DEPTH-MAP SUPER-RESOLUTION FOR ASYMMETRIC STEREO IMAGES

Diogo C. Garcia [*], *Camilo Dorea* [#] *and Ricardo L. de Queiroz* [#]

[*] Electronic Engineering Faculty at Gama, [#] Dept. of Computer Science
University of Brasilia, DF, Brazil
Email: diogo@image.unb.br, camilo@cic.unb.br, queiroz@ieee.org

ABSTRACT

We propose a mixed-resolution coding architecture for stereo color-plus-depth images, where encoding is performed at a low resolution, except for one of the color images. Super-resolution methods are proposed for the depth maps and for the low-resolution color image at the decoder side. Experiments are carried out for several real and synthetic images and reveal a reduction in complexity at the encoder associated with average PSNR gains of the super-resolved low-resolution color image with respect to regular interpolation.

Index Terms— Super-resolution, mixed-resolution, depth map, multiview.

1. INTRODUCTION

Multiview video and 3D video technologies attract great interest from researchers in academia and industry because there are many open questions, such as disparity estimation between views [1], multiview video coding (MVC) [2] and the corresponding computational burden [3, 4].

Based on binocular suppression theory [5–7], bit-rate and computational complexity reduction can be simultaneously achieved by mixed-resolution, or asymmetric coding of the stereo images [8–10]. The objective quality of low-resolution views can be recovered at the decoder side, using high-frequency information from neighboring views [11].

Bit-rate and computational-complexity reduction can also be achieved through low-resolution depth-map encoding [12–14], as depth maps are mainly composed of smooth areas, delimited by sharp boundaries. Down– and up-sampling methods can take advantage of these characteristics to improve the view-rendering quality at the decoder side.

In this paper, we propose a mixed-resolution architecture for stereo-plus-depth images which incorporates a super-resolution method for the low-resolution color image [11] and introduces a novel super-resolution method for the depth maps. In this manner, the bit-rate of the system is further reduced when compared to prior work [11–14] and computational complexity is relieved from the encoder and relocated, in part, to the decoder. At the decoder, the quality of the color-image view at a low-resolution is improved with the aid of the color-image view at a full resolution, as well as the super-resolved depth maps. Results show the effectiveness of depth map super resolution as well as coding efficiency gains with respect to [11].

This work was supported by CNPq under grants 470940/2010-7 and 310375/2011-8.

2. PROPOSED ARCHITECTURE

In order to improve the coding advantages of a mixed-resolution framework, super-resolution methods are proposed within the following architecture. Given views l and r of a scene, the color pairs $\mathbf{V}_l$ and $\mathbf{V}_r$ and depth map pairs $\mathbf{D}_l$ and $\mathbf{D}_r$ are defined. They are encoded and transmitted at a low resolution except for one of the color images, for example, $\mathbf{V}_r$. Separate down-sampling methods are used for the color image and the depth maps as presented in Fig. 1(a).

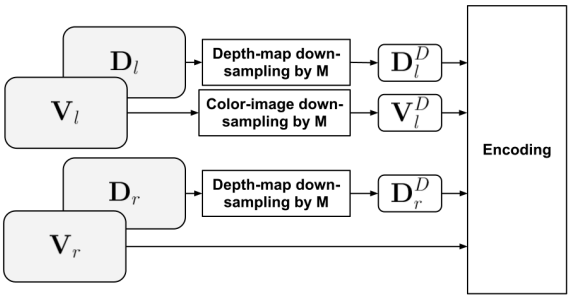
At the decoder side, the down-sampled, decoded depth maps $\tilde{\mathbf{D}}_l^D$ and $\tilde{\mathbf{D}}_r^D$ are mutually super-resolved using the method described in Subsection 2.1. The decoded, low-resolution color image $\tilde{\mathbf{V}}_l^D$ is then super-resolved, based on the method described in Subsection 2.2, using for such the super-resolved depth maps and the decoded, full-resolution color image $\tilde{\mathbf{V}}_r$ as illustrated in Fig. 1(b).

