Chinese Journal of Electronics
Vol.18, No.2, Apr. 2009

New Approach to Multi-Modal Multi-View Video Coding*

ZHANG Yun[1,4], YU Mei[2,3] and JIANG Gangyi[1,2]

(1.*Institute of Computing Technology, Chinese Academic of Sciences, Beijing 100080, China*)

(2.*Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China*)

(3.*National Key Laboratory of Software New Technology, Nanjing University, Nanjing 210093, China*)

(4.*Graduate School of Chinese Academic of Sciences, Beijing 100080, China*)

Abstract — The correlation characteristics of Multi-view video (MVV) are influenced by the content of the video, illumination change, speed of moving objects and cameras, camera distance, frame rate, etc. In this paper, a framework of Multi-modal multi-view video coding (MMVC) is proposed on the basis of correlation analysis to achieve optimal performances among high compression efficiency, low complexity, low memory cost, view scalability and fast random access. Different prediction modes are designed to fit MVV with different correlations and meet different requirements of the Multi-view video coding (MVC). An optimal prediction mode is adaptively selected from the candidate modes according to the correlation characteristics of MVV. Experimental results have proved that MMVC not only has best random accessibility, but also has outstanding performance in compression efficiency, low memory requirement, low complexity and view scalability. MMVC is regarded as the most efficient and balanced MVC scheme among the compared schemes.

Key words — Multi-view video coding, Correlation analysis, Multi-modal multi-view video encoder, Random access, View scalability.

I. Introduction

Multi-view video (MVV) is a collection of multiple viewpoint videos capturing the same scene at different camera locations. The captured scenes can be displayed interactively, which lets the user select the view from multiple angles as if it were 3D and enjoy the feeling of being in the scene[1]. Multi-view video coding (MVC) serves emerging applications, such as free-viewpoint video system, where multiple views of the same scene are coded with possibly high temporal and inter-view correlation between them[2,3].

MPEG has surveyed some of MVC schemes, such as 'Sequential view prediction' (SVP), 'checkerboard decomposition' and so on[4]. The SVP can achieve relatively high compression efficiency by using temporal and sequential inter-view prediction. Oka *et al.* proposed MVC scheme using multi-directional pictures[5], where optimal mode was selected through rate-distortion optimization and multi-reference technology. Mueller *et al.* proposed a MVC scheme using hierarchical B pictures, which shows its superior compression efficiency and temporal scalability[6].

In addition to high compression efficiency, MVC should support fast Random access (RA) in temporal and view dimensions, low coding delay, view scalability as well as low complexity[7]. Recently, more and more importance has been attached to these MVC schemes' functionalities[8−10]. View scalability is defined as the functionality that the same bitstream to be displayed on a multitude of different terminals and over networks with various performance attributes. Moreover, RA is an ability of accessing a frame at a given time with minimal decoded frames and it directly affects the interactive system capabilities that let the user freely change viewing position and direction while downloading and streaming a video content.

Since many existing MVC schemes, such as SVP, MVC scheme using multi-directional pictures, are poor in RA and view scalability, NTT Corporation and Nagoya University proposed Group-of-GOP (GoGOP) scheme to improve random accessibility by adopting multiple intra frames in a 2 Dimensional group-of-picture (2DGOP) at the cost of compression efficiency[8,9]. Liu et al introduced three methods, SP/SI frame in view dimension, multiple representation coding and interleaved view coding, to improve RA[10]. Unfortunately, some of MVC's requirements are conflicting to one another and we cannot expect a single prediction structure to be universally effective for any scene at any time.

In this paper, Multi-modal multi-view video coding (MMVC) is proposed to achieve optimal performances among high compression efficiency, low complexity and high ability of RA. Section II shows some correlation analyses of MVV and describes the problems of traditional MVC schemes. Section III presents the framework of MMVC. Section IV gives experimental results of the proposed framework compared with five typical MVC schemes in compression efficiency, RA, encoding

*Manuscript Received Nov. 2007; Accepted July 2008. This work is supported by the National Natural Science Foundation of China (No.60672073, No.60872094, No.60832003), the Program for New Century Excellent Talents in University (No.NCET-06-0537).

complexity, memory requirement and view scalability. Finally, some conclusions are given.

