SUPER-RESOLUTION FOR DIFFERENTLY EXPOSED MIXED-RESOLUTION MULTI-VIEW IMAGES ADAPTED BY A HISTOGRAM MATCHING METHOD

Thomas Richter and André Kaup

Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany

ABSTRACT

Super-resolution is an important task in the image and video processing domain. In mixed-resolution multi-view scenarios, neighboring high-resolution reference perspectives can be used to increase the image quality of a given low-resolution target view. By using corresponding depth information, the required high-frequency part can be projected from a reference view onto the image plane of the target perspective. However, the contrast and thus the amount of high-frequency information in a reference view varies with the cameras exposure settings. As a consequence, the resulting super-resolution quality drops in case of exposure time variations between the different views. By incorporating a histogram matching method, the required high-frequency part can be efficiently adapted to the exposure settings of the target view. The simulation results show that the proposed adaption leads to an average PSNR gain of 0.63 dB for differently exposed mixed-resolution multi-view setups.

Index Terms— Multi-view, mixed-resolution, super-resolution, histogram matching, DIBR

1. INTRODUCTION

Ranging from home entertainment devices over professional film productions to security and surveillance scenarios, multi-camera setups get more and more important, leading to immersive viewing experiences and modern applications, such as autostereoscopic displays [1] or free viewpoint television (FTV) [2]. As a consequence, an increased interest in the area of multi-view image and video processing can be observed.

Typically, the flexibility and utilizability of multi-view approaches rises with the number of given camera perspectives. However, a larger number of cameras consequently leads to increasing costs and higher complexity regarding data transmission and storage. One way to restrict the required complexity is the usage of mixed-resolution (MR) setups, as shown in Figure 1, where the scene is captured by multiple cameras, providing images with various spatial resolutions and thus with different image qualities. Besides savings in financial and computational aspects, the usage of MR setups can be also motivated by the binocular suppresion

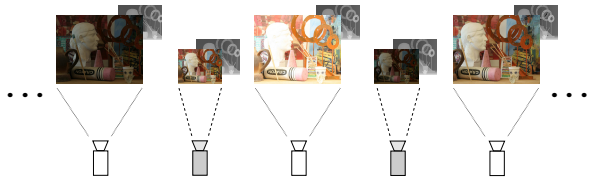

Fig. 1. MR setup for an MVD format: A scene and the corresponding depth information is taken by a set of low- and high-resolution cameras with different exposure times.

theory, stating that the human visual system combines two views with different spatial resolution, such that the perceived quality approximates the quality of the high-resolution view [3]. However, for applications like FTV, high-quality images are required from all available perspectives.

In order to increase the image quality of a given low-resolution input image, super-resolution (SR) approaches can be applied [4]. Coarsely, SR methods can be subdivided into single-image (SISR) and multi-image (MISR) approaches. SISR methods only rely on the low-resolution input image itself, maybe supported by proper dictionaries. For that, the authors in [5] assume that patches in natural images redundantly recur many times at different image scales. In [6], a database is used to learn the relationship between low- and corresponding high-resolution image patches. More recent dictionary-based approaches, including geometric dictionaries, deformable patches, or the usage of large internet-scale databases are proposed in [7]-[9]. In contrast, MISR approaches exploit information from multiple observations of the same scene. For low-resolution videos, sub-pixel shifts between neighboring frames are required for super-resolving the desired target image [4]. In [10] and [11], available high-resolution key frames are exploited by considering the video stream as temporal MR sequence. Finally, hybrid methods, such as [12], are available and aim at combining the basic concepts of both, SISR and MISR approaches.

For the considered case of MR multi-view setups, neighboring high-resolution views can be exploited for SR. Considering the widely used multi-view video plus depth format (MVD) [13], the missing high-frequency content can be extracted and projected from neighboring high-resolution reference views onto the image plane of the low-resolution target

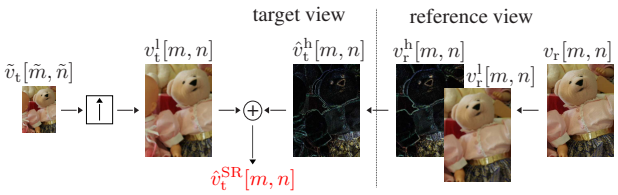

Fig. 2. SR based on high-frequency synthesis for an MR stereo setup.

view [14]. In order to account for potential depth inaccuracies, a robust extension, based on displacement compensation and high-frequency extrapolation, has been proposed in our previous work [15]. Up to now, the resulting SR performance has been only investigated for multi-view images captured with equal exposure times. However, differently exposed multi-view images can occur, either, by accident, if the individual exposure time is only controlled by the respective camera, or on purpose, for applications such as high-dynamic-range (HDR) video [16], [17].