2.1. Depth-map super-resolution

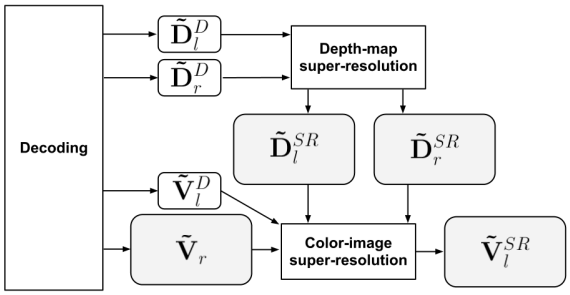
As in typical super-resolution [15], pixels in both depth maps are registered, displaced and combined to create a larger map. To do so, a pinhole camera model is assumed in which point coordinates from one view are projected onto the 3D space and then back to the other view, establishing a correspondence. Given any view i of the scene, the 3×3 intrinsic parameters $\mathbf{A}_i$, 3×3 rotation matrix $\mathbf{R}_i$, 3×1 translation vector $\mathbf{t}_i$ and corresponding depth map $\mathbf{D}_i$ are used to project pixel location (u, v) into world coordinates (x, y, z) [16]:

$$[x, y, z]^T = \mathbf{R}_i^{-1}\{D_i(u, v)\mathbf{A}_i^{-1}[u, v, 1]^T - \mathbf{t}_i\}. \quad (1)$$

The 3D points are then re-projected onto a new view j, yielding coordinates (u', v'):



(a)



(b)

Fig. 1: Proposed architecture: (a) given views l and r, the color image $\mathbf{V}_l$ and the depth maps $\mathbf{D}_l$ and $\mathbf{D}_r$ are down-sampled prior to encoding while color image $\mathbf{V}_r$ is encoded at full resolution; (b) at the decoder side, $\tilde{\mathbf{D}}_l^D$ and $\tilde{\mathbf{D}}_r^D$ are mutually super-resolved and, with the aid of $\tilde{\mathbf{V}}_r$, used to super-resolve $\tilde{\mathbf{V}}_l^D$ resulting in $\tilde{\mathbf{V}}_l^{SR}$.

$$[u' * w', v' * w', w']^T = \mathbf{A}_j\{\mathbf{R}_j[x, y, z]^T + \mathbf{t}_j\}. \quad (2)$$

Using this point correspondence among views, the proposed depth-map super-resolution method works as follows. At the encoder side, the depth maps are down-sampled by a factor M, creating $\mathbf{D}_l^D$ and $\mathbf{D}_r^D$. At the decoder side, the decompressed depth maps $\tilde{\mathbf{D}}_l^D$ and $\tilde{\mathbf{D}}_r^D$ are up-sampled by the same factor, rendering low-pass versions $\tilde{\mathbf{D}}_l^L$ and $\tilde{\mathbf{D}}_r^L$. For each of these depth maps, pixels from the adjacent depth map are forward mapped to the current depth map, based on Eqs. (1) and (2) and subject to a consistency check among maps. In this manner, information from the adjacent map is consistently transferred to the current map. Finally, these enhanced maps are filtered by an $N \times N$-pixel median filter to form the super-resolved versions as depicted in Fig. 2.

The consistency check is performed prior to forward mapping and aimed at isolating imperfections which often plague depth maps. These may be caused by down-sampling, compression quantization, occlusions or imprecise estimation. Considering, without loss of generality, $\tilde{\mathbf{D}}_l^L$ as reference, point (u, v) in view l is mapped to (u', v') in view r through Eqs. (1) and (2). $\tilde{D}_r^L([u'], [v'])$ is selected (where [.] is the *round* operation) and mapped back to view l determining (u'', v''). If (u'', v'') falls inside a $1-$pixel radius around (u, v), $\tilde{D}_l^L(u, v)$ is considered a consistent pixel. If (u'', v'') falls outside this radius, then $\tilde{D}_l^L(u, v)$ is considered inconsistent.

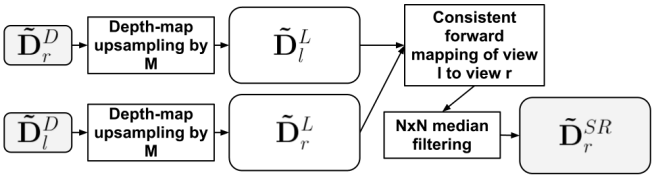


Fig. 2: Depth-map super-resolution method, using view r as reference. The down-sampled, decoded depth maps $\tilde{\mathbf{D}}_l^D$ and $\tilde{\mathbf{D}}_r^D$ are up-sampled to their original resolutions, rendering their respective low-pass versions $\tilde{\mathbf{D}}_l^L$ and $\tilde{\mathbf{D}}_r^L$. Next, $\tilde{\mathbf{D}}_l^L$ is consistently mapped to $\tilde{\mathbf{D}}_r^L$, in order to add sub-pixel information, and filtered by $N \times N$-windowed median filtering, resulting in the super-resolved version $\tilde{\mathbf{D}}_r^{SR}$. The same method is applied for view l.

