RECONSTRUCTION OF MULTIVIEW IMAGES TAKEN WITH NON-REGULAR SAMPLING SENSORS

Thomas Richter, Markus Jonscher, Wolfgang Schnurrer, Jürgen Seiler, and André Kaup

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

ABSTRACT

Increasing spatial image resolution is a widely discussed area in the field of image processing. In this paper, we present an efficient reconstruction approach for high-resolution images, taken with irregularly shielded low-resolution sensors in a multiview setup. The approach is based on the sparsity assumption, meaning that natural images can be efficiently represented in a transform-domain using only few coefficients. Utilizing information from adjacent cameras results in a better reconstruction quality for the central high-resolution view. Since neighboring camera perspectives might differ in illumination, the information from adjacent views has to be adapted to the view to be reconstructed. The simulation results show that a proper incorporation of information from neighboring views leads to a PSNR gain of up to 2.20 dB compared to a state-of-the-art singleview reconstruction approach.

Index Terms— Multiview, non-regular sampling, image reconstruction, sparsity, depth-image based rendering

1. INTRODUCTION

The growing popularity of applications like Free Viewpoint Television (FTV) and 3D Video (3DV) leads to an increased number of multiview image and video recordings. Besides larger costs regarding the camera array, a higher number of cameras results in large power requirements and leads to a higher complexity with respect to data storage and transmission. A common way to reduce complexity of the required camera array is to synthesize some of the views using depth-image based rendering (DIBR) [1]. Besides the usual case, where the reference cameras all have the same spatial resolution, a mixed-resolution (MR) setup, consisting of low- and high-resolution cameras, can be an appropriate alternative in many scenarios. Thereby, inter-view super-resolution approaches can be applied to increase the image sharpness of low-resolution views prior to the synthesis of unknown intermediate perspectives [2]-[3]. However, even in MR setups, spatially high-resolution reference cameras are required.

To avoid the need of high-resolution sensors in multiview scenarios, the idea of non-regular sampling can be used. The main idea is to take only a subset of pixels during image acquisition. This is based on the sparsity assumption of natural images, meaning that images can be sparsely represented in a transform-domain using only few transform coefficients. The desired high-resolution image has to be reconstructed afterwards. The authors in [4] presented a reconstruction approach based on compressed sensing (CS) for images with randomly distributed sampling points. A combination of CS and displacement compensation has been proposed by [5] in order to reconstruct non-regular images within multiview setups. However, choosing sampling points randomly still requires an underlying high-resolution image sensor.

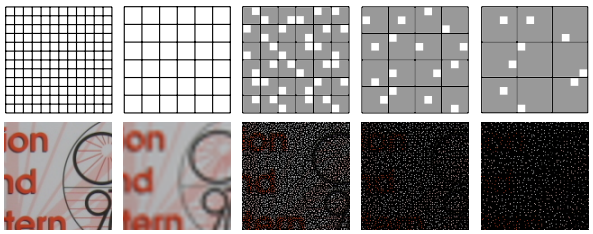

Fig. 1. Top row: from left to right: high-resolution sensor, low-resolution sensor and irregularly shielded low-resolution sensors with sampling factors $s = \{\frac{1}{4}, \frac{1}{9}, \frac{1}{16}\}$ (white areas being sensitive to light). Bottom row: corresponding image samples

In contrast to that, a new image acquisition architecture for low-resolution sensors based on non-regular sampling has been proposed [6]. Thereby, three out of four quadrants of each pixel of a low-resolution sensor are shielded irregularly, resulting in only one quadrant still being sensitive to light. The distribution of unshielded area leads to an irregular sampling pattern on a high-resolution grid. However, the sensor still has the image acquisition complexity and power requirements of a low-resolution one. In order to restore the desired image in high-quality, a reconstruction scheme, named Frequency Selective Extrapolation (FSE), has been used [7].

Fig. 1 visualizes in the first row from left to right the above mentioned high-resolution sensor, low-resolution sensor and irregularly shielded low-resolution sensors with sampling factors $s = \frac{1}{4}$, $s = \frac{1}{9}$, and $s = \frac{1}{16}$. Thereby, only the white areas are sensitive to light. Corresponding image examples are depicted in the second row of the figure. The sampling factor s describes the number of available sampling

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

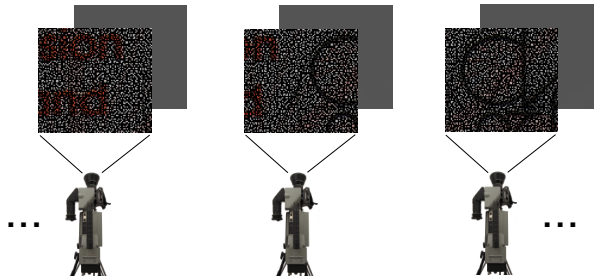

Fig. 2. Considered scenario: A scene is taken by several cameras with irregularly shielded low-resolution sensors. Additionally, the corresponding depth information is available.

points. As can be seen, the smaller s is chosen, the less spatial resolution is required for the underlying low-resolution grid.