Figure 1 shows the considered scenario. Without loss of generality, a scene is captured by a set of low- and high-resolution cameras with different exposure settings. In addition, the depth information is available at each viewpoint. The rest of the paper is structured, as follows. Section 2 introduces the basic concept of SR based on high-frequency synthesis, being the state-of-the-art SR approach for MR-MVD scenarios. Then, Section 3 discusses the proposed adjustment for differently exposed images, based on a histogram matching method. Simulation results are given in Section 4. The paper finally concludes with Section 5.

2. SUPER-RESOLUTION BASED ON HIGH-FREQUENCY SYNTHESIS

The main idea of SR based on high-frequency synthesis [15] (HF-SYN) is depicted in Figure 2. Without loss of generality, the figure shows an MR stereo setup, capturing a scene with a low-resolution camera from the left and a neighboring high-resolution camera from the right. The recorded images are written as $\tilde{v}_\mathrm{t}[\tilde{m},\tilde{n}]$ and $v_\mathrm{r}[m,n]$, where the low-resolution image is indicated by a tilde. The target and reference perspectives are denoted by subscripts t and r and the two spatial image coordinates are described by m and n, respectively.

First, by using an image interpolation method, the target view $\tilde{v}_\mathrm{t}[\tilde{m},\tilde{n}]$ is upsampled to the image dimension of the reference perspective, resulting in a low-frequency image $v_\mathrm{t}^\mathrm{l}[m,n]$. On the other side, the reference view is subdivided into a low- and a corresponding high-frequency part, written as $v_\mathrm{r}^\mathrm{l}[m,n]$ and $v_\mathrm{r}^\mathrm{h}[m,n]$, respectively. The low-frequency image is generated by filtering, downsampling, and interpolation. The high-frequency part is created afterwards by subtracting the low-frequency image $v_\mathrm{r}^\mathrm{l}[m,n]$ from the initial high-resolution reference view $v_\mathrm{r}[m,n]$. Using the idea of depth-image-based rendering (DIBR) [18], the high-frequency part is projected onto the image plane of the target view. For that, let $[m_\mathrm{r},n_\mathrm{r}]$ be a discrete pixel position in the high-frequency part of the reference view. According to

$$Z_\mathrm{t}\begin{bmatrix}m_\mathrm{t}\\ n_\mathrm{t}\\ 1\end{bmatrix}=\mathbf{A}_\mathrm{t}\left(\mathbf{R}_\mathrm{t}\left(\mathbf{R}_\mathrm{r}^{-1}\left(Z_\mathrm{r}\cdot\mathbf{A}_\mathrm{r}^{-1}\begin{bmatrix}m_\mathrm{r}\\ n_\mathrm{r}\\ 1\end{bmatrix}-\mathbf{t}_\mathrm{r}\right)\right)+\mathbf{t}_\mathrm{t}\right), \tag{1}$$

position $[m_\mathrm{r},n_\mathrm{r}]$ is projected onto the target view, resulting in a new position $[m_\mathrm{t},n_\mathrm{t}]$, where the intrinsic camera matrices are written as $\mathbf{A}$ and the extrinsic parameters consist of rotation matrices $\mathbf{R}$ and translation vectors $\mathbf{t}$. Again, the target and reference views are indicated by the subscripts t and r, respectively. The physical depth value is written as Z_r and is computed from the corresponding depth map entry $d_\mathrm{r}[m_\mathrm{r},n_\mathrm{r}]$. Applying (1) to each pixel position of the high-frequency image $v_\mathrm{r}^\mathrm{h}[m,n]$ results in the synthesized high-frequency part $\hat{v}_\mathrm{t}^\mathrm{h}[m,n]$. Finally, after projection, the synthesized high-frequency image is added to the target low-frequency part $v_\mathrm{t}^\mathrm{l}[m,n]$, leading to the SR result $\hat{v}_\mathrm{t}^\mathrm{SR}[m,n]$.