II. Correlation Characteristics of Multi-View Video Sequences

MVV is generated by many cameras which simultaneously capture the same scene from different directions. Therefore, MVV not only contains temporal redundancy but also has large degree of inter-view redundancy. Correlation characteristic of MVV sequences have been analyzed based on block matching method, as shown in Fig.1. In the figure, the current coded frame is marked as 'F', 'T' denotes temporally preceding frames of the F-frame, and 'V' represents frames at the same instant of the F-frame in the neighboring views. Blocks in the F-frame are predicted from the V-and T-frames by block matching. The numbers of most matched blocks from the T-frames or V-frames are counted, respectively, so as to analyze correlations of different sequences.

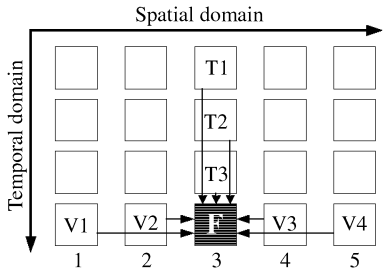


Fig. 1. Correlation analysis of MVV sequence

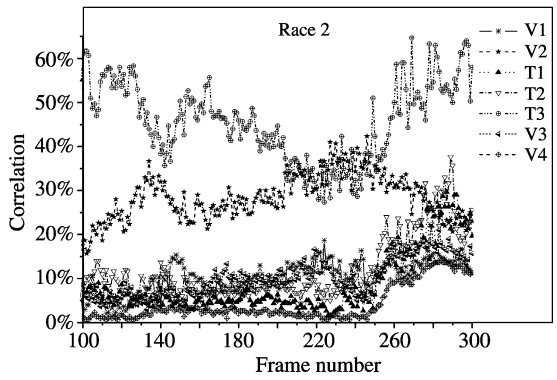


Fig. 2. Correlations of MVV sequence "race 2"

It is noticed that the temporal correlation will be the dominant in the sequences where the objects move slowly or camera distance is large. "Crowd", "race 1", "objects 1" and "Aquarium" are such kind of sequences, and their percentage of temporal correlation is from 86.1% to 91.2%. By contrast, temporal correlation decreases to 19.2% for "Xmas" sequence whose camera distance is very small, that is, the inter-view correlation is the dominant. Besides the above two kinds of MVV, there is another kind of sequences in which the temporal correlation and the inter-view correlation are balanced, so we call it as hybrid correlation.

Fig.2 shows correlations of MVV sequence "race 2". The x-axis is the frame number, while y-axis indicates the percentages of blocks in F-frame referenced from V1, V2, V3, V4, T1, T2, and T3 respectively, as shown in Fig.1. It is seen that correlations of MVV vary along the time axis. For instance, from the 200-th to 250-th frame of "race 2" sequence, inter-view correlation becomes stronger than temporal correlation because cameras move fast with the car, as shown in Fig.2. For "flamenco 1" sequence, although the temporal correlation is the dominant at the most time, there are two periods in which the inter-view correlation is stronger than temporal correlation due to the lighting change. For "objects 1" sequence, there is regular impulse with respect to V2, caused by the regular flicker of lamps. From the above analysis results, it is clear that the correlations of MVV are influenced by the content of the video, illumination change, speed of moving objects and cameras, camera distance, frame rate and so on. The instantaneous change of illumination, high-speed motion will reduce the temporal correlation; while large camera distance will reduce the inter-view correlation.

Because of the non-stationary property of video stream, we cannot expect a single prediction structure to be universally effective at any time for any scene. The conventional approaches with single prediction structure can hardly remove inter-view redundancies efficiently when fast RA and flexible view scalability are expected to be achieved.

III. The Framework of MMVC

1. The framework of MMVC

Fig.3 gives the framework of MMVC, which is able to use different prediction mode to encode MVV according to the correlation characteristic of current MVV. The MMVC encoder consists of four modules. They are module of predication mode selection, MVC module, mode updating trigger, and module of correlation analysis.

