Super-Resolution Stereo- and Multi-View Synthesis from MonocularVideo Sequences

Sebastian Knorr, Matthias Kunter, and Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Einsteinufer 17, Berlin, Germany
E-mail: {knorr, kunter, sikora}@nue.tu-berlin.de

ABSTRACT

This paper presents a new approach for generation of super-resolution stereoscopic and multi-view video from monocular video. Such multi-view video is used for instance with multi-user 3D displays or auto-stereoscopic displays with head-tracking to create a depth impression of the observed scenery. Our approach is an extension of the realistic stereo view synthesis (RSVS) approach which is based on structure from motion techniques and image-based rendering to generate the desired stereoscopic views for each point in time. The extension relies on an additional super-resolution mode which utilizes a number of frames of the original video sequence to generate a virtual stereo frame with higher resolution. The algorithm is tested on several TV broadcast videos, as well as on sequences captured with a single handheld camera and sequences from the well known BBC documentation "Planet Earth". Finally, some simulation results will show that RSVS is quite suitable for super-resolution 2D-3D conversion.

1. INTRODUCTION

Extending visual communication to the third dimension by providing the user with a realistic depth perception of the observed scenery instead of flat 2D images has been investigated over decades. 3DTV is in the focus of many researchers worldwide. Recent progress in related research areas may enable various 3D applications and systems in the near future [1]. Especially the innovations regarding the 3D display technology are tremendous. 3D displays are entering professional and consumer markets. However, the film industry still adheres to traditional capture techniques with a single camera, i.e. the conversion of existing 2D content into super-resolution 3D is highly interesting for instance for content owners. Movies may be reissued in 3D in the future

Figure 1: Example of an auto-stereoscopic display

(see Figure 1), e.g. the *Star Wars* episodes are currently being converted entirely into 3D.

Many fundamental algorithms have been developed to extract the 3D information from monocular video sequences during the last years [2]-[11]. Some of them are dealing with the reconstruction of complete 3D models from the captured scenery [2]-[5]. Others just intend to render stereoscopic views either by estimating planar transformations [6] or via dense depth maps for each frame of the sequence using *depth-image-based rendering* (DIBR) [7]-[11]. Available *structure from motion* (SfM) techniques from the first category estimate the camera parameters and sparse 3D structure quite well, but they fail to provide dense and accurate 3D modeling as it is necessary to render high quality stereoscopic views for consumer markets. On the other hand, dense depth estimation as necessary for DIBR is still an error prone task and computationally very expensive to get time consistent depth maps for each frame of the sequence.

In this paper, we present a new approach for generation of super-resolution stereo and multi-view video from monocular video based on RSVS [12]. It combines both the powerful algorithms of SfM [2] and *image-based*

rendering (IBR) [12] without relying on dense depth estimation.

Most available 3D display systems need 2 views corresponding to the human eye distance to create a depth perception, which is also known as stereo video. However, more advanced systems use multiple views (e.g. 8 views that display the same scene from different viewpoints). The presented algorithm is capable to generate stereo video in its basic mode, but it is also capable to generate multi-view video. We will show that the approach is quite suitable for converting existing 2D video material into multi-view with higher resolution. To our knowledge it is the first time that an approach for generation of super-resolution multi-view video from monocular video is reported.

First, sparse 3D structure and camera parameters are estimated with SfM for the monocular video sequence (grey cameras in Figure 2). Then, for each original camera position (blue in Figure 2) a corresponding multi-view set is generated (red in Figure 2). This is done by estimating planar transformations (homographies) to temporal neighboring views of the original camera path. Surrounding original views are utilized to generate the multiple virtual views with IBR. Hence, the computational expensive calculation of dense depth maps is avoided. Moreover, the occlusion problem is almost nonexistent. Whereas DIBR techniques always have to inter- or extrapolate disclosed parts of the images when shifting pixels according to their depth values, our approach utilizes the information from close views of the original camera path, i.e. occluded regions become visible within the sequence.

