Stereoscopic 3D from 2D Video with Super-Resolution Capability

Sebastian Knorr, Matthias Kunter and Thomas Sikora

Communication Systems Lab, Technische Universität Berlin, Einsteinufer 17, Berlin, Germany

***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. Subjective quality measurements with 25 real and 3 synthetic sequences were carried out to test the performance of RSVS against simple time-shift and depth image-based rendering (DIBR). Our approach heavily enhances the stereoscopic depth perception and gives a more realistic impression of the observed scenery. Simulation results applying super-resolution show that the image quality can further be improved by reducing motion blur and compression artifacts.**

***Index Terms*—stereoscopic imaging, 2D/3D conversion, structure-from-motion, super-resolution stereo, image-based rendering**

I. 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 maturing and 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 into super-resolution 3D is important for content owners. Movies may be reissued in 3D in the future.

Many fundamental algorithms have been developed to reconstruct 3D scenes from monocular video sequences [2]-[24]. These algorithms can roughly be divided into two categories: methods that tend to create a complete 3D model of the captured scene [2]-[10], and methods that just render stereoscopic views [11]- [24].

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.

The authors are with the Technische Universität Berlin, 10623 Berlin, Germany (e-mail: knorrr@nue.tu-berlin.de)

This work was developed within 3DTV (FP6-PLT-511568-3DTV), a European Network of Excellence funded under the European Commission IST FP6 programme.

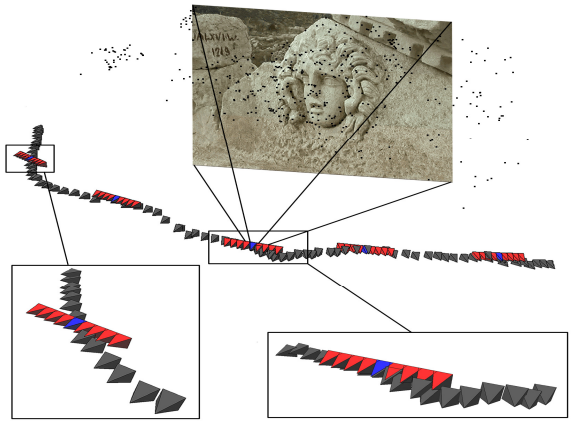

Fig. 1. Multi-view synthesis using SfM and IBR; dark gray: original camera path, red: virtual stereo cameras, blue: original camera of a multi-view camera setup

For the second category, *depth-image-based rendering* (DIBR) [11]-[17] seems to be the most promising technique both for stereoscopic view synthesis and for transmission in 3D-TV broadcast systems [25][26]. DIBR approaches render new virtual views via dense depth maps for each frame of the sequence by shifting image pixels according to their assigned depth. On the other hand, dense depth estimation is still an error prone task and computationally very expensive. In [17] a semi-automatic approach for dense depth estimation was introduced using a machine learning algorithm (MLA) for assigned keyframes and depth tweening between these frames.

Other approaches, e.g. [18][19], are using motion parallax or spatio-temporal interpolation to generate the desired stereoscopic views. Ross [20] introduced a very simple but (for some video sequences) effective technique for stereoscopic depth impression using binocular delay. Finally, in [21] planar transformations on temporal neighboring views are utilized to virtually imitate a parallel stereo camera rig. But, in any case, time consistency along the sequence is heavily dependent on the 3D scene, since a stereo rig is not correctly modeled.

In this paper, we present a new approach for generation of super-resolution stereo and multi-view video from monocular video based on *realistic stereo view synthesis* (RSVS) [22]. It combines both the powerful algorithms of SfM [2] and the idea of *image-based rendering* (IBR) [27] to achieve photo-consistency without relying on dense depth estimation.

