Fast depth coding techniques using early termination scheme

Fengsui Wang*, Guanling Wang

College of Electrical Engineering, Anhui Polytechnic University, Wuhu, 241000, China

Received 1 March 2014, www.cmnt.lv

Abstract

Depth video coding is a new technique that permits lower storage and transmission bandwidth compared with multi-view video coding (MVC). Therefore, fast depth video coding is necessary to reduce the computation complexity of the encoder for realizing the practical use. This paper proposed a fast encoding algorithm for depth video coding based on early termination scheme using texture and depth video correlation. Based on the observation that the Direct mode and Inter16×16 mode were highly possible to be the optimal mode, the proposed algorithm first computed the rate distortion cost of the Direct mode and compared with an adaptive threshold. If this rate distortion cost was smaller than the adaptive threshold, Direct mode was selected as the optimal mode. Otherwise, our approach proceeded with the early termination scheme and further checked whether the current microblock belonged to the low motion region using motion complexity for excluding impossible modes. Experimental results have shown that our proposed algorithm can significantly reduce encoding time with a negligible loss of coding efficiency of depth video, compared with the original joint multi-view video coding encoder.

Keywords: depth video coding, mode decision, texture-depth correlation, early termination

1 Introduction

With the new development in three dimensional display technologies, three dimensional TV (3DTV) realizes the human dream of seeing the realistic scene as the natural world. Free viewpoint TV (FTV) even makes it possible to offer the users free and interactive viewpoint selection within a certain range [1,2]. These new multimedia applications lie at the multi-view video, which is constitutive of more than one viewpoint by using multiple cameras from different viewpoints simultaneously capturing the same scene. Hence, it is a main challenge for multi-view video system to store and transmit the huge data amount induced by multiple views in the current limited network condition. To this end, a promising 3-D scene representation format, called as video plus depth, has been introduced and standardized by moving picture experts group (MPEG) [3], which has received increased attention and is considered as a next generation FTV format [4]. Figure 1 shows an example of the video plus depth format for multi-view sequence "Newspaper". This format uses 2-D texture video and the corresponding pre-pixel depth map to represent 3-D video, and allows to capture few viewpoints at the encoder side and to synthesize one or more "virtual" views of the real-world scene by making use of depth image-based rendering (DIBR) techniques [5]. Moreover, this representation format is both bandwidth efficient and backward compatible with the existent 2-D video. However, since the additional depth map makes greater amount of data and complexity compared with the traditional 2-D video coding systems, as a result, heavier computational burden for texture and depth video coding is introduced in depth video coding. Therefore, it is essential to develop efficient 3-D video coding techniques for practical applications.

FIGURE 1 An illustration of video plus depth format for sequence "Newspaper", consisting of texture video (left) and corresponding depth video (right)

Various video plus depth algorithms have been presented to compress depth video data in [6-8]. Merkle et al. [6] proposed a depth video coding method based on platelet, considering that the sharp boundaries of video objects were crucial to the synthesized view and the image was divided into four blocks and model functions using quad tree decomposition. A mesh based depth coding method was proposed in [7] by exploring a hierarchical triangle mesh. Liu et al. [8] proposed two depth compression techniques of trilateral filter and sparse dyadic mode to remove the coding artefacts and reconstruct the depth map, using the structure similarity between depth and corresponding texture video.

Multi-view video systems presented by MPEG usually adopt MVC to compress multi-view plus depth video because of better compatibility and acquiring higher coding efficiency. As a result, there exists tremendous coding computational complexity in encoder because

