AN IMAGE-BASED RENDERING (IBR) APPROACH FOR REALISTIC STEREO VIEW SYNTHESIS OF TV BROADCAST BASED ON STRUCTURE FROM MOTION

Sebastian Knorr and Thomas Sikora

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

ABSTRACT

In the past years, the 3D display technology has become a booming branch of research with fast technical progress. Hence, the 3D conversion of already existing 2D video material increases more and more in popularity. In this paper, a new approach for realistic stereo view synthesis (RSVS) of existing 2D video material is presented. The intention of our 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 reconstruct the desired stereo views for each video frame. 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— Machine Vision, Image-based Rendering, Structure-from-Motion, Stereo-view Synthesis

1. INTRODUCTION

3DTV technology is currently being investigated in many research labs worldwide [1]. Especially the innovations regarding the 3D display technology are tremendous. However, the film industry still adheres to traditional capture techniques with a single camera. This divergence results in lack of 3D video content. To evade this situation, 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 rendering 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]. In the first category, *structure from motion* (SFM) techniques 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.

This paper proposes a new approach for realistic stereo view synthesis (RSVS) from monocular video sequences (e.g. TV broadcast video and sequences captured with a handheld camera). It combines both, the powerful algorithms of SfM [2] and *image-based rendering* (IBR) [11] to achieve stereoscopic views of high quality. First, the 3D structure and camera parameters are estimated with SfM. Then, the positions of the virtual stereo cameras for each frame of the sequence are calculated. Finally, the virtual stereo images are generated by estimating planar transformations (homographies) to surrounding views of the camera path and using IBR.

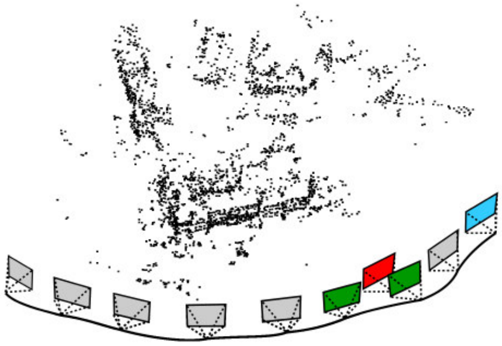


Fig. 1. Example of RSVS: estimated camera path and 3D structure

The main advantage of our approach is that the exact position of the virtual stereo view can be defined within the original camera path as depicted in Figure 1. When the cyan camera represents one frame of the original video sequence, its stereoscopic partner is indicated as red camera which represents the virtual left stereo view. The green cameras in Figure 1 are utilized to reconstruct the virtual stereo view with IBR. Hence, the computational expensive calculation of dense depth maps is avoided.

Another benefit of this approach is the occlusion handling. Whereas DIBR techniques always have to interpolate disclosed parts of the images when shifting pixels according to their depth values, the RSVS approach utilizes the information from surrounding views, i.e. occluded regions become visible within the sequence.

2. BACKGROUND

2.1. Structure from Motion

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

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 achieved by 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 [12]. Since, the feature correspondences between the views and the projection matrices are known, the sparse 3D structure is estimated with triangulation [12]. For a final refinement of the estimated parameters, non-linear minimization can be applied.

2.2. Stereo-view Synthesis using IBR

Once the 3D structure and the camera path are determined, a virtual stereo camera can be defined for each frame of the original video sequence (see Figure 1). With the principles of IBR [11] pixel values from surrounding views can be transferred to their corresponding positions in the virtual view. Thus, the virtual stereo image is just a rendered version of original images.

Since the camera parameters of the virtual camera are defined by the desired stereo-view setup, the 3D points of the estimated reconstruction can be projected into the virtual view (see Figure 2):

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

with $P_{stereo} = KR\left[I \mid \widetilde{C}_{stereo}\right]$. K is the internal calibration matrix, R is the rotation matrix which is identical with the rotation matrix of the corresponding original stereo view in a parallel camera setup, I is a 3x3 identity matrix and $\widetilde{C}_{stereo}$ is the position of the camera center in homogeneous coordinates.

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

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

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_{stereo} 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 [12] many robust and non-linear alternatives are introduced.