Most available 3D display systems rely on 2 views (stereo video) to create a depth impression. However, more advanced systems use multiple views (e.g. 8 views showing the same

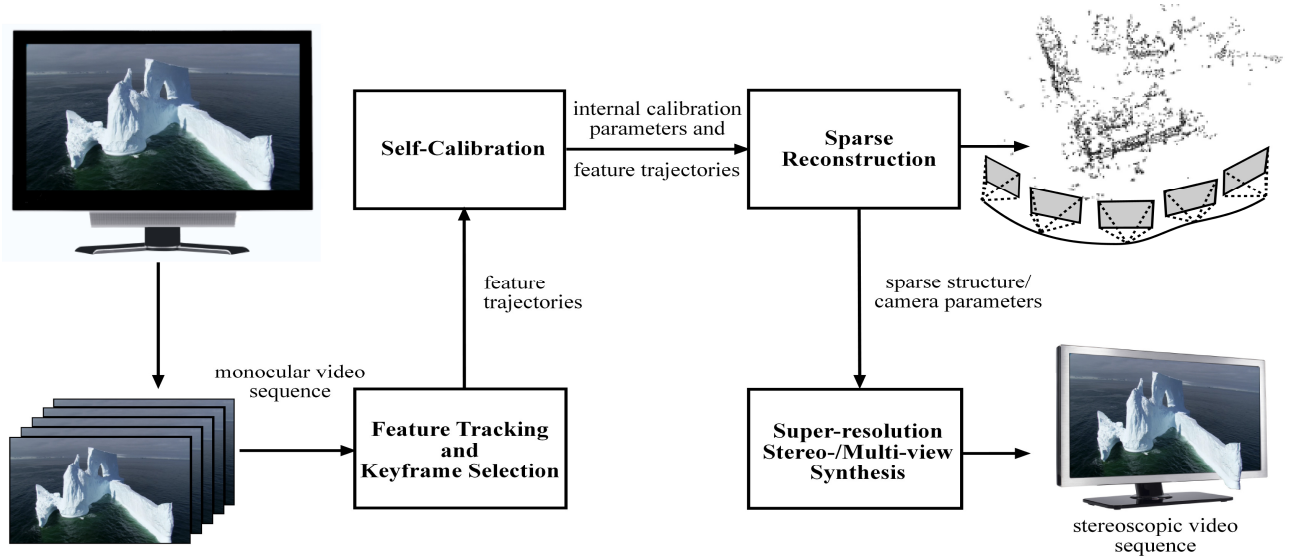

Fig. 2. System overview of the proposed solution

scene from different viewpoints). The presented algorithm is applicable to generate stereo video in its basic mode [22], but it is also capable to generate multi-view video [23]. We will show that the approach is quite suitable for converting existing 2D video material into multi-view with higher resolution [24]. To our knowledge it is the first time that an approach for generation of super-resolution multi-view video from monocular video is presented.

The proposed technique is performed in several stages. First, sparse 3D structure and camera parameters are estimated with SfM for the monocular video sequence (dark grey cameras in Fig. 1). Then, for each original camera position (white in Fig. 1) a corresponding multi-view set is generated (light grey in Fig. 1). This is done by estimating planar homographies (perspective transformations) to temporal neighboring views of the original camera path. Surrounding original views are used to generate the multiple virtual views with IBR. Hence, the computationally 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 a desired resolution higher than the original one. In order to do so, 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 V, the possible reduction of blur and coding artifacts. A complete overview of the proposed conversion system is illustrated in Fig. 2.

The organization of this paper is as follows: The next section describes the fundamentals of SfM as an initial step to estimate the camera path and to define virtual stereo cameras with constant parallax over time. In Section III and IV, the RSVS approach for stereo- and multi-view synthesis and the super-resolution extension are outlined, which are the main contributions of our work. Simulation results are presented in Section V. In Section VI, psycho-visual experiments are carried out to evaluate the performance of RSVS against standard conversion methods. The limitations of our approach are stated in Section VII. Finally, in Section VIII, the paper concludes with a summary and a discussion.

II. Structure-from-Motion Fundamentals

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.

