COMBINING SINGLE-IMAGE AND MULTIVIEW SUPER-RESOLUTION FOR MIXED-RESOLUTION IMAGE PLUS DEPTH DATA

Thomas Richter, Jürgen Seiler, Wolfgang Schnurrer, Michel Bätz, and André Kaup

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg (FAU), Cauerstr. 7, 91058 Erlangen, Germany

ABSTRACT

In mixed-resolution multiview setups, a scene is captured from various viewpoints with cameras having different spatial resolutions. Compared to full-resolution systems, mixed-resolution setups allow for savings with respect to data transmission, storage, and costs. However, for applications like free viewpoint television, high-quality images are required for all available camera perspectives. Therefore, high-resolution cameras can be used to increase the image quality of a neighboring low-resolution view. Due to occlusions, some parts of the scene are invisible in the high-resolution reference views and thus cannot be directly synthesized from the neighboring perspectives. In this paper, we propose to integrate the idea of single-image super-resolution to better handle occluded areas and thus to improve the super-resolution quality for mixed-resolution multiview images. For a downsampling factor of 4, the proposed method achieves an average gain of 0.53 dB with respect to a comparable multiview super-resolution approach.

Index Terms— Multiview, Super-Resolution, Mixed-Resolution

1. INTRODUCTION

Super-resolution (SR) is a widely discussed area in the field of image and video processing and aims at increasing the image quality for a given low-resolution image or video sequence [1]. In single-image SR (SISR), the desired high-resolution output is estimated from the low-resolution input image itself. Therefore, in example-based SR, the core idea is to learn the relationship between given pairs of low- and high-resolution image patches. The learned relationship is used afterwards to estimate the missing high-frequency content for the low-resolution input block [2]. In [3], dictionaries are built by randomly choosing raw patches from a given set of training images. In contrast, [4] uses coupled dictionaries, jointly trained from low- and high-resolution patch pairs to sparsely represent the low-resolution input patch. The authors in [5] perform SR, using the idea of sparse representation combined with a postprocessing step based on natural image prior.

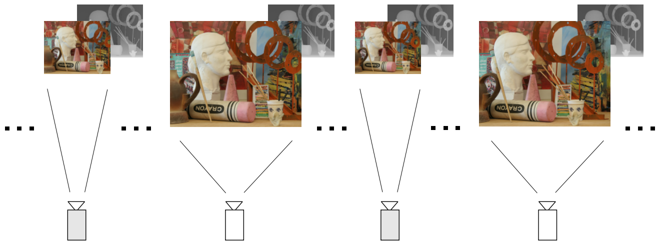

Fig. 1. MR-MVD scenario: A scene and the corresponding depth information is taken by cameras with different spatial resolutions.

For low-resolution video sequences, temporally adjacent frames can be utilized for SR [6]. These approaches typically require sub-pixel motion in order to super-resolve a low-resolution frame.

Besides using temporal information, multiview setups allow for exploiting information from neighboring camera perspectives. Fig. 1 shows an example for a mixed-resolution (MR) multiview video plus depth (MVD) scenario where a scene is taken by multiple cameras having different spatial resolutions. Additionally, the corresponding depth information is available at each viewpoint which can be either estimated [7] or recorded using time-of-flight cameras [8] or devices like the Microsoft Kinect [9]. In [10], a SR approach has been proposed for MR-MVD setups. By using the corresponding depth information, the neighboring high-resolution reference perspectives are projected onto the image plane of the low-resolution view. Afterwards, the required high-frequency information is extracted from the projection results. In our previous work [11], a method has been proposed to synthesize reliable high-frequency information even in the case of inaccurate depth acquisition or erroneous depth calibration. However, due to occlusions, some parts of the scene are only visible in the low-resolution view and cannot be captured from the available high-resolution cameras. Thus, depending on the multiview setup, large image areas might exist for which the missing high-frequency part cannot be directly synthesized. In previous work, e.g. [10] and [11], occluded areas are not explicitely considered. So, as a refinement step, the projected high-frequency information could