In the extended mode, the so called super-resolution mode, the temporal neighboring views are utilized for reconstructing a virtual stereo frame with the desired resolution. That means, each pixel in the super-resolution stereo frame should be located as close to the pixel raster in one of the neighboring views as possible for pixel warping, i.e. the effect of low pass filtering caused by bilinear warping is reduced. Another benefit of this approach is, as will be shown in Section 4, that motion blur and coding artifacts can be reduced.

The organization of this paper is as follows: The next section describes the theoretical background of the RSVS approach. The super-resolution stereo- and multi-view synthesis is detailed in Section 3. In Section 4 simulation results are presented. The limitations of our approach are stated in Section 5. Finally, in Section 6, the paper concludes with a summary and a discussion.

2. BACKROUND

2.1 Camera Calibration and Sparse 3D Structure Estimation Using Structure from Motion

The general intention of SfM is the estimation of the external and internal camera parameters and the structure of a 3D scene relative to a reference coordinate system. SfM requires a relative movement between a static 3-D scene and the camera.

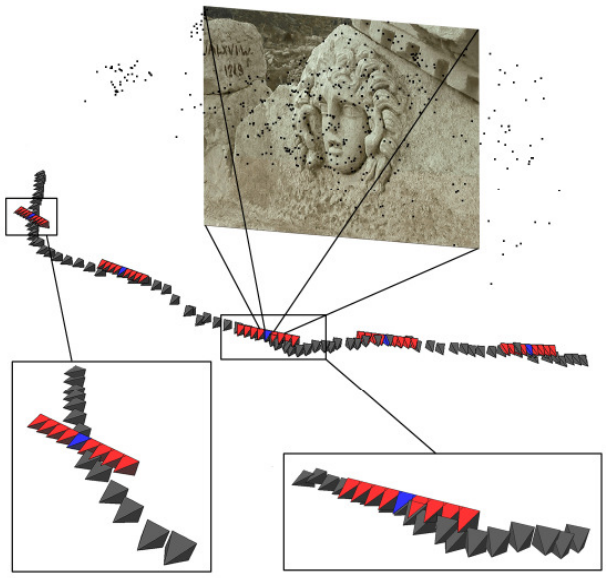

Figure 2: Multi-view synthesis using SfM and IBR; gray: original camera path, red: virtual stereo cameras, blue: original camera of a multi-view camera setup

An initial step in the reconstruction process is to find relations between the views in the video sequence. This geometric relationship, also known as epipolar geometry, can be estimated with a sufficient number of feature correspondences between the views [14]. Once the images are related, the camera projection matrices can be calculated using singular value decomposition [15]. If feature correspondences between the views and projection matrices are known, sparse 3D scene structure can be estimated with triangulation [16], i.e. for a limited number of points the 3D coordinates are available as illustrated in Figure 2. For a final refinement of the estimated parameters, bundle adjustment is often used [17].

Since the performance of the reconstruction is heavily dependent on the initial structure computation, Imre et al. [18] introduced a prioritized sequential 3D reconstruction approach for a fast and reliable structure and motion computation.

If the internal calibration parameters are unknown, which is in general the case for movies and videos captured with a handheld camera, a self-calibration procedure has to be carried out. This can either be done via constraints on the essential matrices as introduced in [19], or by using a stratified approach from projective to metric reconstruction as described in [14] and [15].

2.2 Stereo-/Multi-view Synthesis using IBR

Once 3D structure and camera path are determined, multiple virtual cameras can be defined for each frame of the original

video sequence as depicted in Figure 2. A blue camera corresponds to an original image of a video sequence and the red cameras represent the multiple virtual stereoscopic partners. With the principles of IBR [12] pixel values from temporal neighboring views (grey cameras in Figure 2) can be projected to their corresponding positions in the virtual views. Thus, each of the virtual images is just a rendered version of original images. IBR requires establishment of homographies H between original and virtual views and is done as follows (see Figure 3).

The external parameters of the virtual cameras are defined by the desired multi-view setup. In case of a parallel setup, the rotation matrices of all multiple virtual views are identical to the rotation matrix of the corresponding original view, which is estimated by SfM as described before. The internal parameters are set to be identical as well. Just the translation vector of each virtual view differs with respect to the world coordinate system and the virtual camera distance (see section 2.3 for details on calculation of translation).