Once the homography between the virtual stereo view and the closest view of the video sequence is estimated, all pixel values of the original image can be projected to their corresponding locations in the stereo image.

3. PROPOSED SOLUTION

The input of the proposed algorithm is a 2D video sequence, either recorded with a DVB-T receiver or captured with a single handheld camera. Then, a standard SfM approach is applied to obtain the camera parameters and sparse 3D structure as described in Section 2.

3.1. Determine Positions of the Virtual Views

The first important step to generate a stereo view for each frame of the sequence is the adjustment of the horizontal differences between the left-and right-eye views, the so-called screen parallax values. Since the estimated camera path and the 3D structure are only defined up to scale, it is not clear at this stage whether the camera is close to a small 3D model or far away from a huge 3D scenery. Whereas the average human eye distance is known with approximately 64 mm.

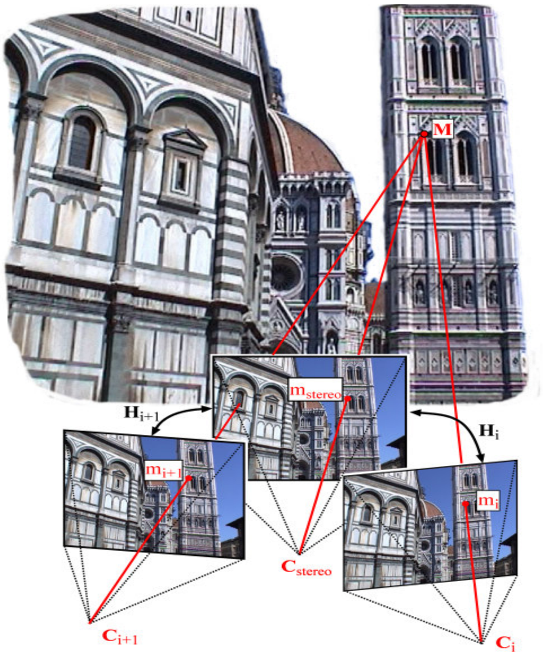


Fig. 2. Stereo view synthesis using IBR

To find a common basis of relative 3D structure and human eye distance, some interaction is required. 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 stereo camera is

$$C_{i,stereo}^m = C_i^m + R_i^{-1} \cdot [\pm t_x \quad 0 \quad 0]^T \ , \tag{4}$$

with $t_x = 64\ mm$, and the camera projection matrix

$$P_{i,stereo}^m = KR\left[I \mid \widetilde{C}_{i,stereo}^m\right] \ . \tag{5}$$

If t_x is fixed, as it is in this case, 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.

3.2. Determine Homographies for IBR

The estimation of the homographies is a critical task, since a large baseline length between the virtual stereo camera and a camera of the original path causes considerable errors if the image reflects a whole 3D scene and not just one plane of the scene [12].

Assuming that the baseline length is small, homographies are determined according to Section 2.2. Using eq. 2, pixel values of the original view can be projected to their corresponding positions in the virtual view. Since these positions do not exactly correspond

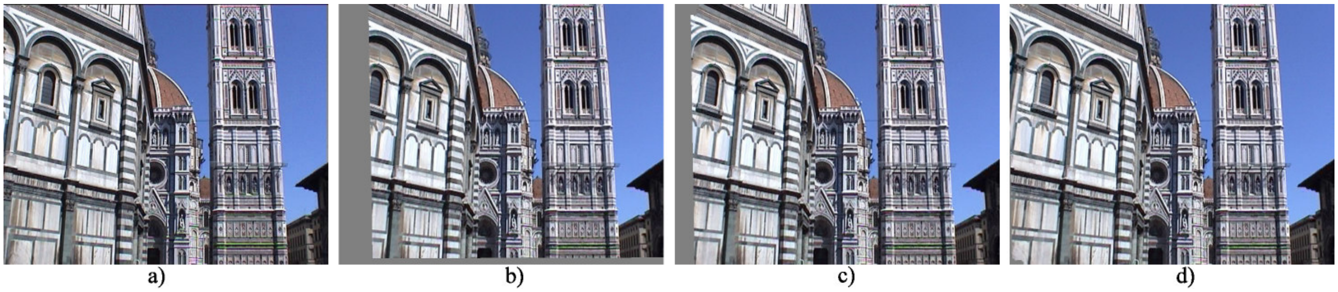


Fig. 3. Padding of pixels with additional frames: a) original left view of the sequence ”Dome”, b) virtual right view, only rendered with the closest view of the camera path, c) virtual right view using 30 and d) 62 frames of the original sequence.