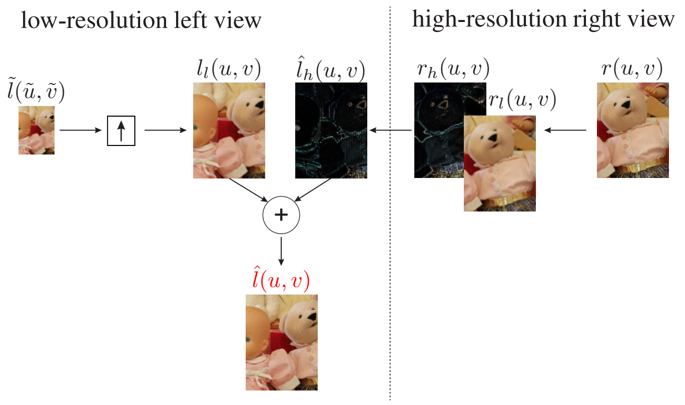

Fig. 2. Basic concept of HF-SYN. The high-frequency part of the reference perspective is projected and added to the upsampled low-resolution image in order to obtain the result $\hat{l}(u,v)$.

be extrapolated from the regions where it is known into the remaining unknown image areas using signal extrapolation techniques, such as [12] or [13]. However, for MR stereo setups, large occluded areas likely occur at the image border making the extrapolation very challenging.

Inspired by [14], in this paper, we propose to incorporate the idea of SISR in order to improve the SR quality for MR multiview images, especially in occluded areas. The rest of the paper is structured as follows. The basic concept of SR based on high-frequency synthesis is discussed in Section 2. The proposed combination is explained in Section 3. Simulation results are given in Section 4. Finally, the conclusions are presented in Section 5.

2. SUPER-RESOLUTION BASED ON HIGH-FREQUENCY SYNTHESIS

The basic concept of high-frequency synthesis (HF-SYN) [11], as used for this work, is depicted in Fig. 2. While the approach can be easily adapted to different multiview scenarios, the idea is shown for an MR stereo setup. Regarding multiview SR, stereo setups are the most challenging scenario, since only one neighboring high-resolution view is available, leading to large occluded areas.

Without loss of generality, the scene is taken by a high-resolution camera from the right, indicated by $r(u,v)$ and a low-resolution camera from the left, written as $\tilde{l}(\tilde{u},\tilde{v})$. The image coordinates on the low- and high-resolution grids are written as $(\tilde{u},\tilde{v})$ and (u,v), respectively. First, the low-resolution image $\tilde{l}(\tilde{u},\tilde{v})$ is enlarged to match the spatial resolution of the reference camera, resulting in an upsampled low-resolution image $l_l(u,v)$. Then, the reference perspective is divided into a low- and a corresponding high-frequency part. The low-frequency part $r_l(u,v)$ is obtained by filtering, downsampling, and interpolation. The corresponding high-frequency part, written as $r_h(u,v)$, is computed afterwards by subtracting $r_l(u,v)$ from $r(u,v)$. Since this work considers a video plus depth scenario, depth-image-based rendering (DIBR) [15] can be used to project the high-frequency part from the reference view onto the image plane of the low-resolution camera perspective.

Therefore, let $(u_\mathrm{r},v_\mathrm{r})$ be a pixel position in the high-frequency image $r_h(u,v)$ with subscript r indicating the right reference perspective. According to

$$\begin{pmatrix} x_\mathrm{w} \\ y_\mathrm{w} \\ z_\mathrm{w} \end{pmatrix} = \mathbf{R}_\mathrm{r}{}^{-1} \left(z \cdot \mathbf{A}_\mathrm{r}{}^{-1} \begin{pmatrix} u_\mathrm{r} \\ v_\mathrm{r} \\ 1 \end{pmatrix} - \mathbf{t}_\mathrm{r} \right), \tag{1}$$

the position is converted into three-dimensional world coordinates $(x_\mathrm{w},y_\mathrm{w},z_\mathrm{w})$. The intrinsic camera matrix is denoted as $\mathbf{A}$, while the extrinsic camera parameters are written as rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. The physical depth value, denoted as z, is computed from the corresponding depth map entry $d_\mathrm{r}(u_\mathrm{r},v_\mathrm{r})$. In a second step, the obtained 3D coordinates are projected onto the image plane of the low-resolution view via

$$z_\mathrm{l} \cdot \begin{pmatrix} u_\mathrm{l} \\ v_\mathrm{l} \\ 1 \end{pmatrix} = \mathbf{A}_\mathrm{l} \left(\mathbf{R}_\mathrm{l} \begin{pmatrix} x_\mathrm{w} \\ y_\mathrm{w} \\ z_\mathrm{w} \end{pmatrix} + \mathbf{t}_\mathrm{l} \right), \tag{2}$$

where subsricpt l denotes the upscaled left view. Finally, the discussed projection leads to the synthesized high-frequency image $\hat{l}_h(u,v)$ which is added to the low-resolution view $l_l(u,v)$ in order to create the desired high-resolution image $\hat{l}(u,v)$.