Then, the 3D points M obtained by SfM can be projected into each virtual view as depicted in Figure 2 resulting in image coordinates m_{multi}:

$$m_{multi} = P_{multi} M \,, \tag{1}$$

with $P_{multi} = KR\left[I \,\middle|\, -\tilde{C}_{multi} \right]$. K is the internal calibration matrix, R is the rotation matrix, I is a 3x3 identity matrix and $\tilde{C}_{multi}$ is the position of the camera center in inhomogeneous coordinates (see section 2.3).

Corresponding 2D points of original images m_i and virtual images m_{multi} are approximately related through the homography H between both views, if the distance (baseline) between the virtual camera and the original camera is small:

$$m_i = H_i m_{multi} \,. \tag{2}$$

H is a 3x3 matrix and therefore it contains 9 entries, but is defined only up to scale. Correspondences are available from the estimated sparse 3D structure, meaning that for a number of 3D points M the corresponding image positions m_i and m_{multi} are known directly from eq. 1. Thus H can be estimated from eq. 2 with a minimum number of four point correspondences. In Hartley and Zisserman [15] many robust and non-linear alternatives with more than four point correspondences are introduced.

Once the homography between a virtual view to be generated and the closest original view (see section 2.3) of the video sequence is estimated, all pixel values of the original image can be projected to their corresponding locations in the virtual image using eq. 2. Since these positions do not exactly correspond with the pixel raster, bilinear interpolation is performed on the pixel values.

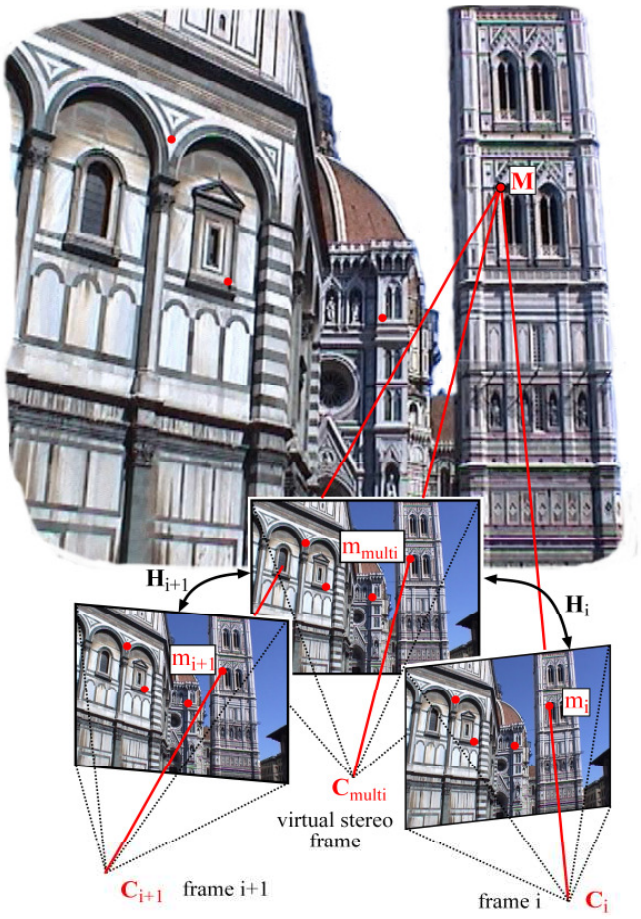

Figure 3: Stereo-/multi-view synthesis using IBR

In general, the closest original view does not cover the whole scene that should be visible with the virtual stereo camera as depicted in Figure 4b). This is particularly the case when the orientation of both cameras differs significantly. To fill the missing parts of the virtual stereo image, additional surrounding views (e.g. frame i+1 in Figure 3) have to be taken into account (see Figure 4c) and 4d)).