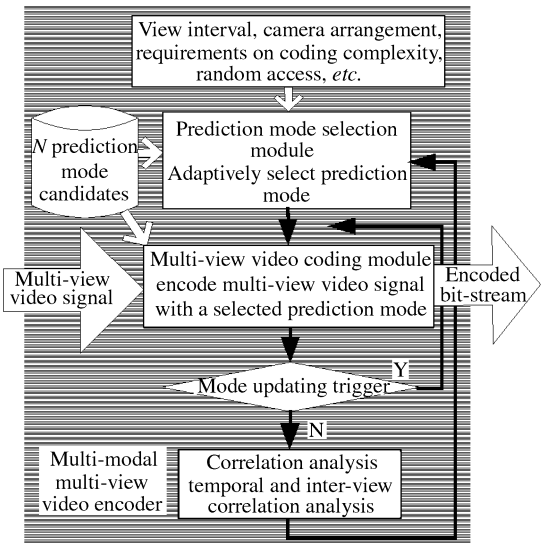


Fig. 3. Flowchart of MMVC

At the beginning, an initial predication mode is selected

from N candidate modes in terms of parameters of camera array such as camera distance, camera arrangement (parallel/convergence setup or other arrangements), or requirements on coding complexity, RA, *etc.* The input MVV is encoded with the selected predication mode in MVC module; meanwhile, the correlation characteristic of current MVV is analyzed. The updating trigger is in charge of mode updating, it determines whether the prediction mode should be changed or not. If the updating is activated, another appropriate prediction mode will be selected from the N candidates according to the results of correlation analysis, otherwise the selected predication mode will be kept working until the mode updating trigger is active again.

2. Predication modes for MVC

In the MMVC framework, three predication modes are designed to encode MVV with different correlation characteristics. With respect to the mentioned three kinds of MVV with different correlation characteristic, we designed three predication modes, that is, Temporal predication mode (TPM), Spatial predication mode (SPM) and Hybrid predication mode (HPM). Fig.4 gives an example of the three types of prediction modes with 5 views and 7 instants in a 2DGOP.

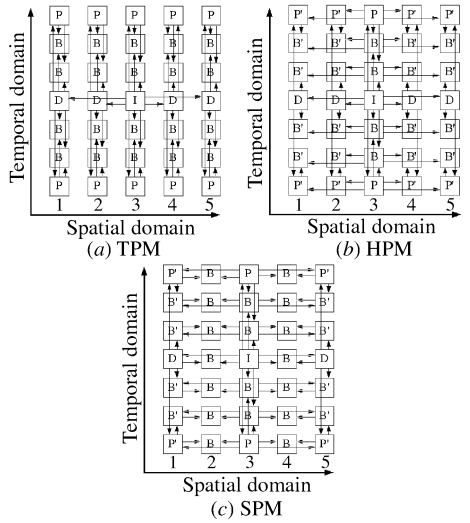


Fig. 4. An example of the three kinds of prediction modes

In Fig.4, I-frame (Intra-predicted frame) is set at the center of the 2DGOP, and the 2DGOP is divided into four regions so as to improve the encoder's ability of RA and parallel processing, because the average path length of reference relationship has been shorten. In the figure, D-frame is predicated with Disparity compensation prediction (DCP); P-frame is predicted with Motion compensation prediction (MCP); B-frame is bi-directionally predicted with MCP or DCP; P′-frame is predicted from D-frame and P-frame; B′-frame is predicted from D-frame and B-frame, or B-frame and P′-frame, thus both of P′-frame and B′-frame have MCP and DCP. In an inter-predicted frame, if the efficiency of MCP or DCP is unsatisfied in rate-distortion optimization process, intra-block is introduced.

TPM in Fig.4(a) is suitable for MVV with more temporal correlation, because more temporal predictions are efficiently utilized to eliminate temporal correlation of MVV. Similarly, SPM is designed for MVV with more inter-view correlation, while HPM suits for the MVV with hybrid correlation. The three prediction modes in Fig.4 have the same sub-prediction-structure, *i.e.* 9 gray frames. These 9 frames are encoded before the rest frames in a 2DGOP. The sub-prediction-structure and coding order enable the MVC encoder to analyze correlation characteristic of the MVV signal. And an appropriate prediction mode is then selected from mode candidates according to the correlation analysis.

The advantage of the above structure is that the correlation analysis can be directly completed in the encoding process without additional computational complexity, and the results of correlation analyses can be used to select prediction mode for the current 2DGOP immediately.