3. PROPOSED COMBINATION OF SINGLE-IMAGE AND MULTIVIEW SUPER-RESOLUTION

The concept of HF-SYN, as discussed in the previous section, mainly suffers from two aspects. First, due to occlusions, some image parts might not be visible in the reference perspectives. Thus, for those areas, the missing high-frequency information cannot be directly synthesized. For handling these occlusions, the high-frequency information could be extrapolated from the regions where it is known into the remaining unknown areas. However, depending on the configuration of the MR array, large unknown areas may exist, especially at the image border in case of stereo setups, making the extrapolation very challenging. As a second drawback, HF-SYN typically requires error-free depth information for projecting the high-frequency components and leads to annoying visual artifacts in case of inaccurate depth information [11].

Fig. 3 shows the block diagram of the proposed SR approach for MR multiview images. Therefore, the basic HF-SYN method is extended by first, a consistency check for rejecting invalid synthesized high-frequency information and second, the incorporation of SISR for a novel way of handling

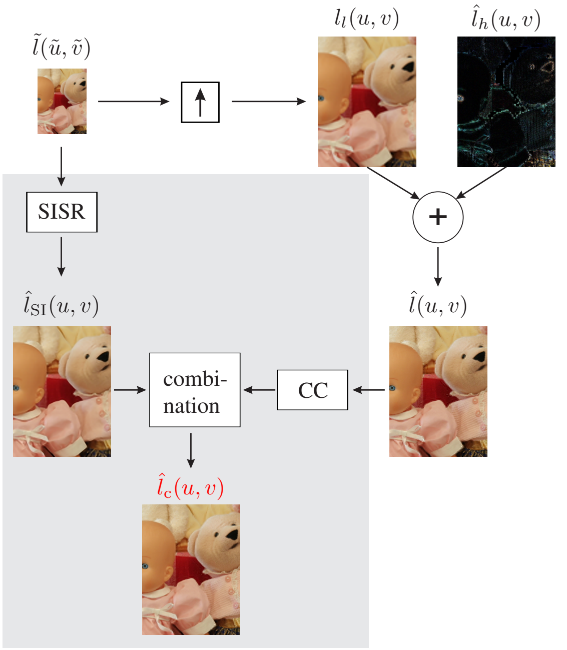
$\tilde{l}(\tilde{u},\tilde{v})$
$l_l(u,v)$
$\hat{l}_h(u,v)$
SISR
+
$\hat{l}_{\mathrm{SI}}(u,v)$
$\hat{l}(u,v)$
combi-nation
CC
$\hat{l}_{\mathrm{c}}(u,v)$

Fig. 3. Proposed integration of SISR for MR multiview images. The extension to HF-SYN is highlighted in gray with CC denoting the consistency check.

occluded areas. While the consistency check is discussed in Section 3.1, the proposed integration of SISR is explained in Section 3.2.

3.1. Consistency check

Due to inaccurate depth information, which likely occurs at object boundaries, the corresponding high-frequency parts may be projected to wrong positions. Since these projection errors cannot contribute to a convincing SR result, they have to be rejected. Therefore, let $(u_{\mathrm{r}}, v_{\mathrm{r}})$ be a pixel position in the right high-frequency image $r_h(u,v)$. By DIBR, the pixel is projected onto the image plane of the left low-resolution view, resulting in a position $(u_{\mathrm{l}}, v_{\mathrm{l}})$. The warped high-frequency information is then added to the low-resolution view at position $(u_{\mathrm{l}}, v_{\mathrm{l}})$, resulting in $\hat{l}(u_{\mathrm{l}}, v_{\mathrm{l}})$. By comparing $\hat{l}(u_{\mathrm{l}}, v_{\mathrm{l}})$ to the original pixel value $r(u_{\mathrm{r}}, v_{\mathrm{r}})$, the reliabilty of the underlying depth information can be verified as follows.