2.3 Determine positions of the virtual views

The virtual parallel camera setup requires definition of the horizontal distance between the views, the so-called *screen parallax* values. Since the estimated camera path and 3D structure are only defined up to scale, it is not clear at this stage if the camera is close to a small 3D model or far away from a huge 3D scenery. The average human eye distance is known with approximately 64 mm, and the virtual views shall have the same distance from each other. Therefore the process requires some initial user interaction. The first frame of the sequence can be used to define the distance t_s between the camera and the dominant scene in meters. Without loss of generality, the world coordinate system is located in the

Figure 4: Padding of pixels with additional frames: a) original left view of the sequence "Dome", b) virtual right view, only rendered with the closest view of the camera path, c) virtual right view using 30 and d) 62 frames of the original sequence.

Figure 5: Multi-view synthesis of the "Statue" sequence. Middle: original view, left: virtual left views (t_x = -64, -128, -192, and -256 mm), right: virtual right views (t_x = 64, 128, 192, and 256 mm)

centroid of the sparse 3D point cloud. Thus, the absolute position of all cameras regarding the world coordinate system can be determined with

$$C_i^m = t_s \frac{C_i}{\|C_1\|}, \tag{3}$$

where $\|C_1\|$ is the vector norm of the first camera.
The position of each corresponding virtual camera is

$$C_{i,multi}^m = C_i^m + R_i^{-1} \cdot \begin{bmatrix} \pm t_x \\ 0 \\ 0 \end{bmatrix}, \tag{4}$$

and the camera projection matrix

$$P_{i,multi}^m = KR\left[I \,\middle|\, -\tilde{C}_{i,multi}^m \right]. \tag{5}$$

t_x is the average human eye distance of 64 mm when synthesizing just one stereoscopic view. In the multi-view camera setup, t_x is a multiple of 64 mm depending on the number of desired virtual views (e.g. see Figure 5 for the case of 8 virtual views). With t_x fixed, the screen parallax can be changed indirectly by setting t_s, i.e. decreasing t_s increases the screen parallax.

Once, the positions of the virtual cameras are defined, the closest original views need to be determined to employ IBR. Therefore, the Euclidean distances between each virtual camera and all original cameras are calculated and sorted in ascending order.

Figure 5 shows 8 virtual views of the handheld sequence "Statue" generated with the proposed solution and its corresponding original view in the middle.

3. SUPER-RESOLUTION STEREO-/MULTI-VIEW SYNTHESIS

The previous section described the fundamental RSVS approach to convert a monocular video sequence into a stereo- or multi-view sequence for auto-stereoscopic displays or multi-user 3D displays. Figure 4 demonstrates that in general more than one view is needed to set up a virtual stereo frame. Thus, the additional views can be used to increase the resolution of the stereo frame as well. Spatial image super-resolution is a very intensively studied topic because it improves the inherent resolution limitation of captured *low resolution* images (LR images) [20],[21]. The main objective is to construct one ore more *high resolution* (HR) images by processing several LR images, captured by different cameras or in our case at different points in time. This can be achieved by estimating the inverse of the observation model which relates LR images to HR images [20].

3.1 Bilinear Warping

Depending on the desired resolution, a virtual super-resolution stereo frame for each original frame has to be set up. Without loss of generality we increase the resolution of the original video sequence with factor 1.5, i.e. an input video in PAL format (720x576 pixel) results in a 1080x864 pixel stereo output video.

For each pixel in a stereo frame we determine the position in surrounding views as described in section 2.2. The pixel which lies closest to the pixel raster has the best properties for bilinear warping, since the low pass characteristics, which is always present during bilinear warping, can be reduced. In Figure 6 an example of this process is given. Let's say frame *i* is the closest original view to the virtual stereo view, the calculated pixel position is quite far from the quantized pixel raster, i.e. bilinear interpolation would increase the low pass effect. In frame *i*+1 the pixel lies almost directly on the pixel raster. Hence, the pixel value is quite more suitable for warping because of low pass effect reduction.