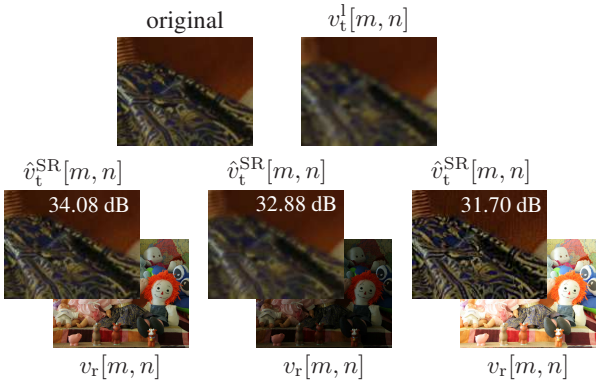

Fig. 3. SR results $\hat{v}_\mathrm{t}^\mathrm{SR}[m,n]$ for HF-SYN, depending on the exposure settings of the reference view $v_\mathrm{r}[m,n]$.

Now, for the special case of differently exposed multi-view images, as required for HDR video, the resulting SR quality highly depends on the exposure settings of the reference camera. This is emphasized in Figure 3. In the first image row, the figure depicts an original high-resolution image detail on the left and the interpolated low-frequency part $v_\mathrm{t}^\mathrm{l}[m,n]$ on the right side. The second row depicts reference views $v_\mathrm{r}[m,n]$ with different exposure settings and details of the resulting SR images $\hat{v}_\mathrm{t}^\mathrm{SR}[m,n]$. In the first case, as shown in the left figure column, the reference image is captured with a medium exposure time, which is equal to the exposure time of the target camera. As a result, it can be seen, that the obtained SR image is very close to the original image. The corresponding PSNR value is 34.08 dB. For the second example, a reference image has been used, captured with a shorter exposure time. Obviously, the reference image $v_\mathrm{r}[m,n]$ is much darker. In addition, the low-contrast reference view provides less high-frequency details, finally leading to a SR result of less sharpness and a PSNR value of 32.88 dB. The

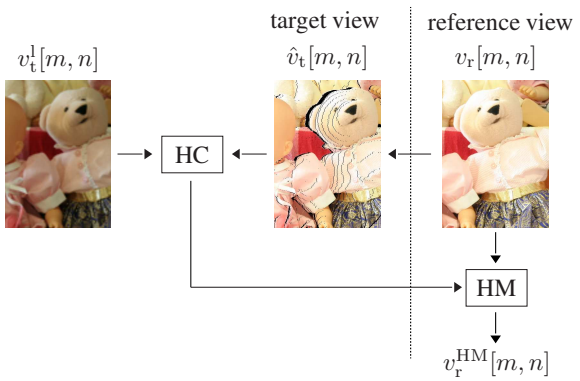

Fig. 4. Histogram-based adaption of $v_{\mathrm{r}}[m,n]$. While the histogram calculation is abbreviated by HC, histogram matching is denoted by HM.

third example, depicted in the rightmost column, illustrates the SR quality for the case of using a high-resolution reference view captured with a larger exposure time. Compared to the medium exposure time, the contrast in most image areas is significantly increased. This leads to amplified high-frequency parts and an unnaturally looking SR output image. In this example, the corresponding PSNR value drops to only 31.70 dB. To conclude, unmatched exposure settings between the conducted views negatively influence the resulting SR quality. For the considered scenario of differently exposed MR multi-view images, the next section shows how to efficiently adapt the synthesized high-frequency part to the exposure settings of the target low-resolution view, using a histogram matching method.

3. PROPOSED ADAPTION BASED ON HISTOGRAM MATCHING

The proposed adaption based on histogram matching is illustrated in Figure 4 for a MR stereo setup, where, without loss of generality, the high-resolution reference view $v_{\mathrm{r}}[m,n]$ is captured with a longer exposure time, compared to the low-resolution target view. For histogram matching, the method from [19] is chosen and applied to the considered MR-MVD scenario.

