From 2D- to Stereo- to Multi-view Video

Sebastian Knorr, Aljoscha Smolic[†], and Thomas Sikora

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

[†] Fraunhofer Institute for Telecommunications/
Heinrich-Hertz-Institut
Einsteinufer 37, Berlin, Germany
E-mail: smolic@hhi.de

ABSTRACT

This paper presents a new approach for generation of 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. The intention of this work is not a real-time conversion of existing video material with a deduction in stereo perception, but rather a more realistic off-line conversion with high accuracy. Our approach is based on structure from motion techniques and uses image-based rendering to generate the desired multiple views for each point in time. The algorithm is tested on several TV broadcast videos, as well as on sequences captured with a single handheld camera. Finally, some simulation results will show the remarkable performance of this approach.

***Index Terms*–** Stereo Vision, Image motion analysis, Rendering, Three-dimensional displays

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. Recent progress in related research areas may enable various 3D applications and systems in the near future [1]. Especially, 3D display technology is maturating. 3D displays are entering professional and consumer markets. Often the content is created directly in some suitable 3D format. On the other hand the conversion of existing 2D content is highly interesting for instance for content owners. Movies may be reissued in 3D in the future.

Therefore such 2D-3D conversion is of high interest, and many fundamental algorithms have been developed to reconstruct 3D scenes from uncalibrated video sequences [2]-[10]. These algorithms can roughly be divided into two categories: methods that tend to get a complete 3D model of the captured scene [2]-[5], and methods that just 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]-[10]. 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 views. On the other hand, dense depth estimation as necessary for DIBR is still an error prone task and computationally very expensive.

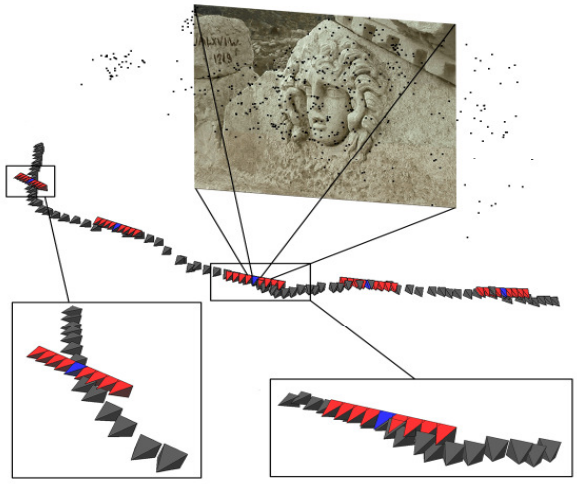

Figure 1: Multi-view synthesis using SfM and IBR; gray: original camera path, red: virtual stereo cameras, blue: original camera of a multi-view camera setup

In this paper, we present a new approach for generation of stereo and multi-view video from monocular video that combines both the powerful algorithms of SfM and *image-based rendering* (IBR) [11] without relying on depth estimation. Most available 3D display systems rely on 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 showing 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. To our knowledge it is the first time that an approach for generation of multi-view video from monoscopic video is reported.

First, sparse 3D structure and camera parameters are estimated with SfM for the monoscopic video sequence (grey cameras in Figure 1). Then, for each original camera position (blue in Figure 1) a corresponding multi-view set is generated (red in Figure 1). This is done by estimating planar transformations (homographies) to temporal neighboring views of the original camera path. Surrounding original views are used to generate the multiple virtual views with IBR. Hence, the computational expensive calculation of dense depth maps is avoided.

Another benefit of this approach is the handling of occlusions. 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. However, our approach has also some limitations, which will be discussed at the end of this paper.

2. IMPLEMENTATION

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 scene and the camera, which is a limitation of our algorithm.

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 [3]. Once the images are related, the camera projection matrices can be calculated using singular value decomposition [2]. If feature correspondences between the views and projection matrices are known, sparse 3D scene structure can be estimated with triangulation [2], i.e. for a limited number of points the 3D coordinates are available as illustrated in Figure 1. For a final refinement of the estimated parameters, non-linear minimization can be applied.