* ***Corresponding author's*** e-mail: fswang@ahpu.edu.cn

compressing the depth video requires nearly the same as complexity of the texture video. Moreover, MVC itself has great computational complexity due to exploiting many new techniques and encoding multiple viewpoints. Hence, fast algorithm is dispensable to reduce the computational complexity while keeping almost the same rate distortion (RD) performance. Recently, fast mode decision algorithms have been proposed to reduce computational burden for MVC-based texture video coding [9-11]. However, they cannot be directly applied to video plus depth video coding, since depth video is different description from the texture video. Hence, only few fast mode decision approaches are proposed for depth video coding [12-14]. Wang et al. [12] presented an early mode termination strategy based on difference detection. If the difference between the current microblock (MB) and the co-located MB in the original frame or reconstructed frame was zero. Then the current MB was considered to be static MB and only Direct mode was tested and all other modes were skipped. Zhang et al. [13] proposed a low complexity multi-view plus depth video coding method utilizing motion information sharing between the depth video and its corresponding texture video in order to reduce computational burden at expense of larger coding performance. Lin et al. [14] suggested a fast mode decision algorithm based on depth information for speeding up video encoding process.

In this paper, a fast early mode termination algorithm for depth video coding is proposed based on texture-depth and spatial correlations. In our method, the RD cost of Direct mode is first checked whether it is below an adaptive threshold for offering an early Direct mode termination chance. If so, Direct mode is selected as the optimal mode and mode decision process is early terminated. Otherwise, our approach further determines whether the current MB belonged to the low motion region for skipping unnecessary modes required to be checked. Experimental results reveal our proposed fast algorithm significantly reduces computational complexity while maintaining almost the same coding efficiency of depth video in comparison with the original joint multi-view video coding (JMVC) encoder.

The rest of the paper is organized as follows. In Section 2, the motivation of this work is described. In Section 3, the proposed fast algorithm for depth video coding is presented in detail. Experimental results are shown and discussed in Section 4. Finally, conclusion is given in Section 5.

2 Motivation

JMVC exploits the hierarchical B picture (HBP) prediction structure to reduce the redundancies existed in temporal and inter-view in multi-view video for improving coding efficiency. Figure 2 shows an illustration of HBP coding architecture of the reference frames from time domain and inter-view domain with 8 views and the length of group of picture (GOP) is equal to 8.

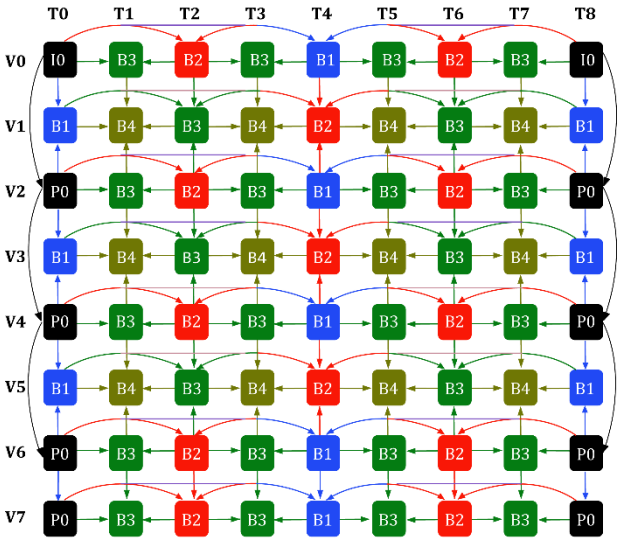

FIGURE 2 Hierarchical B picture architecture

In order to achieve higher coding efficiency, JMVC adopts a wider variable block-size partition, which can effectively encode various video contents. MVC supports the inter modes including Direct, Inter16×16, Inter16×8, Inter8×16 and Inter8×8, where the Inter8×8 can be further divided into Inter 8×8, Inter 8×4, Inter 4×8 and Inter 4×4 sub-modes (jointly denoted as P8×8 in this work), intra modes including Intra16×16, Intra8×8 and Intra4×4 (jointly denoted as Intra in this paper). In order to obtain the optimal mode, JMVC explores full mode decision and Lagrangian rate distortion optimization (RDO) function, namely calculating the RD cost of all the prediction modes and then selecting the one with the minimum RD cost as the optimal mode, where RD cost is:

$$J_{\mathrm{MODE}}(S_k, C_k \mid \lambda_{\mathrm{MODE}}) = \mathrm{SSD}(S_k, C_k) + \lambda_{\mathrm{MODE}} R(S_k, C_k), \quad (1)$$