In this paper, we propose a novel high-resolution reconstruction approach for non-regularly sampled images in a multiview setup. Fig. 2 shows the considered scenario. A scene is taken by several cameras with irregularly shielded low-resolution sensors. Since we consider a multiview video plus depth (MVD) scenario, the corresponding depth information is additionally available at each viewpoint. Usually, the reconstruction quality highly depends on the number of available sampling points. Hence, the depth maps are used in order to project pixels from neighboring reference cameras onto the image plane of the destination view. By adapting the synthesized information to the image to be reconstructed, new sampling points are defined and used as additional support for the final image reconstruction.

The rest of the paper is organized as follows. Section 2 covers the basic principle of using FSE for the reconstruction of irregularly sampled images. Section 3 presents our novel multiview reconstruction scheme. Simulation results are given in Section 4. The paper concludes with Section 5.

2. PREVIOUS WORK: RECONSTRUCTION OF IRREGULARLY SAMPLED IMAGES USING FSE

This section presents the main idea of using the Frequency Selective Extrapolation (FSE) for the reconstruction of images which are only partially known [7].

The reconstruction aims at extrapolating the given image content into the shielded areas, ending up with a dense high-resolution image. The FSE is carried out blockwise on the high-resolution grid, using an optimized processing order, according to [8]. For reconstructing the current block, the extrapolation area $\mathcal{L}$ which is exemplarily shown in Fig. 3, is considered. The extrapolation area can be subdivided into three groups. First, all available sampling points which are grouped in the support area $\mathcal{A}$. Second, all pixel positions that have already been reconstructed, forming the reconstructed area $\mathcal{R}$ and finally all samples that still have to be reconstructed, subsumed in the loss area $\mathcal{B}$. The spatial image coordinates are thereby denoted as (m, n).

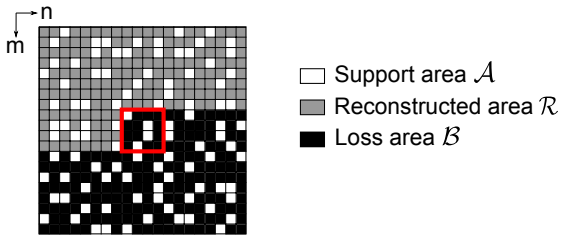

Fig. 3. Extrapolation area $\mathcal{L}$ consisting of support area $\mathcal{A}$, reconstructed area $\mathcal{R}$ and loss area $\mathcal{B}$. The current block to be reconstructed is marked in red.

Then, FSE aims at generating the parametric model

$$g(m,n) = \sum_{k \in \mathcal{K}} c_k \varphi_k(m,n) \tag{1}$$

as weighted superposition of two-dimensional Fourier basis functions $\varphi_k(m,n)$ with corresponding weights c_k. The indices of all available basis functions are summarized in the set $\mathcal{K}$. The model is generated iteratively. In every iteration the basis function that maximizes the decrease of the residual between the current model and the available signal is chosen and the corresponding weight is estimated. In this context, an exponentially decaying weighting function according to

$$w(m,n) = \begin{cases} \hat{\rho}^{\sqrt{\left(m-\frac{M-1}{2}\right)^2+\left(n-\frac{N-1}{2}\right)^2}} & \forall\,(m,n) \in \mathcal{A} \\ \delta\hat{\rho}^{\sqrt{\left(m-\frac{M-1}{2}\right)^2+\left(n-\frac{N-1}{2}\right)^2}} & \forall\,(m,n) \in \mathcal{R} \\ 0 & \forall\,(m,n) \in \mathcal{B} \end{cases} \tag{2}$$

is used in order to control the influence of each individual pixel position on the reconstruction process. Thereby, the speed of decay is controlled by $\hat{\rho}$ and the size of the extrapolation area is written as $M \times N$. Hence, pixels farther away from the origin of the block get less influence on the model generation. The influence of already reconstructed samples is additionally controlled by the factor δ. As proposed in [6], due to the small number of available sampling points, an additional probability is assigned to each basis function. In order to enforce a smooth reconstruction, the probability decreases for basis functions with increasing frequency.