Finding relations between the views in the video sequence is the initial step of our reconstruction process. The geometric relationship, also known as epipolar geometry, can be estimated with a sufficient number of feature correspondences between the views [28]. Once the images are related, the camera projection matrices are calculated using singular value decomposition [29]. If feature correspondences between the views and projection matrices are known, sparse 3D scene structure is estimated with triangulation [30], i.e. for a limited number of points the 3D coordinates are available as illustrated in Fig. 1. For a final refinement of the estimated parameters, bundle adjustment is often used [31].

The following subsections will give a more detailed description of the processing steps needed for our 3-D scene reconstruction approach.

A. Feature Tracking and Keyframe Selection

Ambiguity is the major problem in finding feature correspondences in images. Such image features should be invariant or salient like points lying on edges, corners, line segments, contours, regions or even whole objects. The ambiguity decreases with the information content of the features, i.e. matching an object in two or more views is much

more reliable than matching a single point. However, narrow-baseline applications mostly deal with feature points like corners [32][33], because the extraction has less complexity and the reliability is still very high.

As long as the input is a sequence of consecutive frames, the Kanade- Lukas tracker (KLT) [34] successfully tracks features throughout the sequence. In [35] an extension of the KLT was introduced.

Since a large baseline is needed to relate images, i.e. to estimate the epipolar geometry, consecutive video frames are not really suitable. Torr et al. [36] introduced the Geometric Robust Information Criterion (GRIC) as a robust model selection criterion to detect keyframes within a video sequence. Since the baseline between consecutive frames is small, a 2D motion model H (homography) can be used to transfer features from one frame to corresponding positions in a second frame. If the baseline increases during the tracking process and if the features are part of a 3D scene structure, the projection error increases as well, i.e. the 2D motion model must be upgraded to a 3D motion model F (epipolar geometry). Initializing the first frame of the sequence as keyframe and proceeding frame by frame, the next keyframe is selected if the GRIC value of the motion model F is below the GRIC value of H. The GRIC score is defined as:

$$GRIC = \sum \rho\left(e_i^2\right) + \lambda_1 dn + \lambda_2 k\,, \tag{1}$$

where $\rho(e_i^2) = \min\left(e_i^2/\sigma^2,\ \lambda_3(r-d)\right)$. The parameters are defined as follows: d is the dimension of the selected motion model (H has the dimension two and F dimension three), r is the dimension of the data (i.e. four for two views), k is the number of the estimated model parameters (seven for F and eight for H), n is the number of tracked features, σ is the standard deviation of the error on each coordinate and e_i is the distance between a feature point transferred through a planar homography H and the corresponding point in the target image or the Euclidian distance between the epipolar line of a feature point and its corresponding point in the target image (dependent on the selected model M):

$$e_i = D\left(m_i', Mm_i\right). \tag{2}$$

The parameters λ_1, λ_2, and λ_3 are tuning parameters with $\lambda_1=2$, $\lambda_2=log(4n)$ and $\lambda_3=2$ [37].

Hence, the baseline distance between the selected keyframes is sufficient for the estimation of the epipolar geometry.

B. Multi-view Reconstruction

The first step of our structure and motion recovery is the initial structure computation, i.e. taking the feature correspondences from the tracker into account, the fundamental matrix F is estimated between the first two keyframes of the sequence [28]. The *random sample consensus* (RANSAC) [38] is a robust algorithm which selects inliers for the computation of F. Afterwards, the projection matrices P_1 and P_2 are determined with singular value decomposition (SVD). The world frame is aligned with the first camera [39]. Once the projection matrices are known, the 3D points of the feature correspondences are found via optimal triangulation as described in [30].