For that, by using the concept of DIBR, the reference view is projected onto the target image plane, resulting in the synthesized image $\hat{v}_{\mathrm{t}}[m,n]$. Then, the histograms are calculated for the target view $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$ and the synthesized reference image $\hat{v}_{\mathrm{t}}[m,n]$. For $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$, the histogram, denoted as $h_{v_{\mathrm{t}}^{\mathrm{l}}}[\nu]$, is written as

$$h_{v_{\mathrm{t}}^{\mathrm{l}}}[\nu] = \frac{1}{\sum_{m=1}^{M}\sum_{n=1}^{N} s_{\mathrm{t}}[m,n]} \sum_{m=1}^{M}\sum_{n=1}^{N} \delta[\nu, v_{\mathrm{t}}^{\mathrm{l}}[m,n]] \quad \forall [m,n] \mid s_{\mathrm{t}}[m,n]=1, \tag{2}$$

where the image dimension of $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$ is given by MxN. The binary map $s_{\mathrm{t}}[m,n]$ marks all positions with a valid synthesis result, when projecting from the reference view onto the target image plane. Thus, occluding areas are excluded from the histogram calculation. In addition, $\delta[a,b]$ is defined as

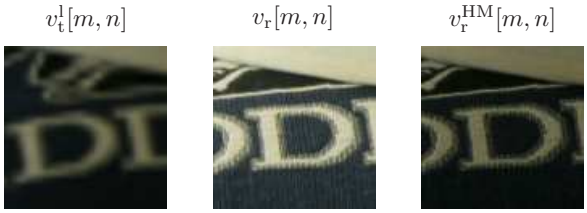

Fig. 5. From left to right: target view $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$, reference view $v_{\mathrm{r}}[m,n]$, and adapted reference view $v_{\mathrm{r}}^{\mathrm{HM}}[m,n]$.

$$\delta[a,b] = \begin{cases} 1, & \text{if } \ a=b \\ 0, & \text{else.} \end{cases} \tag{3}$$

In a next step, the cumulative histogram $c_{v_{\mathrm{t}}^{\mathrm{l}}}[\nu]$ is calculated by

$$c_{v_{\mathrm{t}}^{\mathrm{l}}}[\nu] = \sum_{i=0}^{\nu} h_{v_{\mathrm{t}}^{\mathrm{l}}}[i]. \tag{4}$$

The above described calculations are done in an analogous manner for the synthesized reference view $\hat{v}_{\mathrm{t}}[m,n]$, resulting in the cumulative histogram $c_{\hat{v}_{\mathrm{t}}}[\nu]$.

Now, after getting the cumulative histograms $c_{v_{\mathrm{t}}^{\mathrm{l}}}[\nu]$ and $c_{\hat{v}_{\mathrm{t}}}[\nu]$, the reference view $v_{\mathrm{r}}[m,n]$ is adapted to the exposure settings of the low-frequency target view $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$. For that, a matching function Q is used. According to

$$Q[\nu] = u \quad \text{with} \quad c_{v_{\mathrm{t}}^{\mathrm{l}}}[u] \leq c_{\hat{v}_{\mathrm{t}}}[\nu] < c_{v_{\mathrm{t}}^{\mathrm{l}}}[u+1], \tag{5}$$

the number of pixel value occurences in the synthesized reference view is mapped to the number of occurences in the target view. Then, by

$$v_{\mathrm{r}}^{\mathrm{HM}}[m,n] = Q[v_{\mathrm{r}}[m,n]], \tag{6}$$

the mapping function is applied to the reference perspective $v_{\mathrm{r}}[m,n]$ in order to adapt it to the exposure settings of the target view. The adapted reference view is denoted by $v_{\mathrm{r}}^{\mathrm{HM}}[m,n]$.

The performance of the discussed histogram matching method is exemplarily illustrated in Figure 5. The figure shows from left to right a detail of the low-frequency target view $v_{\mathrm{t}}^{\mathrm{l}}[m,n]$, the reference image $v_{\mathrm{r}}[m,n]$, captured with a larger exposure time, and the adapted reference image $v_{\mathrm{r}}^{\mathrm{HM}}[m,n]$. It can be clearly seen, that the histogram matching method can efficiently adapt a reference image to the exposure settings of the desired target view.

Finally, the high-frequency information is extracted from $v_{\mathrm{r}}^{\mathrm{HM}}[m,n]$ and projected onto the target image plane, as discussed in Section 2. However, since the histogram matching might be inaccurate for over- and underexposed pixels, pixel values falling below a threshold t_{min} or exceeding a threshold t_{max} are excluded from the upcoming synthesis process.