where $J_{\mathrm{MODE}}(S_k, C_k \mid \lambda_{\mathrm{MODE}})$ is RD cost of MODE, S_k and C_k denote the kth original MB and the corresponding reconstructed MB, respectively. λ_{MODE} is the Lagrangian multiplier, and $R(S_k, C_k)$ represents the total bit rate after entropy coding. $\mathrm{SSD}(S_k, C_k)$ denotes the sum of squared difference between the original MB and reconstructed MB. Since MVC supports many prediction modes and RD cost computational load of each mode is quite time-consuming, the computational complexity of the whole mode decision in MVC is extremely high. In addition to mode decision, the optimal motion vector (MV) by exploiting motion estimation (ME) and disparity vector (DV) via utilizing disparity estimation (DE) are also decided by RDO. The RD cost of ME and DE can be computed as follows:

$$J_{\text{MOTION}}(S_k, C_k \mid \lambda_{\text{MOTION}}) = \text{SAD}(S_k, C_k) + \lambda_{\text{MOTION}} R(S_k, C_k), \quad (2)$$

where $\text{SAD}(S_k, C_k)$ is the sum of the absolute difference between the current MB and reference MB. $R(S_k, C_k)$ denotes the total number of bits for motion estimation or disparity estimation, and λ_{MOTION} represents the Lagrangian multiplier.

It is well-recognized that the Direct mode in MVC is fit for coding large block-size region with static or slow motion [10]. And we can observe from Figure 1 that the depth video often includes these regions. In other words, Direct mode should occupy higher possible to be the best mode in depth video coding. For this, the statistics and analysis of optimal modes are conducted for depth video coding by extensive experiments under the JMVC 8.0 on depth video sequences "Newspaper", "Lovebird1", and "Champagne_Tower". The test condition is described as follows: HBP prediction structure, 100 frames for each sequence with GOP = 8, quantization parameter (QP) = 24, 28, 32 and 36, RDO and context-adaptive binary arithmetic coding (CABAC) are enabled, and the search range of the ME and DE is ±64. The statistical result is shown in Figure 3.

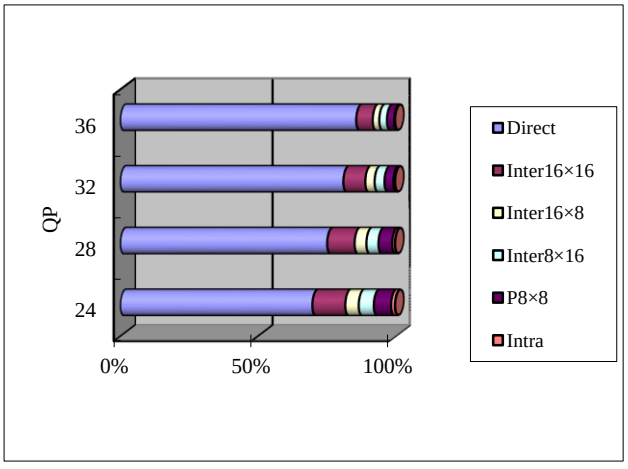

FIGURE 3 Illustration of the optimal mode distribution in depth video coding

We can easily see from Figure 3 that the Direct mode has the highest probability (about 80% on average) to be selected as the best mode. Moreover, this percentage is increased with the increment of QP value. It should be pointed that the Direct mode consumes very low complexity while other modes are time consuming. Therefore, great unnecessary checking process of the remaining modes can be skipped to save encode time if we can in advance decide whether Direct mode is the optimal mode.

In addition, the depth map represents distance from cameras plane to a certain point of scene objects, no more denotes luminance or chrominance. Hence, depth map is with the characteristics both video signal and data on Z-axis in world coordinates. In other words, the motion complexity (such as prediction modes and MVs) from the texture video and the depth video are much similar. Thus, the MV information from the texture video can be used to determine coding information of depth coding. In this paper, we also propose a mode decision method for depth coding by fully using motion vector correlation between the texture video and depth video.