The next step is the updating of the 3D structure and camera motion. First the camera projection matrix for the next keyframe is determined robustly using already existing 3D-2D feature correspondences as described in [39]. Then, the 3D structure and the camera matrix is refined with additional 2D-2D feature correspondences between the actual frame and the previous one. This procedure is repeated for all keyframes. A final refinement of the structure and motion recovery is done via global nonlinear minimization techniques for all frames, also known as bundle adjustment [31]. The cost function used for the minimization is

$$\min_{P_i, M_j} \sum_{i=1}^{m} \sum_{j=1}^{n} D(m_{ij}, P_i M_j)^2\,, \tag{3}$$

where $D(\cdot,\cdot)$ is the Euclidean distance between the 2D features m_{ij} and the re-projected 3D points M_j.

Since the performance of the reconstruction is heavily dependent on the initial structure computation, Imre et al. [40] introduced a prioritized sequential 3D reconstruction approach for a fast and reliable structure and motion computation. The keyframes are re-ordered according to a priority metric, and the frame pair with the highest priority metric is then used for the initial reconstruction.

C. Self-Calibration

If the internal calibration parameters are unknown, which in general is the case for TV broadcast, home videos or cinema movies, a self-calibration procedure has to be carried out. A method that requires a pairwise calibration (i.e. the fundamental matrix) was introduced in [41]. It is based on the concept of the *absolute conic*. In [42] and [29] a stratified approach from projective to metric reconstruction was described. Finally, in [43] a self-calibration procedure was introduced that estimates the internal camera parameters via constraints on the essential matrices, i.e. the initial focal length α is estimated by the assumption of a unity aspect ratio $\alpha_u = \alpha_v$ and the principal point at the center of the image.

III. Realistic Stereo- and Multi-view Synthesis

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 Fig. 1. A white camera corresponds to an original image of a video sequence and the light grey cameras represent its corresponding multiple virtual views. With the principles of IBR pixel values from temporal neighboring views 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 Fig. 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 III A. for details on calculation of translation).

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

$$m_{multi} = P_{multi} M \,, \tag{4}$$

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.

A. *Determination of the 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 a 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. Without loss of generality, the world coordinate system is located in the centroid of the sparse 3D point cloud. Thus, the absolute position of all cameras regarding the world coordinate system can be determined with

$$C_i^m = t_s \frac{C_i}{\|C_1\|} \,, \tag{5}$$

where superscript m indicates the normalization in meters and $\|C_1\|$ is the vector norm of the first camera. The position of each corresponding virtual camera is

$$C_{i,multi}^m = C_i^m + R_i^{-1} \cdot \begin{bmatrix} \pm t_x \\ 0 \\ 0 \end{bmatrix}, \quad \text{with } t_x = n \cdot 64mm, \tag{6}$$

(n=1, 2, 3, …, N) and the camera projection matrix

$$P_{i,multi}^m = KR\left[I \,\middle|\, -\tilde{C}_{i,multi}^m \right]. \tag{7}$$

N is the number of virtual views that should be generated for each frame of the sequence. 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.

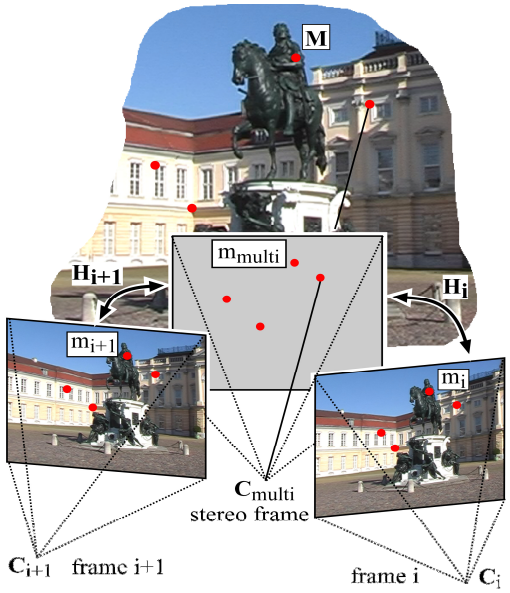

Fig. 3. Stereo-/multi-view synthesis using planar transformations

B. *Determination of the homographies for IBR*

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

$$m_i = H_i m_{multi} \,. \tag{8}$$

H is a 3x3 matrix and therefore it contains 9 entries, but is defined only up to a scale:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{9}$$

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, the first directly from SfM and the second by calculation via eq. 4. Thus, the perspective transformation parameters of H can be estimated from eq. 8 with a minimum number of four point correspondences. Here, a linear estimation with all point correspondences was applied using singular value decomposition. In Hartley and Zisserman [29] many robust and non-linear alternatives are introduced.

