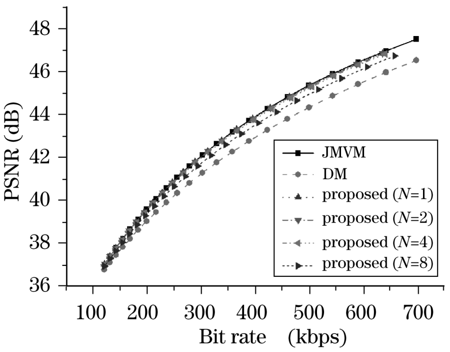

Fig. 7. RD performances comparisons of depth video of Ballet test sequence. PSNR: peak singal-to-noise ratio.

Table 1. Speedup Performances of DM Method and the Proposed Algorithm

QP		22	27	32	37
DM	Ballet	3.43	3.85	4.14	4.48
	Breakdancers	2.25	2.63	2.91	3.39
Proposed (N=1)	Ballet	2.73	2.91	3.03	3.17
	Breakdancers	2.00	2.27	2.43	2.73
Proposed (N=2)	Ballet	2.73	2.92	3.04	3.18
	Breakdancers	2.00	2.27	2.44	2.73
Proposed (N=4)	Ballet	2.75	2.94	3.07	3.21
	Breakdancers	2.00	2.27	2.44	2.73
Proposed (N=8)	Ballet	2.84	3.06	3.22	3.40
	Breakdancers	2.03	2.31	2.50	2.81

To evaluate the performance of the proposed algorithm, experimental comparisons have been made among the full search algorithm in JMVM, DM, and the proposed algorithm. The experiments were performed complying with the common test conditions[10]. Breakdancers and Ballet sequences, provided by Microsoft Research, were adopted as test sequences. Both of them were with 8 views, 1024×768 resolution, and 15-fps frame rate. Each of views includes a texture video and its associated depth video. All tests in the experiments were run on the Intel Xeon 3.2 GHz with 12-GB RAM and the operating system was Microsoft Windows Server 2003. T was set as 5 empirically.

Table 1 lists the speedup performances of DM method and the proposed algorithm. Compared with the full search algorithm, DM method and the proposed algorithm can promote the coding speed by 2.25–4.48 and 2.00–3.48 times, respectively. Figure 7 shows the RD performances of full search algorithm, DM method, and the proposed algorithm with respect to Ballet depth video test sequence. The RD performances of the proposed algorithm with N=1, 2, and 4 are nearly the same as that of the full search algorithm, and much better than that of the DM method. Breakdancers depth test sequence is with the similar comparison results in terms of RD performance. The proposed fast algorithm maintains the high RD performance due to the high macroblock mode prediction accuracy. However, the RD performance of DM method is inferior to that of the proposed algorithm because the lower prediction accuracy of DM method leads to more macroblock mode mismatch deteriorates the RD performance.

The speedup and RD performances of the proposed algorithm are different as N varies. Table 1 also shows that the speedup performances of the proposed algorithm are slightly getting better as N increases. The reason is that the time cost of computing MDF is much less than that of motion and disparity estimation. Figure 7 also illustrates that the RD performance of the proposed algorithm with N=8 is inferior to that of N=1, 2, or 4, especially at high bit rates. Thus, N=4 are reasonable for the test sequences in the proposed algorithm due to previous detailed overall analyses on the speedup RD performances.

In conclusion, a fast macroblock mode selection algorithm is proposed to reduce computational complexity of multiview depth video coding. Firstly, a joint coding scheme is presented, in which the macroblock modes of texture video are used to predict the macroblock modes of its associated depth video. Then, a feasible prediction mechanism is designed after careful consideration on the tradeoff between coding time and RD performance. Additionally, a MDF is defined to discriminate the macroblocks in discontinuous depth regions which are important for virtual view rendering. The effective prediction mechanism and the MDF are adopted in the proposed algorithm. Finally, the fast algorithm is tested. Experimental results show that the proposed algorithm promotes the coding speed by 2.00–3.40 times in comparison with the testing benchmark JMVM while the proposed algorithm maintains the high RD performance.

This work was supported by the National Natural Science Foundation of China (Nos. 60872094 and 60832003), the Project of Chinese Ministry of Education (No. 200816460003), and the National “863” Project of China (No. 2009AA01Z327).

References

1. N. Sgouros, I. Kontaxakis, and M. Sangriotis, Appl. Opt. **47,** D28 (2008).
2. J.-S. Jang, S. Yeom, and B. Javidi, Opt. Eng. **44,** 127001 (2005).
3. P. Merkle, A. Smolic, K. Müller, and T. Wieganol, in *Proceedings of International Conference on Image Processing* I-201 (2007).
4. P.-K. Park and Y.-S. Ho, Opt. Eng. **47,** 047401 (2008).
5. R. Fu, F. Shao, G. Jiang, and M. Yu, Chin. Opt. Lett. **6,** 654 (2008).
6. H. Yang, Y. Chang, J. Huo, L. Xiong, and S. Lin, Acta Opt. Sin. (in Chinese) **28,** 1073 (2008).
7. P. Merkle, A. Smolić, and K. Müller, IEEE Trans. Circ. Syst. Video Tech. **17,** 1461 (2007).
8. Z. Peng, G. Jiang, M. Yu, and Q. Dai, Int. J. Image Video Process. *2008* 393727 (2008).
9. Z. Peng, G. Jiang, and M. Yu, Acta Opt. Sin. (in Chinese) **29,** 1216 (2009).
10. Y. Su, A. Vetro, and A. Smolic, ISO/IEC JTC1/SC29WG11 and ITU-T SG16 Q6 (2006).