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 with the virtual stereo camera as depicted in Figure 3b. 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 have to be taken into account (see Figure 3c and 3d).

A final aspect of this iterative process is the fact that some stopping criteria have to be defined, because it is not always possible to fill the whole virtual stereo image. The first stopping criterion is the transfer error when calculating the homographies. We take the median of the transfer errors ϵ_k

$$\underset{k}{\textbf{median}}\ \epsilon_k, \quad \epsilon_k = d\left(m_{i,k}, H_i m_{stereo,k}\right). \tag{6}$$

If this value is higher than a predefined threshold, no additional views are considered. $d\left(\cdot,\cdot\right)$ is the Euclidean distance between the original 2D feature positions and the transferred ones.

The second criterion is the degree of image reconstruction. If more than 99.5 % of the virtual image is covered with pixel values from surrounding views, the stereo view synthesis is completed.

4. 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 is used for all sequences to generate the stereo views. In Figure 3, the stereo view synthesis for an original left view of the ”Dome”-sequence generated with 62 frames of the original sequence was already presented. Figure 4a shows the anaglyph stereo-images of the TV broadcast video ”Pyramid” using RSVS. The distance t_s between the camera and the dominant scene was set to 8 meters.

Figures 4b and 4c show two anaglyph stereo-images (”Medusa” and ”Dome”) of video sequences captured with a handheld camera and converted with RSVS. Here, the distance t_s was set to 3 meters and 8 meters, respectively. The stereo images demonstrate the remarkable performance of our approach.

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.

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

Sequence	Resolution	t_s in meters	avg. # of views
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

5. SUMMARY AND CONCLUSIONS

This paper presented a new approach for realistic stereo view synthesis (RSVS) of monocular video sequences. 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.

The main advantage of this approach is that the exact position of a virtual stereo camera can be defined within the original camera path. Thus IBR can be utilized to generate the desired stereo views, i.e. a computational expensive dense depth estimation is avoided. Furthermore, the occlusion problem, which is always present in dense depth estimation, does almost not exist.

A final advantage of RSVS is that photo realism is achieved with no additional operations, since the photometric properties of a scene are determined entirely by the original frames of the reference sequence.

Nevertheless, this approach has some limitations which have to be mentioned. The most important one is that the scene has to be static, i.e. moving objects within the scene would disturb the stereoscopic depth perception. Furthermore, the camera movement has restrictions as well. If the camera moves only in a forward- or backward direction, a stereo view synthesis with this approach would fail. The case of a camera movement in up- and down direction can be solved by transposing the frames by 90 degrees. A final limitation is that a larger screen parallax increases the probability of a divergence between the camera path and the position of the virtual stereo cameras. Hence, a planar transformation might not be valid any longer.

For future research, we will apply moving object segmentation techniques to employ RSVS even on dynamic scenes.

6. 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.

7. REFERENCES

[1] Oliver Schreer, Peter Kauff, and Thomas Sikora, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, p. 364, John Wiley & Sons, 2005.

[2] T. Jebara, A. Azarbayejani, and A. Pentland, "3D structure from 2D motion," *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 66–84, 1999.

[3] M. Pollefeys, "3D modeling from images," Tutorial notes, tutorial organized in conjunction with ECCV 2000, Dublin, Ireland, June 2000.

[4] C. Tomasi and T. Kanade, "Shape and motion from image streams: a factorization method," *Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[5] S. Knorr, E. Imre, B. zkalayci, A. A. Alatan, and T. Sikora, "A modular scheme for 2D/3D conversion of tv broadcast," in *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," in *Proc. of the SPIE: Stereoscopic Displays and Virtual Reality Systems XII*, San Jose, CA, USA, March 2005.

[7] 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," in *Proc. of the SPIE: Stereoscopic Displays and Applications IX*, San Jose, CA, USA, 1998.