Once consistent pixels have been determined, $\tilde{\mathbf{D}}_l^L$ is forward mapped onto $\tilde{\mathbf{D}}_r^L$ in order to enhance the latter. Consistent pixels of $\tilde{D}_l^L(u, v)$ project onto the corresponding location (u', v') of view r. Since position (u', v') is generally of sub-pixel precision, the closest integer position $\tilde{D}_r^L([u'], [v'])$ is replaced by $\tilde{D}_l^L(u, v)$, if $[u']\ mod\ M > 0$ and if $[v']\ mod\ M > 0$ (where $\alpha\ mod\ \beta$ is the remainder of α/β). This condition guarantees that the original information from $\tilde{\mathbf{D}}_r^D$ is kept. Thus, $\tilde{\mathbf{D}}_r^L$ is only updated with information from $\tilde{\mathbf{D}}_l^L$ at pixel positions where depth values have been created through interpolation within the up-sampling process. Analogously, the consistent pixels of $\tilde{\mathbf{D}}_r^L$ are used to enhance $\tilde{\mathbf{D}}_l^L$ through forward mapping.

After the projections of $\tilde{\mathbf{D}}_l^L$ onto $\tilde{\mathbf{D}}_r^L$ and $\tilde{\mathbf{D}}_r^L$ onto $\tilde{\mathbf{D}}_l^L$, the resulting depth maps are subject to $N \times N$ median filtering, in order to filter some of the noise and maintain sharp edges, rendering the super-resolved versions $\tilde{\mathbf{D}}_l^{SR}$ and $\tilde{\mathbf{D}}_r^{SR}$.

2.2. Color-image super-resolution

Following the depth-map super-resolution method, $\tilde{\mathbf{D}}_l^{SR}$ and $\tilde{\mathbf{D}}_r^{SR}$ are used in conjunction with $\tilde{\mathbf{V}}_r$ to super-resolve $\tilde{\mathbf{V}}_l^D$ as depicted in Fig. 3. The up-sampled version of $\tilde{\mathbf{V}}_l^D$, $\tilde{\mathbf{V}}_l^L$, represents a low-pass version of $\tilde{\mathbf{V}}_l$, so that the missing high-frequency components of $\tilde{\mathbf{V}}_l$ are to be estimated, using high-frequency information from $\tilde{\mathbf{V}}_r$ and pixel-correspondence information from $\tilde{\mathbf{D}}_l^{SR}$ and $\tilde{\mathbf{D}}_r^{SR}$. These estimates are added to $\tilde{\mathbf{V}}_l^L$, creating the super-resolved version $\tilde{\mathbf{V}}_l^{SR}$ [11].

To do so, we use the backward projection of $\tilde{\mathbf{V}}_r$ onto $\tilde{\mathbf{V}}_l$, based on Eqs. (1) and (2), to create $\tilde{\mathbf{V}}_{l|r}$. Only consistent pixels are projected, following the consistency check described in Section 2.1. Since the projected position (u', v') is usually calculated in sub-pixel precision, $\tilde{V}_{l|r}(u, v)$ is obtained through bilinear interpolation of $\tilde{\mathbf{V}}_r$ at the four closest integer positions to (u', v'). Inconsistent pixels are copied from $\tilde{\mathbf{V}}_l^L$, so that $\tilde{V}_{l|r}(u, v) = \tilde{V}_l^L(u, v)$ if $\tilde{D}_l^{SR}(u, v)$ and $\tilde{D}_r^{SR}([u'], [v'])$ are inconsistent.