3.2 Smoothness Constraint for Pixel Warping

The previous subsection indicated that the pixel closest to the pixel raster in one of the surrounding views is most suitable for pixel warping. This is not always true if the pixel belongs to a view far from the virtual stereo view, because the planar transformation errors increase with the baseline length between the views. To avoid this, we consider a smoothness constraint for pixel warping.

First, we calculate the pixel values in all desired views (e.g. 8 closest views) with bilinear interpolation. Then we determine the median of this pixel values with

$$I_{med}(x, y) = \underset{\forall i}{\textbf{median}}\ I_i(x, y), \tag{6}$$

where *I* is the color value of the pixel in each frame *i*. Pixel values with an absolute deviation from the median higher than a predefined theshold are removed and not considered in further processing steps. Finally, for the remaining pixels, we take the one which lies closest to the pixel raster for bilinear warping..

4. SIMULATION RESULTS

In subsection 4.1 we present some results of our RSVS approach tested on five TV broadcast videos, as well as on five sequences captured with a single handheld camera. Then, in subsection 4.2, we show some preliminary results of stereo images converted from the well known BBC documentation "Planet Earth" and, finally, the super-resolution mode is tested on two sequences captured with a handheld camera (subsection 4.3). A parallel camera setup was used for all sequences to generate the desired stereo- and multi-views. The simulation results show the remarkable performance of the conversion process.

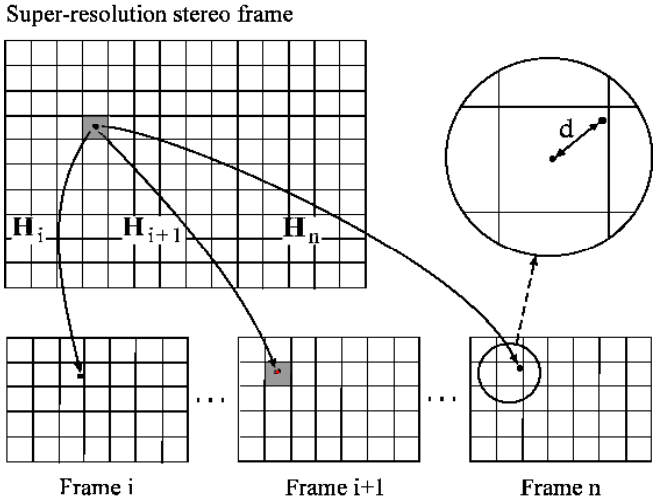

Figure 6: Super-resolution stereo-/multi-view synthesis

4.1 Stereo-/Multi-view Synthesis using Standard RSVS

In Figure 4, the stereo view synthesis for an original left view of the "Dome"-sequence generated with 62 frames of the original sequence is already presented. Figure 5 shows 8 virtual views of the handheld sequence "Statue" generated with the proposed solution and its corresponding original view in the middle.

Table 1: Distance settings and average number of used frames for rendering a stereo frame

Sequence	resolution in pixel x pixel	distance settings t_s in meters	avg. # of surround. views per stereo view
TV broadcast			
Pyramid	720 x 405	8	2.13
Vase	720 x 405	5	1.69
Cliff	720 x 576	10	3.04
Wall	720 x 576	10	10.50
Canyon	720 x 576	10	14.38
handheld			
Medusa	448 x 358	3	14.04
Dome	720 x 576	8	61.24
Facade	720 x 576	6	37.11
Statue	720 x 576	8	53.70
Church	576 x 720	6	8.14

In Table 1, the distance settings and the average number of used frames to render each stereo view of a sequence is presented for all test data sets. Due to a linear camera movement with almost no camera rotations, the average number of frames needed to render the stereo views is quite low for the TV broadcast sequences "Pyramid" and "Vase", i.e. 2.13 and 1.69, respectively. For the handheld sequences "Medusa", "Dome", Statue" and "Facade", this number is relative high because of unsteady camera movement and camera rotations.

4.2 Stereo View Synthesis from BBC documentation "Planet Earth"

Figures 7 and 8 show some results for sequences from the BBC documentation "Planet Earth" converted with RSVS. In Figure 7 a red-cyan stereo image pair from the series "Ice Worlds" is presented. Figure 8 gives an example stereo image pair from the series "Jungles", respectively.