3 Proposed fast algorithm

The depth video describes the depth information from the objects of the corresponding texture video, which usually contains abundant static or smooth regions with distinct boundary. Since the same video content is respectively presented in the texture video and the depth video, there exists a strong texture-depth correlation, such as mode and motion vector correlation. Thus, we can design our depth coding algorithm utilizing these sharing correlation information from the texture video. In addition, the depth video also contains a great deal of spatial correlation in the same frame. Our fast algorithm is proposed by making full use of these correlations as follows.

3.1 Early Direct mode termination

In the proposed method, we first employ prediction mode correlation between texture and depth video to design an early Direct mode termination scheme. Coding information of the previously encoded texture picture can be effectively shared and reused by using the correlation between the current MB in depth video and the corresponding MBs in texture video, due to content similarity between texture video and depth video. Intuitively, for those static or low motion objects, if cameras are fixed, the corresponding certain point depth values in depth map will maintain constant or consistent, not be affecting from the lightening mutation or fluctuation. Considering these cases, the prediction modes of the current MB in depth video can be inferred from that of the corresponding MBs in texture video using the texture-depth correlation, because the block-size partition of an MB can reflect its motion complexity. The MBs in static or low motion regions can be early determined to only select Direct mode as the optimal mode and coding performance is guaranteed. If the modes of low motion MBs are early terminated, computation amount will be greatly reduced. Based on above analysis, the early Direct mode termination method is described as follows.

As shown in Figure 4, MB_0 represents the current MB in the depth video. MB_1, MB_2 and MB_3 are its spatialadjacent MBs (i.e., left, top and top-right MBs in Figure 4), respectively. MB_4 is the corresponding co-located MB in texture video of the current MB, while MB_i for i=5,6,…,12 are the eight neighboring MBs of MB_4.

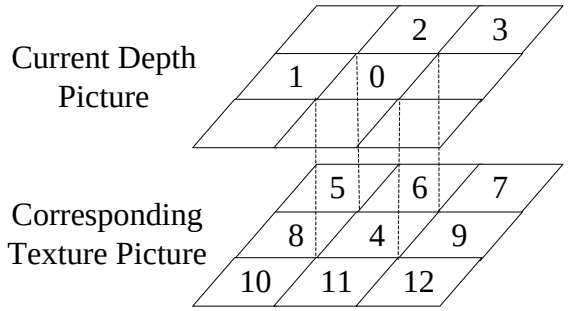

FIGURE 4 Texture-adjacent MBs of the current MB in depth video

The adaptive threshold for early termination in our proposed algorithm is the summation of the average value of RDO, J_m, and the minimal value of the RD cost of Inter16×16 in the corresponding texture frame, J_{min}. The current MB in depth frame corresponds to a co-located MB and neighboring eight blocks in the previously coded texture frame as shown in Figure 4. J_m can be computed from the RD cost of the twelve corresponding MBs by making full use of the texture-depth correlation between the current depth video and corresponding texture video, and spatial correlation among spatial-adjacent MBs in depth video, that is:

$$J_m = \frac{1}{N}\sum_{i=1}^{N} J_i \,, \tag{3}$$

where N is the number of the MBs. The adaptive threshold TH_1 can be derived from Equation (4), namely:

$$TH_1 = J_m + J_{\min} \,. \tag{4}$$

The proposed method firstly records the minimal RD cost of Inter16×16 mode for all the twelve MBs as shown in Figure 4. Then the adaptive threshold can be obtained by Equation (4), which is employed to decide if the Direct mode should be skipped by utilizing the average value of the RD cost of the previously coded spatial-adjacent MBs and co-located MB and its neighboring MBs in the corresponding texture video, plus the minimal RD cost value of the previously coded MBs for Inter16×16 mode. After acquiring this adaptive threshold, if the RD cost of the Direct mode for the current MB in depth video, J_D, is smaller than the adaptive threshold TH_1, and the optimal mode of the previously coded MBs is Direct mode, an early mode termination will be performed and Direct mode is selected as the optimal mode and mode decision process is early terminated.