To account for potential illumination inconsistencies across different views, the consistency check is conducted in the YCbCr color space. Therefore, the maximum absolute difference is computed for the two color components Cb and Cr, excluding the luminance Y. If the difference exceeds a pre-defined threshold p, the corresponding high-frequency information is rejected, otherwise the pixel position $(u_{\mathrm{l}}, v_{\mathrm{l}})$ is marked in a map $m(u,v)$. The influence of the discussed consistency check is visualized in Fig. 4. While the left side shows the result of basic HF-SYN, the output after applying the consistency check is depicted in the middle image. The corresponding mask $m(u,v)$, marking all positions with valid high-frequency information in white color is shown on the right side.

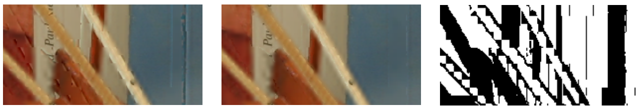

Fig. 4. Result of HF-SYN before (left) and after (middle) applying the consistency check. Mask $m(u,v)$, marking pixel positions with valid high-frequency information is shown on the right side.

3.2. Integration of single-image super-resolution

Typically, due to occlusions and the rejection of invalid information, the synthesized high-frequency image is not completely known. Since a complete high-pass image is desired for SR, the high-frequency content could be extrapolated into the occluded areas. However, the handling of large connected occluded parts is a very challenging task for common extrapolation approaches, such as [12] or [13]. Therefore, we propose to replace the high-frequency extrapolation by integrating the idea of SISR. As visualized in Fig. 3, the low-resolution input image $\tilde{l}(u,v)$ is not only enlarged by interpolation, resulting in the interpolated image $l_l(u,v)$, but also by applying an SISR approach, resulting in a super-resolved image $\hat{l}_{\mathrm{SI}}(u,v)$. Bascically, any kind of SISR method can be used. Regarding the image quality, especially for larger downsampling factors, SISR is typically better than pure interpolation but worse than directly synthesizing high-frequency information from neighboring perspectives. Therefore, the proposed combination of single-image and multiview SR is written according to

$$\hat{l}_{\mathrm{c}}(u,v) = \begin{cases} \hat{l}(u,v), & \forall (u,v)|m(u,v)=1 \\ \hat{l}_{\mathrm{SI}}(u,v), & \forall (u,v)|m(u,v)=0 \end{cases}, \quad (3)$$

where the combined SR result is denoted as $\hat{l}_{\mathrm{c}}(u,v)$ and the information from SISR is used to estimate the missing high-frequency information for both, occluded areas and pixel positions where the high-frequency information has been rejected by the consistency check.

4. SIMULATION RESULTS

The proposed combination of single-image and multiview SR has been tested for the datasets *aloe*, *art*, *baby1*, *bowling2*, *cloth1*, *cloth3*, *dolls*, *laundry*, *middl*, *moebius*, and *reindeer* [16], [17]. For all datasets, camera views 1 and 5 have been chosen as left and right views, respectively. The left low-resolution view has been simulated by filtering and downsampling. Therefore, a Lanczos kernel has been used and the downsampling factor has been varied between 2 and

Table 1. PSNR evaluation for all considered datasets and different downsampling factors in dB.