The reconstruction quality of the above described FSE highly depends on the number of available sampling points. Therefore, the next section shows how information from neighboring cameras can be utilized in order to achieve a better reconstruction quality for the desired central high-resolution view.

3. PROPOSED MULTIVIEW RECONSTRUCTION

The basic overview of our proposed reconstruction scheme is given in Fig. 4. Without loss of generality, the scene is taken by three low-resolution cameras with irregularly shielded sensors. The images are therefore denoted as left image $l(m,n)$,

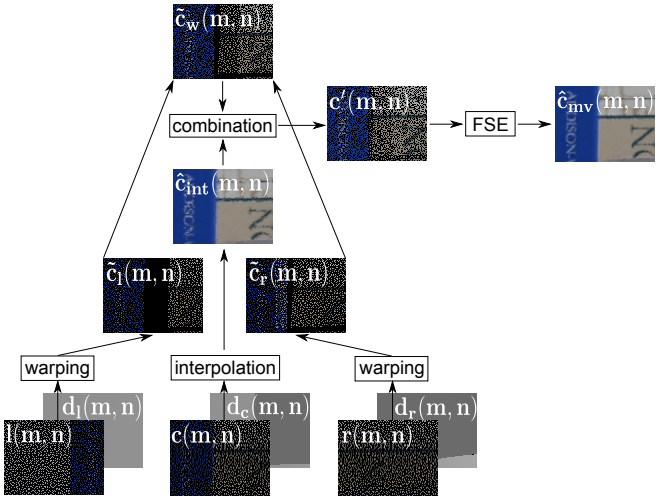

Fig. 4. Proposed multiview reconstruction scheme.

central image $c(m,n)$ and right image $r(m,n)$ while the corresponding depth maps are indicated by $d_l(m,n)$, $d_c(m,n)$, and $d_r(m,n)$, respectively. In a first step, the sampling points of the neighboring reference views are projected onto the image plane of the central view $c(m,n)$ using DIBR, as presented in [1]. Therefore, let (m_l, n_l) be a pixel position from a sampling point of the left reference view. Note that the warping process is equivalent for the right view. By utilizing the corresponding depth map entry, the projection can be written as

$$z_c \begin{pmatrix} m_c \\ n_c \\ 1 \end{pmatrix} = \mathbf{A}_c \left(\mathbf{R}_c \mathbf{R}_l^{-1} \left(z \cdot \mathbf{A}_l^{-1} \begin{pmatrix} m_l \\ n_l \\ 1 \end{pmatrix} - \mathbf{t}_l \right) + \mathbf{t}_c \right) \quad (3)$$

The intrinsic camera matrices are denoted as $\mathbf{A}$, $\mathbf{R}$ describes the rotation matrices and $\mathbf{t}$ are the translation vectors of the cameras with respect to the origin. The left and center camera are indicated by the corresponding subsrcipts l and c, respectively. The physical depth value, denoted as z, is computed from the corresponding depth map entry $d_l(m_l, n_l)$ [1].
By applying (3) to the sampling points of both reference views $l(m,n)$ and $r(m,n)$, two synthesized images $\tilde{c}_l\,(m,n)$ and $\tilde{c}_r\,(m,n)$ are obtained and can be considered as additional sampling information for the central destination view.

For combination of the two synthesis results, a simple depth-based comparison is used. A pixel is taken from one of the reference views, if no information is available from the other one. If sampling points from both reference images are projected to the same location in the desired central view, the one which is more to the foreground occludes the background pixel. The combination of $\tilde{c}_l\,(m,n)$ and $\tilde{c}_r\,(m,n)$ is denoted as $\tilde{c}_w\,(m,n)$.

The most intuitive way to reconstruct the central image would be to apply the above described FSE on an extended set of sampling points, consisting of both, original sampling points from $c(m,n)$ and projected sampling information from $\tilde{c}_w\,(m,n)$. For later comparison, this approach is indicated by FSE-DIBR. However, adjacent views might differ in illumination and the used depth information can not be assumed to be error-free. Since erroneous sampling points would negatively affect the final reconstruction quality, the synthesized information in $\tilde{c}_w\,(m,n)$ has to be adapted to the desired central view to be reconstructed. Therefore, the irregularly sampled central view $c(m,n)$ is first interpolated using fast bilinear interpolation. This results in a coarse approximation of the desired high-resolution image. The interpolated image is denoted as $\hat{c}_{\text{int}}(m,n)$ and is further used to adapt the information from $\tilde{c}_w\,(m,n)$ according to