3.2 Fast mode decision

Generally, the depth map has similar characteristics to the texture video as shown in Figure 1. The boundaries of the texture and depth video have similar shape, and the directions of object movements are the same in both texture and depth video coding. In other words, there exists high motion vector correlation between the texture and depth video. Hence, we can explore the motion vectors information from the corresponding texture video frame to reduce coding complexity of the depth map.

Firstly, the MV set $\{mv_1, mv_2, mv_3,\ldots, mv_{12}\}$ for the current MB in depth video is established as shown in Figure 4, by employing spatial-adjacent MBs and the co-located MB and its eight adjacent MBs in the texture video frame. Where $mv_i=(x_i,y_i)$ is the MV of the corresponding MB_i, for i=1,2,…,12, respectively. The motion complexity of the current MB in depth video can be obtained based on Equation (5), that is:

$$MC = \frac{1}{N}\cdot\sum_{i=1}^{N}\frac{1}{\beta_i}\left(|x_i|+|y_i|\right), \tag{5}$$

where MC denotes the motion complexity of the current MB in depth video, and i indicates the index of MB. N is the number of MBs, and β_i is a weight factor for the corresponding MB_i. If the optimal mode of the corresponding MB_i is Inter16×16, β_i is equal to 1. If the optimal mode of the corresponding MB_i is Inter16×8 or Inter8×16, β_i is equal to 2. By such analogy, weight factor β_i is documented in Table 1. In fact, it can be observed from Table 1 that an Inter16×16 block-size is standardized as a basic unit for various macroblock partition modes.

TABLE 1 Weight factors of different MB modes

Mode	**16×16**	**16×8**	**8×16**	**8×8**	**8×4**	**4×8**	**4×4**
β_i	1	2	2	4	8	8	16

In general, the larger the motion complexity, the more complex motion the MB is. Thus, every MB can be classified into two types (i.e. low motion and complex motion) according to the value of MC. If motion complexity MC is less than or equal to TH_2, then the current MB in depth map belongs to the region with low motion, otherwise, the current MB belongs to complex motion region. In order to determine the threshold TH_2, extensive experiments are performed based on various different depth video sequences under different MC. It can be found that when MC is less than or equal to 1, Inter16×16 mode is with high percentage (more than 90%) to be the optimal mode, hence TH_2 is set to 1 in our experiment. Therefore, if the early Direct mode termination condition is not granted, a fast mode decision method is adopted in our work. For a depth MB with low motion, it usually contains static or slow motion content, and the best mode is often Direct mode or Inter16×16 mode. For a depth MB with complex motion, various prediction modes are possible to be chosen as the optimal mode. Based on the above analysis, if the current depth MB is with low motion region, only Direct mode and Inter16×16 modes are checked, while MBs with complex motion test all prediction modes to select the one with the minimal RD cost as the optimal mode.

The above-mentioned early Direct mode termination and fast mode decision are summarized and depicted in flowchart as follows:

Step 1: Check whether the current depth MB is located in an anchor picture. If so, go to Step 6; otherwise, go to Step 2.

Step 2: Calculate the RD cost of Direct mode, J_D, and the adaptive threshold TH_1 according to Equation (4).

Step 3: If $J_D < TH_1$ and the optimal mode of the previously coded MBs is Direct mode, perform an early Direct mode termination, go to Step 7. Otherwise, go to Step 4.

Step 4: Computer the motion complexity of the current depth MB, *MC*, according to Equation (5).

Step 5: If $MC \leq TH_2$, only Direct mode and Inter16×16 modes are checked, and other modes are skipped, then go to Step 1; otherwise, go to Step 6

Step 6 Perform full mode decision and all the modes are checked to select the one with the minimum RD cost as the optimal mode, then go to Step 1 and proceed with next MB.

Step 7: The Direct mode is selected as the optimal mode and the mode decision process is early terminated, then go to Step 1 and proceed with next MB.