2.2 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 1. A blue camera corresponds to an original image of a video sequence and the red cameras represent its corresponding multiple virtual views. With the principles of IBR [11] pixel values from temporal neighboring views can be projected to their

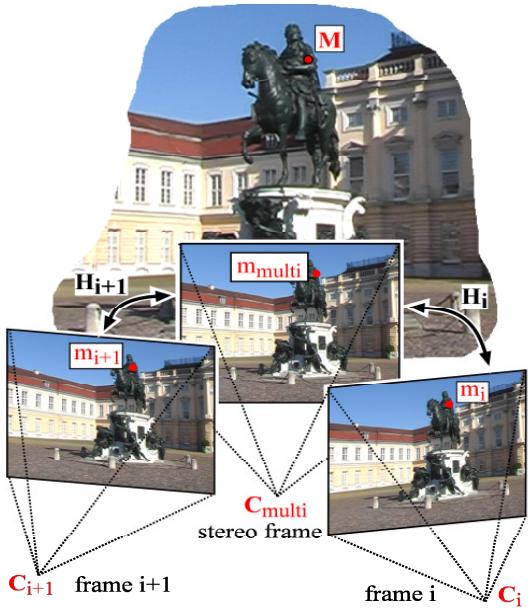

Figure 2: Stereo-/multi-view synthesis using IBR

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 2).

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 homogeneous coordinates (see section 2.3).

Corresponding 2D points of original images m_i and virtual images m_{multi} are 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

Figure 3: 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)

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, the first directly from SfM and the second by calculation via eq. 1. Thus *H* can be estimated from eq. 2 with a minimum number of four point correspondences. In Hartley and Zisserman [2] many robust and non-linear alternatives 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.

In general, the closest original view does not cover the whole scene that should be visible in the virtual view. This is particularly the case when the orientation of both cameras differs significantly. To fill the missing parts of the virtual image, additional temporal neighboring views have to be taken into account.

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 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. 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

$$C_{i,multi}^m = C_i^m + R_i^{-1} \cdot \begin{bmatrix} \pm t_x \\ 0 \\ 0 \end{bmatrix}, \quad \text{with } t_x = \frac{n}{2} \cdot 64mm \tag{4}$$

(n=2, 4, 6, …, N) and the camera projection matrix

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

N is the number of virtual views that should be generated for each frame of the sequence (e.g. N is set to 8 in Figure 3). 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.

3. SIMULATION RESULTS

The algorithm is tested on five TV broadcast videos, as well as on five sequences captured with a single handheld camera. A parallel camera setup was used for all sequences to generate the multiple stereo views. Figure 3 shows 8 virtual views of the handheld sequence "Statue" generated with the proposed solution and its corresponding original view in the middle. The resolution was reduced from 720x576 pixel to 640x512 pixel because the virtual views couldn't be filled completely with pixel values from surrounding views for increasing t_x.

Two anaglyph stereo-images ("Statue" and "Dome") are presented in Figure 4. The distance t_s between the camera and the dominant scene was set to 10 meters and 8 meters, respectively (t_x = 64 mm). Similar results were obtained with the TV broadcast sequences as well as with two of the remaining three handheld sequences. Only the conversion of the handheld sequence "Medusa" (see Figure 1) results in some transformation errors for large baselines

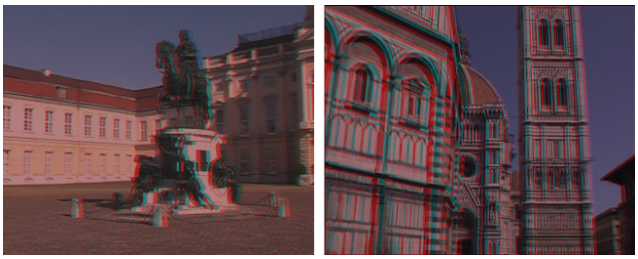

Figure 4: Red/cyan anaglyph stereo-image pairs of the sequences "Statue" (t_s = 10m) and "Dome" (t_s = 8m)

between the camera path and the virtual views as depicted in Figure 1 on the bottom left.

4. SUMMARY AND CONCLUSIONS

This paper presented a new approach for generation of stereo and multi-view video from monocular video. To our knowledge it was the first time that generation of 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 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.

The algorithm was tested on several data sets, five TV broadcast sequences and five sequences captured with a single handheld camera. In the previous section, the simulation results show the remarkable performance of the conversion process.

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 1 on the bottom left. Hence, a planar transformation might not be valid any longer.

Despite these restrictions 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.

5. 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.

6. 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] R. Hartley, and A. Zisserman, "Multiple view geometry", Cambridge University Press, UK, 2003

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

[8] 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

[9] 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

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

[11] L. MacMillan, "An image based approach to three-dimensional computer graphics", PhD dissertation, University of North Carolina, 1997