4.3 Stereo-/Multi-view Synthesis using Super-resolution RSVS

Two example figures show the performance of the super-resolution mode of our approach. In Figure 9, an up-sampled virtual stereo frame using Lanczos-filtering and a super-resolution virtual stereo frame (each of size 1080x864 pixels) of the "Dome"-sequence (original frame size see Table 1) are presented. Four close-ups should stress the differences between both frames. Figure 9c shows some typical artifacts when dealing with interlaced PAL video and up-sampling: Sawtooth pattern can be noticed along edges resulting from de-interlacing. Additionally, aliasing artifacts become more visible in the up-sampled frame, which can be seen on the top of the right arc in Figure 9c and more clearly in Figure 9e. In the super-resolution case, these artifacts are strongly reduced.

In Figure 10, just a super-resolution frame of the "Statue"-sequence (original frame size see Table 1) is presented with the same settings. The reduction of the previous mentioned artifacts is also visible in the super-resolution frame. Furthermore, it can be seen that super-resolution has two more advantages than just up-sampling the virtual stereo frame: Ghosting effects resulting from the compression and motion blur caused by very unsteady camera movements are strongly reduced in the super-resolution case as well (see close-ups in Figure 10).

5. LIMITATIONS

Nevertheless, this approach has some limitations. The most important one is that the scene has to be static, i.e. moving objects within the scene would disturb the depth perception. Furthermore, there are restrictions on camera movement. If the camera moves only in a forward- or backward direction, this approach for virtual view synthesis fails. The case of a camera movement in up- and down direction can be handled by transposing the frames by 90 degrees. A final limitation is that a larger screen parallax increases the divergence between the camera path and the position of the virtual views as depicted in Figure 2 on the bottom left. Hence, a planar transformation might not be valid any longer. To overcome this problem, a reduction of the stereo effect in such parts of the sequence should be carried out, i.e. the baseline between stereoscopic views must be decreased smoothly.

Figure 7: Anaglyph stereo image pair of the series "Ice Worlds" [Source: BBC documentation "Planet Earth"]

Figure 8: Anaglyph stereo image pair of the series "Jungles" [Source: BBC documentation "Planet Earth"]

6. SUMMARY AND CONCLUSIONS

This paper presented a new approach for generation of super-resolution stereo and multi-view video from monocular video, i.e. we extended our previous work on RSVS with a super-resolution mode. To our knowledge it was the first time that generation of super-resolution multi-view video from monocular video was addressed. Thus, the algorithm is suitable for offline content creation for conventional and advanced 3D display systems with minimum user assistance.

The main advantage of this approach over available DIBR algorithms is that planar transformations are utilized