Table 1. PSNR evaluation for all considered datasets and downsampling factors of 2 and 4 in dB.

	aloe	*art*	*baby1*	*cloth1*	*dolls*	*midd1*	*rocks1*
downsampling factor: 2							
BIC	34.41	36.40	40.22	38.14	35.17	39.86	41.84
HF-SYN [15]	38.82	**38.22**	41.78	42.59	**37.87**	40.91	43.17
proposed	**39.23**	38.06	**42.27**	**43.00**	37.74	**41.92**	**43.96**
downsampling factor: 4							
BIC	29.14	31.62	35.57	31.87	30.45	35.61	37.19
HF-SYN [15]	34.93	34.55	39.22	38.55	34.46	38.33	40.10
proposed	**35.74**	**34.62**	**39.85**	**39.15**	**34.58**	**39.41**	**41.21**

Table 2. PSNR evaluation for all considered datasets and downsampling factors of 2 and 4 in dB. PSNR evaluated for pixels which are synthesized from only one reference view.

	aloe	*art*	*baby1*	*cloth1*	*dolls*	*midd1*	*rocks1*
downsampling factor: 2							
BIC	35.75	36.24	37.99	37.39	35.94	39.14	42.67
HF-SYN [15]	37.16	37.07	38.47	40.14	**37.71**	38.94	42.86
proposed	**38.92**	**37.23**	**39.88**	**42.60**	37.56	**41.07**	**44.56**
downsampling factor: 4							
BIC	30.47	31.52	33.10	31.08	30.99	33.96	38.14
HF-SYN [15]	33.10	33.38	35.54	35.51	33.95	35.64	39.86
proposed	**35.41**	**33.72**	**37.50**	**38.71**	**34.18**	**38.14**	**42.24**

4. SIMULATION RESULTS

The proposed SR scheme for differently exposed multi-view images has been tested for the datasets *aloe*, *art*, *baby1*, *cloth1*, *dolls*, *midd1*, and *rocks1* [20]. For all considered datasets, *view3* has been chosen as target view and *view1* and *view5* have been taken as reference perspectives, respectively. The exposure settings have been chosen, as follows. A short exposure time (*Exp0*) for *view1*, a medium exposure time (*Exp1*) for *view3*, and a large exposure time (*Exp2*) for *view5*. The low-resolution images have been simulated by filtering and downsampling. For that, an average filter has been used and the downsampling factor has been set to 2 and 4 in both spatial image dimensions. For upsampling to the original image size, bicubic interpolation has been chosen. The thresholds $t_{\min}$ and $t_{\max}$ have been set to 3 and 252.

Table 1 summarizes the PSNR evaluation for both downsampling factors and all considered multi-view datasets. The PSNR values are given for bicubic interpolation (BIC), the state-of-the-art method HF-SYN and the proposed SR scheme. Averaged over all datasets and a downsampling factor of 2, the proposed method results in a mean PSNR gain of 2.88 dB over BIC and 0.40 dB over HF-SYN. A maximum gain of 1.01 dB, compared to HF-SYN, has been achieved for the *midd1* dataset. For a downsampling factor of 4, the respective mean gains are 4.73 dB and 0.63 dB, compared to BIC and HF-SYN, respectively.

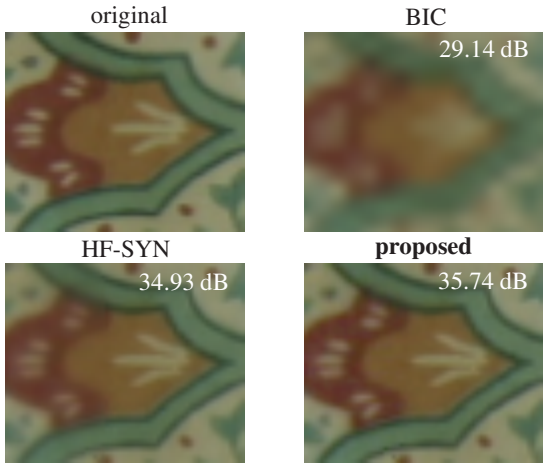

Fig. 6. Visual SR comparison between BIC, HF-SYN, and the proposed method.