3. Mode-updating trigger and correlation analysis module

The mode updating trigger is in charge of mode updating. It adaptively determines whether the prediction mode should be changed or not. Let m_i be the number of Intra blocks (I-block) in the i-th frame predicted with MCP, d_j be the number of I-blocks in the j-th frame predicted with DCP, N_m and N_d be the numbers of frames predicted with MCP or DCP. The correlation representation coefficient, η_c, is defined as

$$\eta_c = \frac{1}{N_m}\sum_{i=1}^{N_m} m_i \Big/ \frac{1}{N_d}\sum_{j=1}^{N_d} d_j \tag{1}$$

If η_c is larger than 1, it indicates that the current 2DGOP of MVV possesses more inter-view redundancies so that the inter-view prediction is more efficient than temporal prediction. On the contrary, if η_c is smaller than 1, the current 2DGOP of MVV holds more temporal redundancies and temporal prediction is more efficient than inter-view prediction. In order to select appropriate prediction mode from the candidates, thresholds of η_c are defined to distinguish the correlation characteristic of current 2DGOP. For prediction modes given in Fig.4, two thresholds T_1 and T_2 ($0 \le T_1 \le 1 \le T_2$) are defined for prediction mode selection. (1) If $\eta_c < T_1$, TPM, the mode shown in Fig.4(a), will be used to encode current 2DGOP; (2) If $T_1 \le \eta_c \le T_2$, it means that temporal correlation is close to inter-view correlation, thus the HPM, prediction mode shown in Fig.4(b), will be selected; (3) If $\eta_c > T_2$, SPM, the mode shown in Fig.4(c), will be used.

Since the numbers of I-blocks in frames are directly output from the encoder, the above adaptive trigger does not bring any extra computational complexity for the encoder except calculation of Eq.(1) for each 2DGOP whose complexity is almost neglectable.

IV. Experimental Results and Analysis

1. Compression efficiency comparison

The experiments are implemented on H.264/AVC (JM8.6, main profile), and test MVV sequences include "Aquarium", "flamenco 1", "race 2", and "Xmas". For each sequence, ten 2DGOPs (*i.e.* 350 frames) are utilized. The four sequences, as shown in Fig.5, are jointed together as one MVV sequence so as to simulate scene switching of MVV. Here, "Xmas" is

down-sampled to 320 × 240, which is the original image size of the other three sequences, and the camera distance is with 30mm.

Fig. 5. Joint MVV sequence

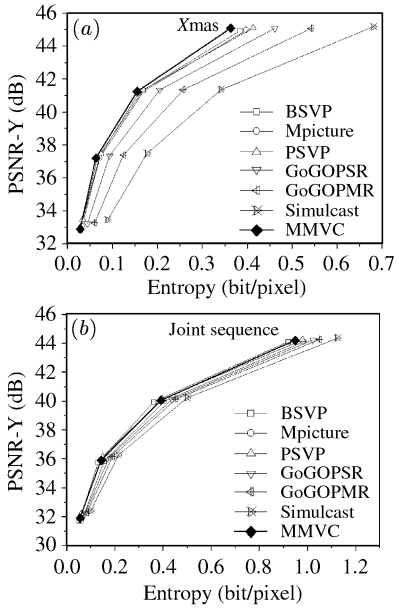


Fig. 6. Compression efficiency comparison. (a) Rate-distortion performance of 'Xmas' sequence; (b) Rate-distortion performance of the joint sequence

Fig.6 gives comparisons on compression efficiency. 'BSVP′ and 'PSVP′ denote SVP using P and B pictures[4], respectively. 'GoGOP SR' and 'GoGOP MR' represent GoGOP coding structures[8,9] utilizing single reference and multiple references, respectively. 'Mpicture' is the MVC scheme using multi-directional pictures[5]. Additionally, 'Simulcast' denotes simulcast scheme[11]. 'MMVC' indicates the proposed MMVC scheme. For "Xmas" in which inter-view correlation is the dominant, MMVC adaptively selects SPM as prediction structure and outperforms any other schemes over 0.5∼4dB, as shown in Fig.6(a). For other sequences, 'BSVP′ achieves the best rate-distortion performance in most cases. 'MMVC' is almost the same as 'BSVP′ in compression efficiency for the test sequences and better than other schemes, including 'GoGOP′, 'Mpicture' and 'PSVP′. Fig.6(b) illustrates compression efficiency of the joint sequence. Even though 'MMVC' is a bit inferior to 'BSVP′, but it is better than other schemes in compression efficiency.

2. Other performances comparisons

Besides compression efficiency, we use other six parameters to evaluate the performances of MVC schemes, including computational complexity, RA, view scalability and memory requirement.