$$c'\,(m,n) = \begin{cases} c\,(m,n)\,, & \forall\,(m,n) \in \mathcal{S}_c \\ t \cdot \tilde{c}_w\,(m,n) + & \\ +(1-t) \cdot \hat{c}_{\text{int}}\,(m,n)\,, & \forall\,(m,n) \in \mathcal{S}_{\tilde{c}_w} \wedge \bar{\mathcal{S}}_c \\ 0, & \text{else} \end{cases} \quad (4)$$

where $c'(m,n)$ is the extended central image including sampling information from the reference views. $\mathcal{S}$ defines the set of available sampling points and t describes a weighting factor in the range between 0 and 1. Thus, for the initially shielded pixel positions on the high resolution grid, new sampling points are obtained as weighted average of combined warped neighboring information and the interpolated image $\hat{c}_{\text{int}}(m,n)$. The larger t is chosen, the more weight is assigned to the warped information from the neighboring reference cameras. Setting $t = 1$ results in FSE-DIBR.

However, potentially unreliable depth information might especially occur at depth boundaries. Therefore, the above described weighted averaging is only done for pixel positions which have at least a distance of p pixels to the nearest depth edge.

After incorporating additional information from the reference camera perspectives, the FSE algorithm is applied on the new extended set of sampling points $c'(m,n)$, resulting in the final reconstruction result $\hat{c}_{\text{mv}}(m,n)$.

4. SIMULATION RESULTS

The proposed reconstruction scheme has been tested for the multiview datasets *art*, *books*, *dolls*, and *moebius* [9]. For each dataset, the views indexed by 1, 3, and 5 have been chosen. The sampling factor s has been varied in the range of $\frac{1}{4}$ to $\frac{1}{16}$. The irregularly shielded low-resolution images were created by multiplying the original high-resolution images with corresponding sampling masks. For the FSE, we have used 100 iterations and a blocksize of 4. The extrapolation area has been 28 samples wide. The weighting function has declined with $\hat{\rho} = 0.7$. Already reconstructed samples have been additionally weighted by a factor $\delta = 0.8$. For the basis functions, the FFT size has been set to 32×32. The weighting factor t has been set to 0.6, 0.7, and 0.8 for $s = \frac{1}{4}$, $s = \frac{1}{9}$, and $s = \frac{1}{16}$, respectively. Since the interpolation quality

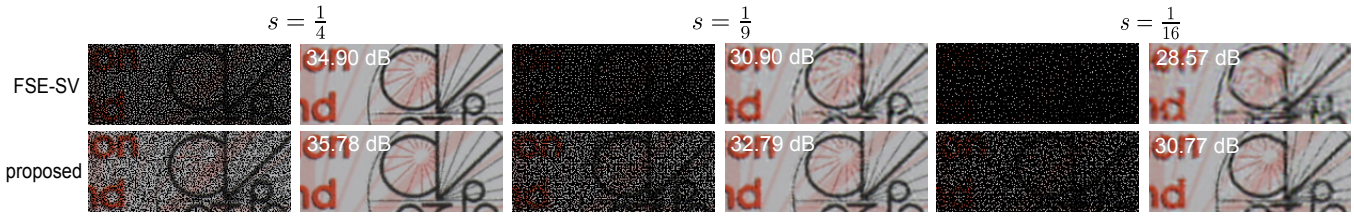

Fig. 5. Visual comparison: From left to right: $s = \{\frac{1}{4}, \frac{1}{9}, \frac{1}{16}\}$, top row: non-regular input and result of singleview reconstruction (FSE-SV), bottom row: non-regular input and result of proposed multiview reconstruction

also depends on the number of available sampling points, the information from adjacent views is weighted higher, the less information from the central view is available.

First, the approach has been tested with full-resolution (FR) depth maps. However, high-resolution depth images are typically obtained via stereo matching approaches which assume images to be completely known. Thus, the method has been also tested with simulated Time-of-flight (ToF) depth data. Since ToF cameras only provide a low spatial resolution, the originally high-resolution depth maps have been downsampled by a factor of 8 in both spatial dimensions. In order to remain sharp edges, the downsampled depth maps have been resized using nearest neighbor interpolation. The depth distance parameter p has been set to 2 for FR depth maps and to 5 for the ToF depth images.