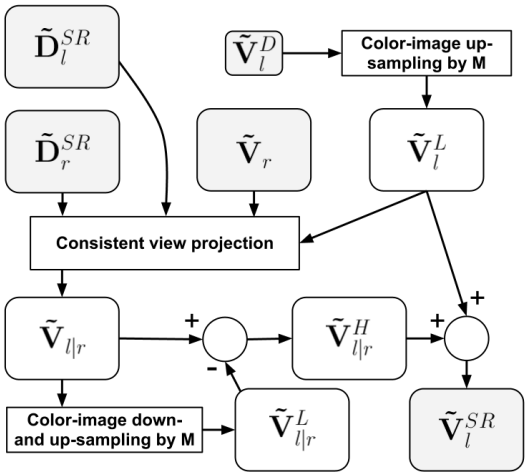


Fig. 3: Color-image super-resolution method. The super-resolved depth maps $\tilde{\mathbf{D}}_l^{SR}$ and $\tilde{\mathbf{D}}_r^{SR}$ are used to backward project $\tilde{\mathbf{V}}_r$ onto view l creating $\tilde{\mathbf{V}}_{l|r}$. Its low-pass version $\tilde{\mathbf{V}}_{l|r}^L$ is obtained by down– and up-sampling and its corresponding high-pass version $\tilde{\mathbf{V}}_{l|r}^H$ is found by subtracting $\tilde{\mathbf{V}}_{l|r}^L$ from $\tilde{\mathbf{V}}_{l|r}$. The super-resolved image $\tilde{\mathbf{V}}_l^{SR}$ results from adding $\tilde{\mathbf{V}}_{l|r}^H$ to $\tilde{\mathbf{V}}_l^L$.

$\tilde{\mathbf{V}}_{l|r}$ is then down– and up-sampled, generating its low-pass version $\tilde{\mathbf{V}}_{l|r}^L$. The high-frequency components are given by $\tilde{\mathbf{V}}_{l|r}^H = \tilde{\mathbf{V}}_{l|r} - \tilde{\mathbf{V}}_{l|r}^L$, which are then added to $\tilde{\mathbf{V}}_l^L$, creating its super-resolved version $\tilde{\mathbf{V}}_l^{SR}$. The process is illustrated in Fig. 3.

3. EXPERIMENTAL RESULTS

The proposed architecture was tested for a wide range of bit-rates with several real and synthetic stereo images and corresponding depth maps, as listed in Table 1. Frame 0 was used for all video sequences and, due to a lack of high-frequency content in the original images, *Ballet*, *Breakdancers*, *Cafe*, *Pantomime* and *Poznan Street* were resized *a priori*, as indicated in Table 1 along with the employed view numbers.

Table 1: Test sequences.

Sequence	Resolution	Full-res. view #	Low-res. view #
Barn2 [1]	416×368	2	6
Sawtooth [1]	432×368	2	6
Venus [1]	432×368	2	6
Ballet [17]	512×384	0	1
Breakdancers [17]	512×384	0	1
Cafe [18]	960×528	37	39
Pantomime [19]	640×480	2	3
Poznan Street [20]	960×544	3	4

The adopted codec was the H.264/AVC reference software JM 17.2 [21] operating in Intra mode to serve as a proof of concept even though one could extend the implementation to Inter coding and to other similar codecs such as HEVC [22]. Rate-control and RD optimization were disabled, and the quantization parameters (QPs) in the set $QP = \{22, 27, 32, 37\}$ were used for both color and depth. For the color-image down– and up-sampling, a high-performance, 6-tap Lanczos interpolation filter was used. The coding results were evaluated through average luma PSNR gains between curves [23] considering bit-rates of both views and depths.

In multiview services, such as free-viewpoint television, only the color images are meant for final consumption. Thus, the quality gains of the proposed architecture were measured in terms of PSNR differences between the super-resolved images $\tilde{\mathbf{V}}_l^{SR}$ resulting from usage of the various depth-map enhancement scenarios. The bit-rate for each case was considered as the total bit-rate of the stereo color-plus-depth image pairs at low or full resolutions, accordingly.

Three different comparisons were carried out. In the first comparison (C_1), in order to assess the penalties incurred by subsampling color and depth maps, the proposed architecture was compared to coding $\mathbf{D}_l$ and $\mathbf{D}_r$ at full resolution, so that no depth-map processing was required [11]. In the second comparison (C_2), in order to assess the depth-map super-resolution method, the proposed architecture was compared to simple interpolation in which depth maps $\tilde{\mathbf{D}}_l^L$ and $\tilde{\mathbf{D}}_r^L$ were directly applied to color-image super-resolution. Both C_1 and C_2 use the aforementioned Lanczos-based method for depth-map down– and up-sampling. In the third comparison (C_3), in order to verify the effect of the interpolation method, the setup of comparison C_2 was repeated, but with a different depth interpolation method. A median-based approach was applied, where each pixel of the down-sampled depth maps corresponds to the median value from $M \times M$-pixel blocks of the original map, and up-sampling is done through nearest-neighbour up-sampling followed by 3×3-pixel median filtering.