As illustrated in Figure 3, the contrast and thus the extractable high-frequency parts tend to be reduced for shorter exposure times and amplified for larger exposure times. Thus, the two effects typically equalize for positions which are synthesized from both reference views. As a consequence, the gain of the proposed adaption is largest for pixels which are synthesized from only one reference view and are occluded in the other one. The corresponding PSNR evaluation is given in Table 2. Compared to BIC and HF-SYN, the proposed method achieves mean gains of 2.39 dB and 1.35 dB for a downsampling factor of 2 and even 4.38 dB and 1.85 dB for a downsampling factor of 4.

Figure 6 finally illustrates the visual SR performance of the proposed method for a downsampling factor of 4. The figure shows an image detail of the *aloe* dataset and the respective results for BIC, unrefined HF-SYN, and the histogram-based adaption. It can be seen, that, especially at the left image border, where the pixels have been synthesized from only one reference view, the proposed method provides a clear visual SR gain, compared to HF-SYN.

5. CONCLUSION

In mixed-resolution multi-view scenarios, a low-resolution image can be super-resolved by projecting high-frequency information from a neighboring reference view onto the image plane of the target perspective. For various applications, such as high-dynamic-range video, the different views are captured with different exposure times. However, the contrast and thus the amount of available high-frequency information in a reference view varies with the cameras exposure settings. In this paper, a novel histogram-based refinement method has been proposed, aiming at adapting the reference images to the exposure settings of the target view. The simulation results illustrate that, besides a noticeable gain in visual quality, an average PSNR gain of 0.63 dB is achieved.

6. REFERENCES

[1] H. Urey, K.V. Chellappan, E. Erden, and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proc. of the IEEE*, vol. 99, no. 4, pp. 540–555, Apr. 2011.

[2] M. Tanimoto, M.P. Tehrani, T. Fujii, and T. Yendo, "Free-Viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67–76, Jan. 2011.

[3] R. Blake, "Threshold Conditions for Binocular Rivalry," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3(2), pp. 251–257, 2001.

[4] 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.

[5] D. Glasner, S. Bagon, and M. Irani, "Super-Resolution from a Single Image," in *Proc. IEEE Int. Conference on Computer Vision*, Kyoto, Japan, Sep.-Oct. 2009, pp. 349–356.

[6] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-Resolution Through Neighbor Embedding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June-July 2004, pp. I–275 – I–282 Vol.1.

[7] S. Yang, M. Wang, Y. Chen, and Y. Sun, "Single-Image Super-Resolution Reconstruction via Learned Geometric Dictionaries and Clustered Sparse Coding," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4016–4028, Sep. 2012.

[8] Y. Zhu, Y. Zhang, and A.L. Yuille, "Single Image Super-resolution Using Deformable Patches," in *IEEE Int. Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2917–2924.

[9] L. Sun and J. Hays, "Super-resolution from internet-scale scene matching," in *IEEE Int. Conference on Computational Photography*, Seattle, WA, USA, Apr. 2012, pp. 1–12.

[10] F. Brandi, R. de Queiroz, and D. Mukherjee, "Super-Resolution of Video using Key Frames and Motion Estimation," in *Proc. IEEE Int. Conference on Image Processing*, San Diego, CA, USA, Oct. 2008, pp. 321–324.

[11] E.M. Hung, R.L. de Queiroz, F. Brandi, K.F. de Oliveira, and D. Mukherjee, "Video super-resolution using codebooks derived from key-frames," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1321–1331, Sep. 2012.

[12] 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 *Proc. IEEE Int. Conference on Image Processing*, Québec City, Canada, Sep. 2015, pp. 58–62.

[13] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-View Video plus Depth Representation and Coding," in *Proc. IEEE Int. Conference on Image Processing*, San Antonio, TX, USA, Sep. 2007, pp. I–201 – I–204.

[14] 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.

[15] T. Richter, J. Seiler, W. Schnurrer, and A. Kaup, "Robust Super-Resolution for Mixed-Resolution Multiview Image plus Depth Data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 5, pp. 814–828, May 2015.

[16] V. Ramachandra, M. Zwicker, and T. Nguyen, "HDR imaging from differently exposed multiview videos," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, May 2008, pp. 85–88.

[17] M. Bätz, T. Richter, J. Garbas, A. Papst, J. Seiler, and A. Kaup, "High dynamic range video reconstruction from a stereo camera setup," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 191–202, September 2013.

[18] 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.

[19] U. Fecker, M. Barkowsky, and A. Kaup, "Histogram-based prefiltering for luminance and chrominance compensation of multiview video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 9, pp. 1258–1267, Sep. 2008.

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