4 Experimental results

The proposed fast algorithm for depth coding including early Direct mode termination and fast mode decision is evaluated through simulation studies. JMVC 8.0, the reference software of MVC, is selected as the experimental platform to implement our approach. The results tested on six common-used depth video sequences released by the MPEG are shown in Table 2, which represent a wide range of motion complexity and frame sizes with 1024×768 (Book_Arrival, Newspaper, Lovebird1, Breakdancers), 1280×960 (Champagne_ Tower) and 1920×1088 (Pozan_Hall2), respectively. Three different views from each depth video sequence are chosen for experiment. The first and third views are used as the reference views respectively. The second view is used as the auxiliary view. The experimental setup is described as follows:

1) HBP coding structure and GOP length=8.
2) Each test sequence encodes 100 frames.
3) RDO and CABAC enabled.
4) QP = 24, 28, 32 and 36, respectively.
5) The search range is set as ±64.

Experimental results of the proposed fast algorithm are shown in Table 2. The Bjøntegaard delta PSNR (i.e., BDPSNR) and Bjøntegaard delta bit rate (i.e., BDBR) [10] are used to evaluate the averaged PSNR and bit rate changes between the proposed algorithm and JMVC, respectively. The time reduction ratio ΔT is defined as follows:

$$\Delta T = \frac{T_{proposed} - T_{JMVC}}{T_{JMVC}} \times 100\%, \quad (6)$$

where $T_{proposed}$ and T_{JMVC} represent the encoding time of the proposed algorithm and JMVC, respectively. The positive values represent increments whereas the negative values represent decrements.

TABLE 2 Experimental results of our proposed fast algorithm

Sequences	Views	BDPSNR(dB)	BDBR(%)	ΔT(%)
Book_Arrival	8,9,10	-0.04	+0.91	-69.13
Newspaper	2,3,4	-0.06	+1.48	-63.36
Lovebird1	6,7,8	-0.08	+1.64	-75.10
Breakdancers	0,1,2	-0.03	+0.79	-52.86
Champagne_Tower	39,40,41	-0.05	+1.18	-69.67
Pozan_Hall2	5,6,7	-0.04	+1.03	-64.59
Average		-0.05	+1.17	-65.78

One can see from Table 2 that the proposed fast algorithm can greatly reduce the computational complexity while keeping almost the same coding efficiency, compared with the full mode decision in MVC reference software. It has reduced encoding time about 65.78% on average, with maximum of 75.10% in "Lovebird1". The loss of coding efficiency is negligible in our proposed algorithm: only 0.05 dB PSNR loss on average and 1.17% increment in the bit rate on average. Therefore, our method can significantly reduce encoding time while keeping a good RD performance.

Table 3 compares the proposed fast algorithm with that proposed in reference [12]. The proposed algorithm shows better comparison results. Compared with algorithm presented in reference [12], a speed up of 14.97% on average can be acquired in our algorithm. Meanwhile, 0.21 dB BDPSNR improvement and 3.66% BDBR bit rate reduction are achieved by our proposed in comparison with reference [12]. In a word, the proposed method outperforms reference [12] in terms of both coding efficiency maintenance and computation complexity reduction.

TABLE 3 Performance comparisons between reference [12] (A) and the proposed fast algorithm (B)

Sequences	Method	BDPSNR(dB)	BDBR(%)	ΔT(%)
Book_Arrival	A	-0.05	+1.03	-25.35
	B	-0.04	+0.91	-69.13
Newspaper	A	-0.49	+10.77	-53.55
	B	-0.06	+1.48	-63.36
Lovebird1	A	-0.01	+0.10	-75.20
	B	-0.08	+1.64	-75.10
Champagne_Tower	A	-0.54	+7.95	-63.30
	B	-0.05	+1.18	-69.67
Average	A	-0.27	+4.96	-54.35
	B	-0.06	+1.30	-69.32

For a better illustration, Figure 5 shows the time-saving ratio comparison between method in reference [12] and the proposed method. It can be easily seen from Figure 5 that the proposed fast method can reduce much more computational complexity for various depth video test sequences.