(1) **Computational complexity.** We estimate the computational complexity of a MVC prediction structure by using the minimum number of reference frames of a 2DGOP, *i.e.* $PN_{\min}$.

(2) **Random accessibility.** Let $x_{i,j}$ be the number of frames which have to be decoded before the frame at (i,j) position is decoded in a 2DGOP with n time instants and m views. Let $p_{i,j}$ be the probability of the frame at (i,j) position being selected by a user, then the RA cost F_{av} and the maximum number of pre-decoded frames $F_{\max}$ are defined by

$$F_{\text{av}} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}p_{i,j} \tag{2}$$

$$F_{\max} = \max\{x_{i,j} | 0 < i \le n, 0 < j \le m\} \tag{3}$$

F_{av} and $F_{\max}$ indicate the average and maximum path length of RA.

(3) **Memory requirement.** Decoded picture buffer (DPB), which is used to store the reference frames, possesses most memory cost in H.264/AVC. Assume that each scheme adopts the optimal coding order to minimize the DPB size, represented by $DPB_{\min}$ here.

(4) **View scalability.** In this paper, we define two cost variables, F_{SV} and F_{DV}, to represent the average number of compulsorily decoded frames for a 2DGOP when single view or double views are displayed, respectively. Let O_n be a set of the frames in a 2DGOP and $X_{i,j}$ be a set of the compulsory decoded frames when the frame at (i,j) position is displayed, thus $X_{i,j} \subseteq O_n$. Suppose ρ_j is the probability that the user will watch the j-th view, and $\rho_{j,k}$ is the probability that both j-th view and k-th view will be accessed. F_{SV} and F_{DV} are defined as

$$F_{SV} = \sum_{j=1}^{m} [Card(U_{i=1}^{n} X_{i,j}) \cdot \rho_j] \tag{4}$$

$$F_{DV} = \sum_{j=1}^{m}\sum_{k=j+1}^{m} [Card[U_{i=1}^{n}(X_{i,j} U X_{i,k})] \cdot \rho_{j,k}] \tag{5}$$

where "$Card$" is cardinality of a set. Here, we assume that the view switching among the views is an equiprobable event, that is $\rho_j = 0.2$ and $\rho_{j,k} = 0.1$.

The performances of MMVC are associated with the selected prediction mode, which varies with the correlation of the encoded sequence. According to the correlation characteristics of the joint MVV sequences, MMVC adaptively selects TPM for "Aquarium" and "flamenco 1", HPM for "race 2" and SPM for "Xmas". We use the average value of encoding performances for each sequence to represent the performance of MMVC. As we can see from Table 1, MMVC performs best in random accessibility and the number of pre-decoded frames reduces about 9%∼300% compared with other schemes. Additionally, MMVC is a bit inferior to 'Simulcast' but much better than 'GoGOP′, 'Mpicture', 'PSVP′ and 'BSVP′ in complexity with 41%∼94% improvements, memory requirement with 40%∼220% improvements and view scalability with 37%∼92% improvements in F_{SV}, 25%∼62% improvements in F_{DV}. Although 'Simulcast' outperforms 'MMVC' in these four aspects, the compression efficiency of 'Simulcast' is the lowest among the compared schemes and it is much lower

than that of 'MMVC'. The gap is about 1 ∼ 4dB depending on MVV sequences. Therefore, 'MMVC' is the most efficient and balanced MVC scheme over all performance.

Table 1. Performance comparison among MVC schemes

Prediction structure	Random access cost		$PN_{\min}$	$DPB_{\min}$	View scalability	
	F_{av}	$F_{\max}$			F_{SV}	F_{DV}
Simulcast	3.0	6	30	1	7.0	14.0
GoGOP SR	3.6	9	111	16	12.6	21.7
GoGOP MR	4.6	14	114	16	15.4	25.2
PSVP	11.0	34	58	7	21.0	28.0
BSVP	7.5	19	83	7	21.0	28.0
Mpicture	6.0	20	97	16	16.0	22.4
M aquarium	2.2	3	54	3	7.8	14.6
M flamenco 1	2.2	3	54	3	7.8	14.6
V race 2	3.1	5	62	4	12.6	18.2
C xmas	3.5	6	64	5	15.4	21.7
Av. value	2.75	4.25	58.5	5*	10.9	17.3