In terms of computational complexity, comparison C_1 yields an average reduction in total encoding time (color-plus-depth) of 18% and 22%, for $M = 2$ and $M = 4$, respectively, over all tested sequences. As for the decoder, C_1 presents an average increase in color super-resolution processing time of 172% for $M = 2$, and 103% for $M = 4$. None of the implementations were optimized for performance.

Table 2 presents the average luma PSNR gains of the color image for the three comparisons. Results for comparison C_1 show superior performance of the proposed architecture for real image data and inferior performance for synthetic images. For example, for $M = 2$, there is an average gain over full-resolution depth-map coding of 1.39 dB for sequence *Pantomime*, and an average loss of 0.79 dB for sequence *Barn1*. These results indicate that depth maps for real-image data may present erroneous values, which were corrected by the depth-map processing done in the proposed architecture, while depth maps for synthetic sequences were already at a

Table 2: Average Luma PSNR gain [23] of the color image, in dB, for the proposed architecture in different comparisons.

Sequence	M	C_1	C_2	C_3
Pantomime	*2*	1.39	0.08	0.20
	4	2.12	0.27	0.36
Ballet	*2*	0.51	-0.10	-0.09
	4	0.20	-0.04	-0.03
Breakdancers	*2*	0.57	0.03	0.08
	4	0.63	0.09	0.15
Cafe	*2*	0.62	0.25	0.24
	4	0.50	0.46	0.31
Poznan Street	*2*	0.08	0.04	0.04
	4	0.00	0.19	0.22
Barn1	*2*	-0.79	0.03	0.00
	4	-2.12	-0.06	-0.02
Barn2	*2*	-0.50	0.01	0.00
	4	-1.55	-0.05	0.03
Sawtooth	*2*	-0.33	0.08	0.07
	4	-1.70	-0.16	-0.06
Venus	*2*	-0.78	-0.12	-0.01
	4	-1.51	-0.10	0.07

higher quality.

Results for comparison C_2 show that the proposed depth-map super-resolution method offers gains over depth-map interpolation for most real image data, as high as 0.46 dB for sequence *Cafe*, $M = 4$. The only exception is sequence *Ballet*, which presents 0.10 dB and 0.04 losses for $M = 2$ and $M = 4$, respectively. As for the synthetic images, results also indicate small gains, again due to the high quality of the original depth maps. These results show the effectiveness of the proposed depth-map super-resolution method over simple depth-map interpolation.

Results for comparison C_3 are similar to those of C_2. Average gains are registered for all real image data, again, with the exception of *Ballet*. As with C_2, smaller gains are achieved for the synthetic images. These results indicate the robustness of the proposed architecture to changes in the depth-map interpolation method.

Fig. 4 presents details of the interpolated and super-resolved versions for sequence *Cafe*, view 39, frame 0, $QP = 22$. Significant visual quality improvements are perceived from the proposed architecture.

4. CONCLUSION

In this paper, we proposed a mixed-resolution architecture for stereo color-plus-depth images, where encoding is performed at a low resolution, except for one of the color images. Furthermore, super-resolution methods were proposed for the low-resolution depth maps and the low-resolution color image. Experiments over several real and synthetic images reveal a reduction in encoding complexity and average PSNR gains of the proposal when compared to traditional interpolation of the low-resolution color-image.

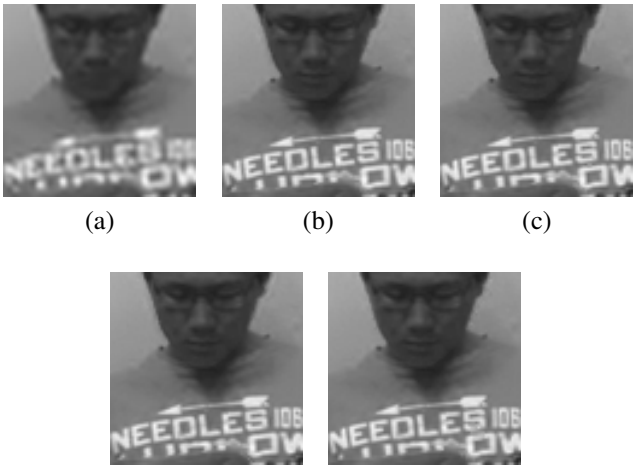


Fig. 4: Detail crops of the luma component of sequence *Cafe*, view 39, frame 0, $QP = 22$: (a) down- and up-sampled by 2 with Lanczos-based method, (b) super-resolved using simple interpolation of depth maps only, (c) super-resolved by proposed architecture using Lanczos-based down– and up-sampling of depth maps, (d) super-resolved by proposed architecture using median-based down– and up-sampling of depth maps, and (e) super-resolved with full-resolution depth maps. Results best seen on a screen.