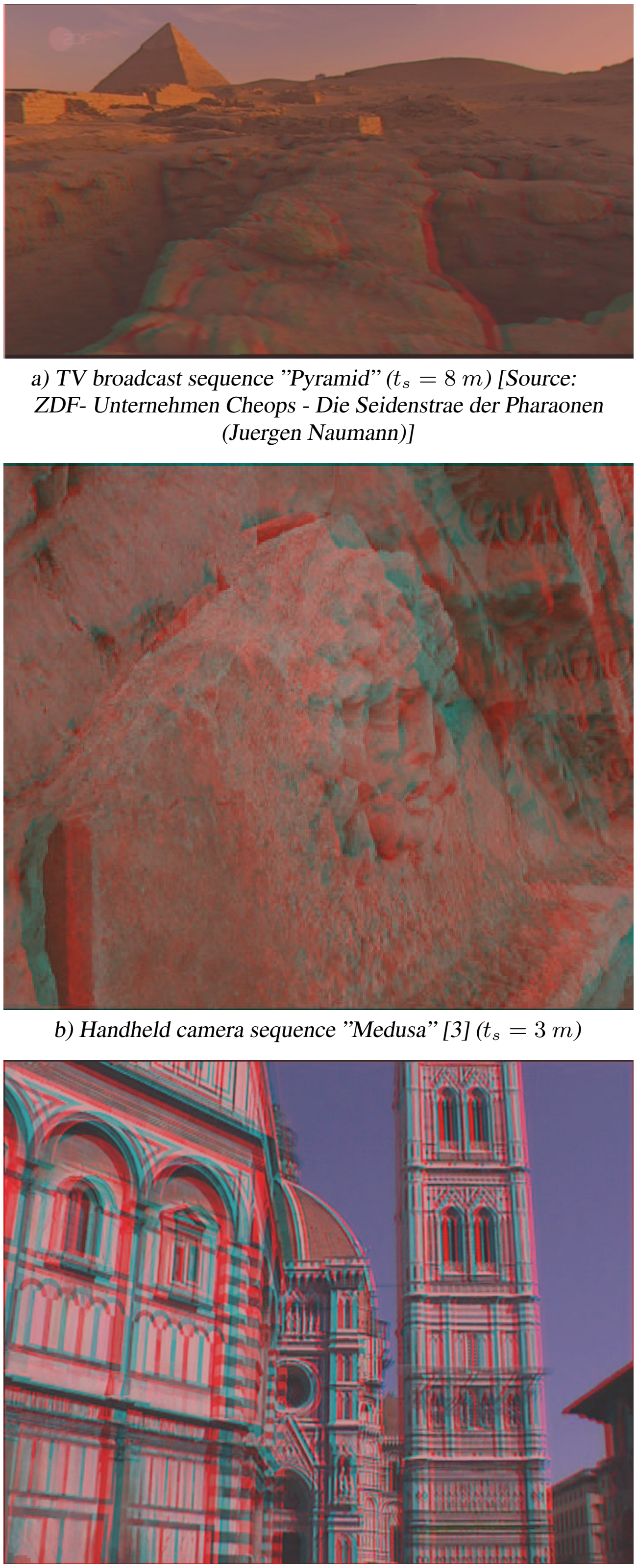
[8] K. Moustakas, D. Tzovaras, and M .G. Strintzis, "Stereoscopic video generation based on efficient structure and motion estimation from a monoscopic image sequence," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 8, pp. 1065 – 1073, August 2005.

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

[10] Christoph Fehn, "Depth-image-based rendering, compression and transmission for a new approach on 3D-TV," in *Proc. of the SPIE: Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, CA, USA, January 2004.

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

[12] R. Hartley and A. Zisserman, *Multiple view geometry*, Cambridge University Press, UK, 2003.



a) TV broadcast sequence "Pyramid" ($t_s = 8\ m$) [Source: ZDF- Unternehmen Cheops - Die Seidenstrae der Pharaonen (Juergen Naumann)]

b) Handheld camera sequence "Medusa" [3] ($t_s = 3\ m$)

c) Handheld camera sequence "Dome" ($t_s = 8\ m$)

Fig. 4. Red/cyan anaglyph stereo-image pairs.