Table 1 gives the PSNR results for all considered datasets, various sampling factors s and different depth resolutions. Thereby, our proposed multiview reconstruction approach has been compared with both, the singleview reconstruction (FSE-SV) [6] and the FSE-DIBR approach. PSNR values are calculated with respect to the desired high-resolution central image. For all simulations, the proposed reconstruction algorithm outperforms the FSE-SV approach and the FSE-DIBR method. Compared to FSE-SV, average gains of 0.85 dB, 1.62 dB, and 1.82 dB have been achieved for the various sampling factors when using ToF depth maps. The maximum gain is 2.20 dB and has been achieved for the *books* dataset and $s = \frac{1}{16}$. Compared to FSE-SV, larger gains can be obtained for larger sampling factors. It has to be emphasized that for $s = \frac{1}{16}$, 93.75% of the desired high-resolution image are unknown and thus have to be reconstructed. If FR depth maps can be assumed, even higher PSNR gains can be achieved, since depth information near edges is more reliable than for ToF depth data. This leads to a maximum gain of 2.44 dB for *books* and $s = \frac{1}{16}$.

The final visual quality can be observed in Fig. 5 for a detail of the *books* dataset. The images correspond to the case of using ToF depth data for the synthesis process. The figure shows from left to right both, the available sampling points and the reconstruction results for all considered sampling factors. Thereby, the first row depicts the singleview reconstruction, while the bottom row shows the proposed multiview reconstruction approach. It can be seen that especially in case of larger sampling factors, the incorporation of neighboring views is highly important in order to obtain additional sampling points and thus to reliably reconstruct sharp edges and image details.

Table 1. PSNR evaluation in dB for all considered data sets, various sampling factors s, and different depth resolutions

		FSE-SV	FSE-DIBR	pro-posed	FSE-DIBR	pro-posed
			FR depth		ToF depth	
$s = \frac{1}{4}$	*art*	37.03	37.14	**37.82**	36.96	**37.61**
	books	34.90	35.38	**35.89**	35.18	**35.78**
	dolls	35.23	35.70	**36.35**	35.56	**36.18**
	moebius	36.18	36.76	**37.38**	36.55	**37.17**
	avg. Δ	-	0.41	**1.03**	0.23	**0.85**
$s = \frac{1}{9}$	*art*	32.83	34.24	**34.51**	33.88	**34.14**
	books	30.90	32.82	**32.99**	32.54	**32.79**
	dolls	31.14	33.03	**33.30**	32.70	**32.96**
	moebius	32.52	34.14	**34.46**	33.65	**33.97**
	avg. Δ	-	1.71	**1.97**	1.35	**1.62**
$s = \frac{1}{16}$	*art*	30.28	32.08	**32.25**	31.63	**31.79**
	books	28.57	30.86	**31.01**	30.61	**30.77**
	dolls	28.72	30.92	**31.12**	30.53	**30.73**
	moebius	30.30	32.23	**32.46**	31.60	**31.82**
	avg. Δ	-	2.07	**2.26**	1.64	**1.82**

As already mentioned, in order to restrict complexity, fast bilinear interpolation has been used for generating the interpolation result $\hat{c}_{\text{int}}(m, n)$. The final reconstruction quality might be further improved if a more powerful but probably also more complex interpolation technique is used instead.

5. CONCLUSION

In this paper, we proposed a novel reconstruction scheme for multiview images taken with non-regular sampling sensors. The approach avoids the need of high-resolution sensors and thus provides an opportunity to reduce the costs in multiview scenarios. The incorporation of neighboring views leads to a significant gain, both in PSNR and visual quality. Considering ToF depth data, our framework achieves a maximum PSNR gain of 2.20 dB for a sampling factor of $s = \frac{1}{16}$ compared to a state-of-the-art singleview reconstruction approach.

6. REFERENCES

[1] 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.

[2] D.C. Garcia, C. Dórea, and R. de Queiroz, "Super-Resolution for Multiview Images using Depth Information," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Hong Kong, China, Sep. 2010, pp. 1793–1796.

[3] 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.

[4] P. Sen and S. Darabi, "A novel Framework for Imaging using Compressed Sensing," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2133–2136.

[5] M. Trocan, T. Maugey, J.E. Fowler, and B. Pesquet-Popescu, "Disparity-Compensated Compressed-Sensing Reconstruction for Multiview Images," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, Singapore, July 2010, pp. 1225–1229.

[6] M. Schöberl, J. Seiler, S. Fössel, and A. Kaup, "Increasing Imaging Resolution by Covering your Sensor," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, Brussels, Belgium, Sep. 2011, pp. 1897–1900.

[7] 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.

[8] J. Seiler and A. Kaup, "Optimized and Parallelized Processing Order for Improved Frequency Selective Extrapolation," in *Proc. IEEE European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, Aug. 2011, pp. 269–273.

[9] 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.