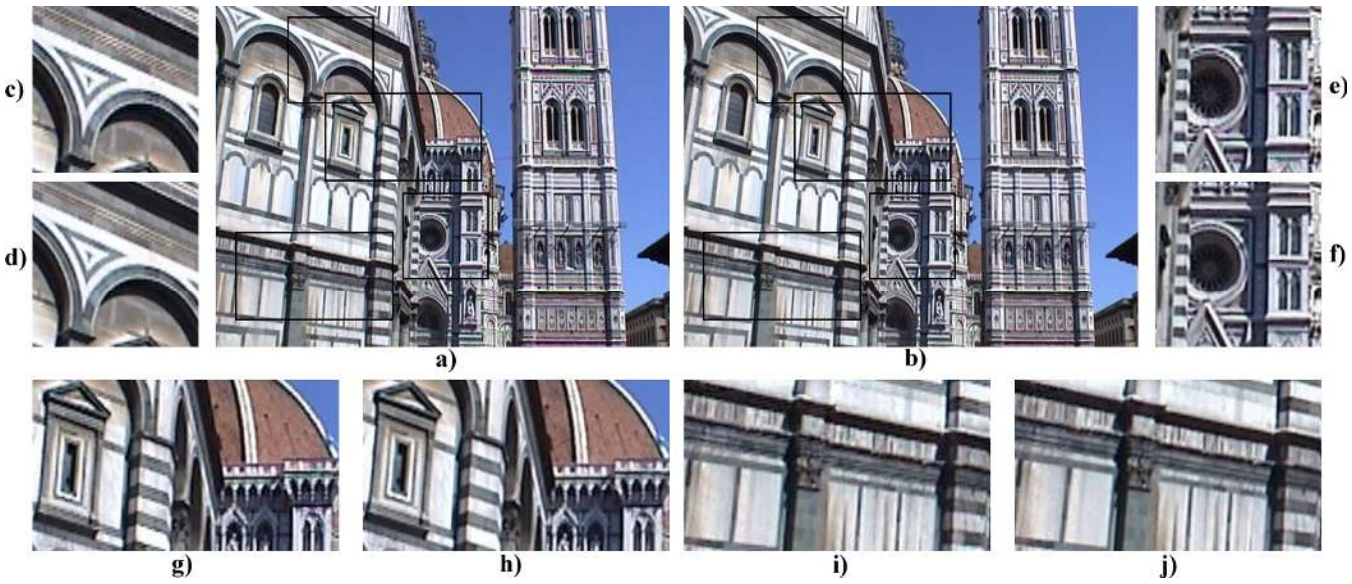

Figure 9: Super-resolution stereo view synthesis of the "Dome"-sequence. a) virtual stereo view up-sampled and b) super-resolution stereo view. c), e), g), i) close-up of the up-sampled and d), f), h), j) close-up of the super-resolution frame.

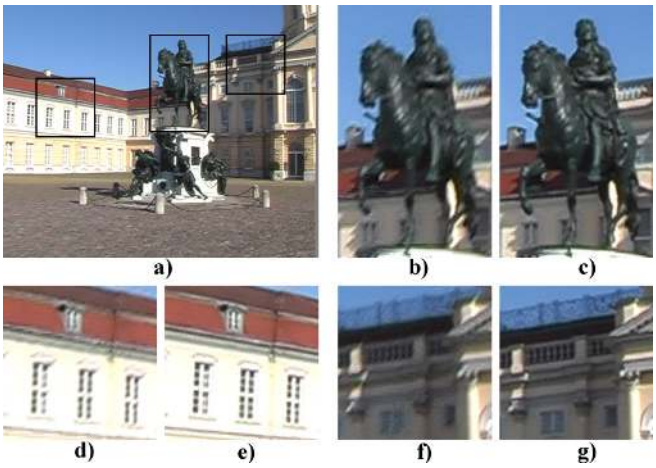

Figure 10: Super-resolution stereo view synthesis of the "Statue"-sequence. a) super-resolution stereo view. b), d), f) close-up of the up-sampled and c), e), g) close-up of the super-resolution stereo frame.

to generate the virtual views from original views, i.e. a computational expensive and error prone dense depth estimation is not needed. Furthermore, the occlusion problem, which is always present in dense depth estimation, does almost not exist. Another advantage is that photo realism is achieved without additional operations, since the photometric properties of a scene are determined entirely by the original frames of the reference sequence. Particularly, super-resolution photo realism with DIBR is not achievable.

The RSVS algorithm was tested first on five TV broadcast sequences and five sequences captured with a single handheld camera. Then, we demonstrated the performance of RSVS on several sequences of the BBC documentation "Planet Earth". Finally, super-resolution RSVS was carried out on two of the test-sequences ("Dome" and "Statue") with a remarkable performance. Here, de-interlacing-, aliasing-, ghosting- and blurring artifacts were reduced significantly.

Despite the restrictions mentioned in the previous section, the presented algorithm is highly attractive as a tool for user-assisted 2D-3D conversion and 3D production systems. High quality conversion and production is still done using semi-automatic software systems. Here, the presented algorithm may help reducing the manual workload.

Future works will focus on subjective quality tests to compare the results with DIBR and with stereoscopic sequences captured with a stereo camera rig.

7. ACKNOWLEDGEMENT

The work presented was developed within 3DTV, a European Network of Excellence (http://www.3dtv-research.org), funded under the European Commission IST FP6 programme.