	aloe	*art*	*baby1*	*bowling2*	*cloth1*	*cloth3*	*dolls*	*laundry*	*midd1*	*moebius*	*reindeer*	avg. Δ PSNR
downsampling factor: 2												
BIC	33.21	37.38	36.08	37.40	35.61	37.19	36.53	36.06	35.49	36.96	36.36	-
HF-SYN	35.34	37.04	37.72	38.15	38.81	38.49	36.66	35.86	35.28	37.06	36.20	0.76
HF-SYN-FSE	35.51	37.43	37.73	38.19	38.83	38.58	36.88	35.97	35.36	37.22	36.32	0.89
Yang [4]	34.84	**38.62**	36.84	38.08	37.06	38.02	**37.55**	**37.29**	**36.26**	**37.76**	**37.09**	1.01
proposed	**36.20**	38.19	**37.97**	**38.61**	**39.06**	**38.75**	37.26	36.43	35.58	37.53	36.66	**1.27**
Kim [5]	34.56	**38.96**	36.89	38.13	36.64	38.02	**37.66**	**37.35**	**36.34**	**37.94**	**37.22**	1.04
proposed	**36.13**	38.28	**37.97**	**38.61**	**39.00**	**38.77**	37.27	36.47	35.58	37.57	36.66	**1.28**
downsampling factor: 4												
BIC	27.11	31.53	31.14	32.87	27.98	30.69	30.61	29.27	30.74	31.90	30.40	-
HF-SYN	30.74	32.35	34.67	34.61	33.82	34.34	32.41	31.25	32.40	33.19	32.19	2.52
HF-SYN-FSE	30.94	32.82	**34.70**	34.70	33.89	34.51	32.73	31.49	32.52	33.48	32.51	2.73
Yang [4]	27.42	32.13	31.26	33.14	28.40	30.94	31.07	29.79	31.21	32.30	30.76	0.38
proposed	**31.22**	**33.24**	34.61	**34.97**	**33.98**	**34.57**	**32.93**	**31.73**	**32.75**	**33.61**	**32.67**	**2.91**
Kim [5]	27.42	32.70	31.48	33.39	28.33	30.94	31.45	30.05	31.65	32.57	30.99	0.61
proposed	**31.24**	**33.54**	**34.73**	**35.13**	**33.96**	**34.64**	**33.12**	**31.86**	**32.97**	**33.74**	**32.81**	**3.05**

4 in both spatial dimensions. Since the image acquisition model of the low-resolution camera cannot be assumed to be known, bicubic filtering has been used for the reference perspective. For later upsampling, bicubic interpolation has been used for both camera perspectives.

For evaluation, the proposed combination has been compared against bicubic interpolation (BIC), high-frequency synthesis, as discussed in Section 2 (HF-SYN), and HF-SYN-FSE where the synthesized high-frequency information has been extrapolated using the Frequency Selective Extrapolation (FSE) from [12]. The proposed combination has been tested for the SISR approaches of [4] and [5].

For FSE, a blocksize of 4 has been used with a support area of 14 samples. The number of iterations has been set to 300 and the FFT size has been set to 32×32. For the discussed consistency check, the threshold p has been chosen as 4. For the approach of [4], dictionaries of size 1024 have been trained for both considered downsampling factors. While the patch size has been set to 5, the number of patches has been set to 100000.

Table 1 summarizes the PSNR evaluation for all considered multiview datasets and different downsampling factors. The last column gives the average gain with respect to BIC. For a downsampling factor of 2, the basic multiview approach HF-SYN achieves an average gain of 0.76 dB compared to BIC. Using FSE for high-frequency extrapolation leads to an additional gain of 0.13 dB. On average, the considered SISR approaches [4] and [5] perform slightly better than HF-SYN and for some test sets, such as *dolls*, *laundry* or *midd1*, they also beat the proposed combination. However, averaged over all datasets, our proposed combination outperforms both, the single-image and the multiview SR approaches. Compared to BIC, the SISR approaches result in average gains of 1.01 dB and 1.04 dB, whereas the proposed combination ends up with gains of 1.27 dB and 1.28 dB for [4] and [5], respectively.

For a downsampling factor of 4, the performance of the considered SISR approaches drops heavily, whereas HF-SYN achieves an averaged gain of 2.52 dB with respect to BIC. Again, the proposed combination outperforms both, the single-image and the multiview approaches, resulting in averaged gains of 2.53 dB and 2.44 dB compared to the SISR approaches. Averaged gains of 0.39 dB and 0.53 dB are achieved with respect to HF-SYN.

Compared to HF-SYN-FSE, the proposed method achieves average gains of 0.39 dB and 0.32 dB for downsampling factors of 2 and 4, respectively. Thus, the simulation results illustrate that the novel integration of SISR is a convincing way of handling occluded areas in MR multiview scenarios.