Note: '*' represents that it is the maximum value for MMVC while encoding MVV

V. Conclusions

Temporal and inter-view correlations of multi-view video sequences vary along the time axis. They are influenced by the content of the video, illumination change, speed of moving objects and cameras, camera distance, frame rate, *etc.* We proposed a framework of Multi-modal multi-view video coding (MMVC) that fully utilize the correlation characteristic of multi-view video so as to achieve low complexity, low memory cost, fast random access and view scalability while maintaining high compression efficiency. Compared with some typical MVC schemes, MMVC can achieve better performance in random accessibility. Additionally, MMVC is better than the compared schemes in complexity for 41%∼94%, memory requirement for 40%∼220% and view scalability for 25%∼92% improvements. MMVC is regarded as the most efficient and balanced multi-view video coding scheme among the compared MVC schemes.

References

[1] S.U. Yoon, E.K. Lee, S.Y. Kim *et al.*, "A framework for representation and processing of multi-view video using the concept of layered depth image", *Journal of VLSI Signal Processing Systems*, Vol.46, No.2, pp.87–102, Mar. 2007.

[2] Y. Kim, J. Kim and K. Sohn, "Fast disparity and motion estimation for multi-view video coding", *IEEE Trans. on Consumer Electronics*, Vol.53, No.2, pp.712–719, May 2007.

[3] Y. Zhang, M. Yu and G. Jiang, "Evaluation of typical prediction structures for multi-view video coding", *ISAST Trans. on Electronics & Signal Processing*, Vol.2, No.1, pp.7–15, 2008.

[4] ISO/IEC JTC1/SC29/WG11 N6909: Survey of algorithms used for MVC. Hong Kong, Jan. 2005.

[5] S. Oka, T. Endo, T. Fujii, "Dynamic ray-space coding using multi-directional picture", *IEICE Technical Report*, pp.15–20, Dec. 2004.

[6] P. Merkle, A. Smolic, K. Mueller *et al.*, "Efficient prediction structures for multiview video coding", *IEEE Trans. on CSVT.*, Vol.17, No.11, pp.1461–1473, Nov. 2007.

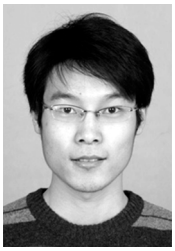
[7] ISO/IEC JTC1/SC29/WG11 N8218: Requirements on multi-view video coding v.7. Poznan, July 2006.

[8] H. Kimata, M. Kitahara, K. Kamikura, "Multi-view video coding using reference picture selection for free-viewpoint video communication", *Picture Coding Symposium*, pp.499–502, San Francisco, USA, Dec. 2004.

[9] H. Kimata, M. Kitahara, K. Kamikura *et al.*, "Low-delay multiview video coding for free-viewpoint video communication", *Systems and Computers in Japan*, Vol.38, No.5, pp.15–29, 2007.

[10] Y. Liu, Q. Huang, D. Zhao *et al.*, "Low-delay view random access for multi-view video coding", *in Proc. IEEE Int'l Symp. on Circuits, and Syst* (*ISCAS 2007*), New Orleans, USA, pp.997–1000, May 2007.

[11] U. Fecker, A. Kaup, "H.264/AVC-compatible coding of dynamic light fields using transposed picture ordering", *EUSIPCO 2005*, Antalya, Turkey, 2005.



ZHANG Yun received B.S. and M.S. degrees in information and electronic engineering from Faculty of Information Science and Engineering, Ningbo University, China, in 2004 and 2007. He is now a Ph.D. candidate at Institute of Computing Technology, Chinese Academy of Sciences of China. His research interests mainly include digital video compression and communications, multi-view video coding and content based video processing.

YU Mei received M.S. degree from Hangzhou Institute of Electronics Engineering, China in 1993, and Ph.D. degree from Ajou University, Korea, in 2000. She is now a professor at Faculty of Information Science and Engineering, Ningbo University, China. Her research interests include image/video coding and video perception.

JIANG Gangyi received M.S. degree from Hangzhou University, in 1992, and Ph.D. degree from Ajou University, Korea, in 2000. He is now a professor at Faculty of Information Science and Engineering, Ningbo University, China. His research interests mainly include video compression and communications, multi-view video coding and image processing. (Email: jianggangyi@126.com)