8. REFERENCES

[1] O. Schreer, P. Kauff, and T. Sikora (Eds.), "3D videocommunication: algorithms, concepts and real-time systems in human centered communication", John Wiley & Sons Ltd, Chichester, England, 2005

[2] T. Jebara, A. Azarbayejani, and A. Pentland, “3D structure from 2D motion”, IEEE Signal Processing Magazine, May 1999, Vol. 16. No. 3, p. 66-84

[3] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool, “Automated reconstruction of 3D scenes from sequences of images”, ISPRS Journal of Photogrammetry and Remote Sensing (55) 4, pp. 251-267, 2000

[4] C. Tomasi, and T. Kanade, ”Shape and motion from Image Streams: A Factorization Method”, Journal of Computer Vision 9(2), pp. 137-154, 1992

[5] S. Knorr, E. Imre, B. Özkalayci, A. A. Alatan, and T. Sikora, “A modular scheme for 2D/3D conversion of TV broadcast” 3rd Int. Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT), Chapel Hill, USA, 2006

[6] E. Rotem, K. Wolowelsky, and D. Pelz, “Automatic video to stereoscopic video conversion”, Proc. of the SPIE: Stereoscopic Displays and Virtual Reality Systems XII, Vol. 5664, pp. 198-206, March 2005

[7] S. Curti, D. Sirtori, and F. Vella, “3D effect generation from monocular view”, Proc. of the Int. Symposium on 3D Data Processing Visualization and Transmission (3DPVT), Padova, Italy, 2002

[8] K. Moustakas, D. Tzovaras, and M .G. Strintzis, "Stereoscopic video generation based on efficient structure and motion estimation from a monoscopic image sequence", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 15, No. 8, pp. 1065 - 1073, August 2005.

[9] K. T. Kim, M. Siegel, and J. Y. Son, “Synthesis of a high-resolution 3D stereoscopic image pair from a high-resolution monoscopic image and a low-resolution depth map“, Proc. of the SPIE: Stereoscopic Displays and Applications IX, San José, USA, 1998

[10] C. Fehn, “Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV”, Proc. of the SPIE: Stereoscopic Displays and Virtual Reality Systems XI, San José, USA, 2004

[11] L. Zhang, J. Tam, and D. Wang, “Stereoscopic image generation based on depth images”, IEEE Int. Conf. on Image Processing (ICIP), Singapore, 2004

[12] S. Knorr, and T. Sikora, “An Image-based Rendering (IBR) Approach for Realistic Stereo View Synthesis of TV Broadcast based on Structure from Motion”, IEEE Int. Conf. on Image Processing (ICIP), San Antonio, USA, 2007 (to be published).

[13] L. MacMillan, “An image based approach to three-dimensional computer graphics”, Ph.D dissertation, University of North Carolina, 1997

[14] M. Pollefeys, “Tutorial on 3D modeling from images”, European Conf. on Computer Vision (ECCV), 2000

[15] R. Hartley, and A. Zisserman, “Multiple view geometry”, Cambridge University Press, UK, 2003

[16] R. Hartley, and P. Sturm, “Triangulation”, Computer Vision and Image Understanding, 68(2): 146-157, 1997

[17] B. Triggs, and P. McLauchlan, R. Hartley, A. Fitzgibbon, “Bundle adjustment - a modern synthesis”, in “Vision Algorithms: Theory & Practice”, Springer-Verlag, 2000

[18] E. Imre, S. Knorr, A. A. Alatan, and T. Sikora “Prioritized sequential 3D reconstruction in video sequences of dynamic scenes”, IEEE Int. Conf. on Image Processing (ICIP), Atlanta, USA, 2006.

[19] P. R. S. Mendonca and R. Cipolla, “A simple technique for self-calibration”, IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), 1999

[20] C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: a technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, May 2003, pp. 21-36.

[21] S. Borman and R. L. Stevenson “Super-resolution from image sequences - a review,” Midwest Symposium on Systems and Circuits, pp. 374-378, 1998.