Fig. 5 finally shows the visual quality of the proposed SR method for image details of the *midd1* and *reindeer* datasets and a downsampling factor of 4. The figure shows from left to right, the original image and the SR results of HF-SYN, HF-SYN-FSE, and [5]. The right most column shows the proposed combination of single-image and multiview SR. For HF-SYN, some image parts cannot be directly synthesized from the reference perspective, resulting in annoying

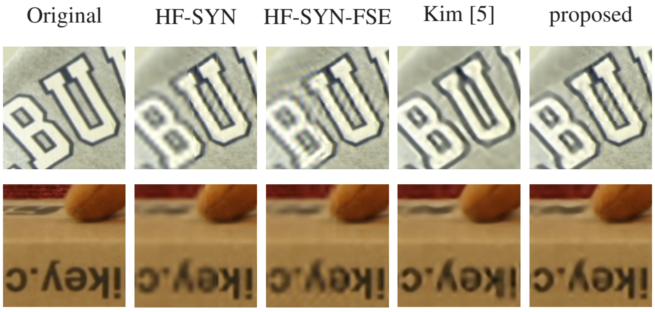

Fig. 5. Visual comparison between the SR approaches HF-SYN, HF-SYN-FSE, Kim [5], and the proposed combination.

visual artifacts at the transistions between low-resolution and super-resolved high-resolution parts. Due to the extent of these occluded areas, the desired high-frequency part cannot be convincingly reconstruced in HF-SYN-FSE. As can be seen for [5], the resulting image quality of SISR is typically better than simple interpolation but worse than extracting the missing high-frequency information from neighboring reference views. Finally, the proposed method combines the benefits of single-image and multiview SR, leading to a remarkable gain in visual quality.

5. CONCLUSION

In this paper, the combination of single-image and multiview super-resolution has been proposed for mixed-resolution image plus depth data. Using the idea of single-image super-resolution leads to a novel way of handling occluded areas in multiview scenarios. The simulation results illustrate that the proposed combination outperforms both, pure single-image and multiview super-resolution methods, resulting in an average gain of 0.53 dB for a downsampling factor of 4 with respect to the reference multiview approach. In addition, the proposed method leads to a remarkable gain in visual quality. For future work, the proposed idea could be extended to mixed-resolution multiview video data.

REFERENCES

[1] S.C. Park, M.K. Park, and M.G. Kang, "Super-Resolution Image Reconstruction: A Technical Overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, May 2003.

[2] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-Based Super-Resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[3] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image Super-Resolution as Sparse Representation of Raw Image Patches," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA, June 2008, pp. 1–8.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image Super-Resolution via Sparse Representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov 2010.

[5] K.I. Kim and Y. Kwon, "Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1127–1133, June 2010.

[6] M. Elad and A. Feuer, "Super-resolution reconstruction of image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 817–834, Sep 1999.

[7] D. Scharstein, R. Szeliski, and R. Zabih, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," in *Proc. IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV)*, 2001, pp. 131–140.

[8] S.B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions," in *Proc. IEEE Computer Science Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Washington, USA, Sep 2004.

[9] H. Jungong, S. Ling, X. Dong, and J. Shotton, "Enhanced Computer Vision With Microsoft Kinect Sensor: A Review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, Oct 2013.

[10] D.C. Garcia, C. Dórea, and R. de Queiroz, "Super Resolution for Multiview Images Using Depth Information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1249–1256, Sep 2012.

[11] T. Richter, J. Seiler, W. Schnurrer, and A. Kaup, "Robust Super-Resolution in a Multiview Setup Based on Refined High-Frequency Synthesis," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, Banff, Canada, Sep 2012, pp. 7–12.

[12] J. Seiler and A. Kaup, "Complex-Valued Frequency Selective Extrapolation for Fast Image and Video Signal Extrapolation," *IEEE Signal Processing Letters*, vol. 17, no. 11, pp. 949 – 952, Nov 2010.

[13] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel Regression for Image Processing and Reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, Feb 2007.

[14] M. Bätz, A. Eichenseer, J. Seiler, M. Jonscher, and A. Kaup, "Hybrid Super-Resolution Combining Example-Based Single-Image and Interpolation-Based Multi-Image Reconstruction Approaches," in *to appear in IEEE Int. Conf. on Image Processing (ICIP)*, Quebec City, Kanada, Sep 2015.

[15] C. Fehn, "Depth-Image-Based Rendering (DIBR) Compression and Transmission for a New Approach on 3D-TV," in *Proc. SPIE Electronic Imaging - Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, Jan 2004, pp. 93–104.

[16] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[17] H. Hirschmüller and D. Scharstein, "Evaluation of Cost Functions for Stereo Matching," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, USA, June 2007, pp. 1–8.