5. REFERENCES

[1] D. Scharstein, and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr.-Jun 2002.

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audio- visual service," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), 2010.

[3] G. Cernigliaro, F. Jaureguizar, J. Cabrera, and N. García, "Fast mode decision for multiview video coding based on scene geometry", in *Int. Conf. Image Processing*, Hong Kong, China, pp. 3429–3432, Sept. 2010.

[4] Z.-P. Deng, Y.-L. Chan, K.-B. Jia, C.-H. Fu, and W.-C. Siu, "Fast iterative motion and disparity estimation algorithm for multiview video coding", in *3DTV-CON*, Tampere, Finland, Jun. 2010.

[5] B. Julesz, "Foundations of Cyclopean Perception", University of Chicago Press, 1971.

[6] P. Aflaki, M. Hannuksela, J. Häkkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video", in *Int. Conf. Image Process.*, Hong Kong, China, pp. 4021–4024, Sept. 2010.

[7] G. Saygili, C. Gurler, and A. Tekalp, "Quality assessment of assymmetric stereo video coding", in *Int. Conf. Image Process.*, Hong Kong, China, pp. 4009–4012, Sept. 2010.

[8] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, and J. Kim, "Asymmetric coding of stereoscopic video for transmission over T-DMB", in *3DTV-CON*, Kos Island, Greece, May. 2007.

[9] H. Brust, G. Tech, K. Mueller, and T. Wiegand, "Mixed resolution coding with inter view prediction for mobile 3DTV", in *3DTV-CON*, Tampere, Finland, Jun. 2010.

[10] Y. Chen, Y.-K. Wang, M. Gabbouj, and M. Hannuksela, "Regionally adaptive filtering for asymmetric stereoscopic video coding", *IEEE Int. Symp. Circuits and Systems*, pp. 2585–2588, May. 2009.

[11] D. Garcia, C. Dórea, R. de Queiroz, "Super resolution for multiview images using depth information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1321–1331, 2012.

[12] K. Oh, S. Yea, A. Vetro, and Y. Ho, "Depth reconstruction filter and down up sampling for depth coding in 3-D video," *IEEE Signal Process. Lett.*, vol. 1, no. 9, pp. 747-750, 2009.

[13] K.-J. Oh, A. Vetro, and Y-S. Ho, "Depth coding using a boundary reconstruction filter for 3-D video systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 350–359, Mar. 2011.

[14] Q. Liu, Y. Yang, R. Ji, Y. Gao, and Li Yu, "Cross-view down/up-sampling method for multiview depth video coding," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 295-298, May. 2012.

[15] S. Baker, and T. Kanade, "Limits on super-resolution and how to break them", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[16] P. Kauff, N. Atzpadin, C. Fehn, M. Mller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability", *Signal Processing Image Communication*, vol. 22, no. 2, pp. 217–234, Feb. 2007.

[17] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation", in *Proc. of ACM SIGGRAPH*, pp. 600–608, Los Angeles, CA, Aug. 2004.

[18] E.-K. Lee, Y.-S. Kang, J.-I. Jung, Y.-S. Ho, "3-D video generation using multi-depth camera system", ISO/IEC JTC1/SC29/WG11, M17225, pp. 001-008, Jan. 2010.

[19] Nagoya University FTV test sequences: *http://www.tanimoto.nuee.nagoya-u.ac.jp/*

[20] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, "Poznań Multiview Video Test Sequences and Camera Parameters," ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050, Xian, China, October 2009.

[21] JM H.264/AVC reference software: *http://iphome.hhi.de/suehring/tml/*.

[22] ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/16, "Joint Call for Proposals on Video Compression Technology," *WG11 Doc. N11113* and *ITU-T Q6/16 Doc. VCEG-AM91*, Kyoto, Jan. 2010.

[23] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves", VCEG-M33 13th Meeting, Austin, TX, Apr. 2001.