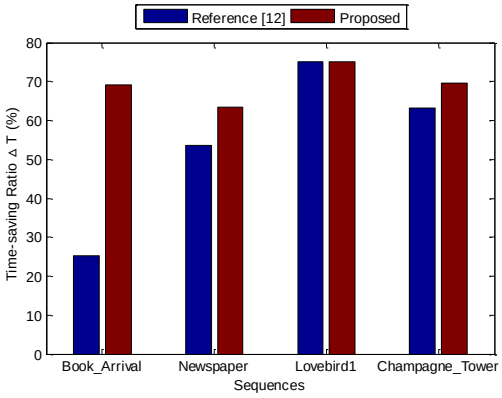

FIGURE 5 Comparison of time-saving ratio between Reference [12] and the proposed algorithm

5 Conclusion

This paper proposes an efficient fast algorithm for depth video coding based on the correlation between the texture video and depth video. For the current MB in depth map, the proposed method firstly provides an early Direct mode termination scheme for skipping the checking process of the remaining time-consuming modes. If the above early termination condition is not met, a fast mode decision method using motion complexity is adopted in order to further speed up encoding process. If the current depth MB is with low motion, only Direct mode and Inter16×16 modes are checked, while MBs with complex motion check all prediction modes. Experimental results have shown that the proposed algorithm achieves 65.78% time-saving on average with only 0.05 dB PSNR loss and 1.17% increment in the total bit rate on average, compared with the full mode decision in MVC.

Acknowledgments

This work was supported by the Scientific Research Foundation for Talents, Anhui Polytechnic University of China (2014YQQ006), and the Major Program of Scientific Research of Anhui Province, China (KJ2013A042).

References

[1] Muller K, Merkle P, Wiegend T 2011 *Proceedings of The IEEE* **99**(4) 643-56
[2] Tanimoto M, Tehrani M P, Fujii T 2012 *Proceedings of The IEEE* **100**(4) 905-17
[3] 3D-AVC Test Model 3 2012 *Joint Collavorative Team on 3-D Video Coding Extension Development of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 1st Meeting Stockholm: Sweden*
[4] Mori Y, Fukushima N, Fujii T, Tanimoto M 2008 *Proceedings of 3DTV conference* Istanbul: Turkey 229-32
[5] Kauff P, Atzpadin N, Fehn C, Muller K, Schreer O, Smolic A, Tanger R 2007 *Signal Processing: Image Communication* **22**(2) 217-34
[6] Merkle P, Morvan Y, Smolic A, Farin D, Muller K, With P.H.N.de, Wiegand T 2009 *Signal Processing: Image Communication* **24**(1-2) 73-88
[7] Kim S Y, Ho Y S 2007 *Proceedings of IEEE International Conference on Image Processing* San Antonio, TX: America 117-20
[8] Liu S, Lai P, Tian D, Chen C W 2011 *IEEE Transactions on Broadcasting* **57**(2) 551-61
[9] Deng Z P, Chan Y L, Jia K B, Fu C H, Siu W C 2012 *IEEE Transactions on Broadcasting* **58**(1) 24-33
[10] Wang F S, Zeng H Q, Shen Q H, Du S D 2013 *Signal Processing: Image Communication* **28**(7) 736-44
[11] Khattak S, Hamzaoui R, Ahmad S, Frossard P 2013 *Signal Processing: Image Communication* **28**(6) 569-80
[12] Wang M, Jin X, Goto S 2010 *Picture Coding Symposium Nagoya: Japan* 502-5
[13] Zhang Q W, An P, Zhang Y, Shen L Q, Zhang Z Y 2011 *IEEE Transactions on Consumer Electronics* **57**(4) 1857-65
[14] Lin Y H, Wu J L 2011 *IEEE Transactions on Broadcasting* **57**(2) 542-50

Authors

Fengsui Wang, 02.02.1981, Suzhou Anhui, China.

Current position, grades: lecturer in Anhui Polytechnic University, China.
University studies: PhD. degree in Circuits and System from Nanjing University.
Scientific interest: image processing, video signal processing, and computer vision.
Publications: 12 papers

Guanling Wang, 01.07.1971, Lujiang Anhui, China.

Current position, grades: associate professor in Anhui Polytechnic University.
University studies: MSc. degree in Detection Technology and Automation from Anhui Polytechnic University.
Scientific interest: image processing and embedded system.
Publications: about 30 papers.