Once the perspective transformation between a virtual view to be generated and the closest original view of the video sequence is estimated, all pixel values of the original image can be projected to their corresponding locations in the virtual image:

$$u' = \frac{a_0 + a_1 u + a_2 v}{1 + c_1 u + c_2 v}$$
$$v' = \frac{b_0 + b_1 u + b_2 v}{1 + c_1 u + c_2 v} \tag{10}$$

a) b) c) d)

Fig. 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.

Fig. 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)

with

$$a_0 = \frac{h_{13}}{h_{33}},\ a_1 = \frac{h_{11}}{h_{33}},\ a_2 = \frac{h_{12}}{h_{33}}$$

$$b_0 = \frac{h_{23}}{h_{33}},\ b_1 = \frac{h_{21}}{h_{33}},\ b_2 = \frac{h_{22}}{h_{33}}$$

$$c_1 = \frac{h_{31}}{h_{33}},\ c_2 = \frac{h_{32}}{h_{33}}$$

Since these positions do not exactly correspond with the pixel grid, 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 Fig. 4b. This is particularly the case when the orientation of both cameras differs significantly. To fill the missing parts of the virtual stereo image, we take additional surrounding views into account (see Fig. 4c and d).

A final aspect of our 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 median transfer error ε_k for all k feature correspondences when calculating the homographies:

$$\underset{\forall k}{median}\ \varepsilon_k\,, \quad \text{with}\ \ \varepsilon_k = D\left(m_{i,k}, H_i m_{multi,k}\right), \tag{11}$$

where $D(\cdot,\cdot)$ is the Euclidean distance between 2D features $m_{\mathrm{i,k}}$ of an original view and corresponding 2D features $m_{\mathrm{multi,k}}$ of the desired virtual view transferred through a planar-homography. If this value is higher than a predefined threshold (e.g. 0.5 pixel), no additional views are considered.

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 virtual view synthesis is completed.

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

IV. Super-Resolution Stereo- and Multi-view Synthesis

The previous section described our 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) [44]-[46]. The main objective is to construct one or 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 [45] and usually consists of three stages: registration, interpolation, and restoration.

A. 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.

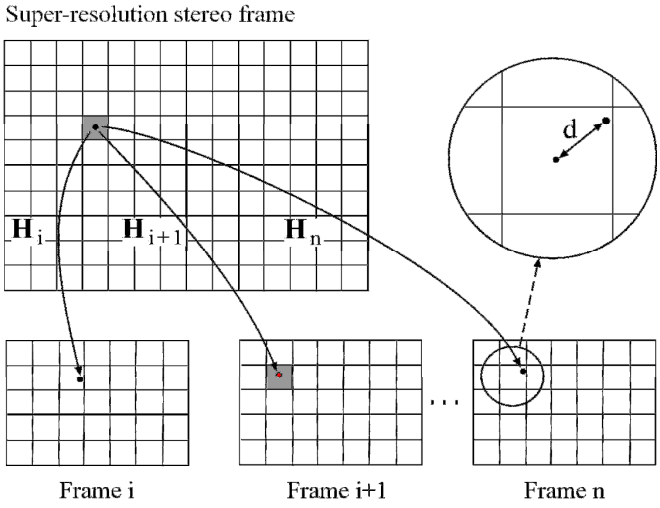

Fig. 6. Super-resolution stereo-/multi-view synthesis

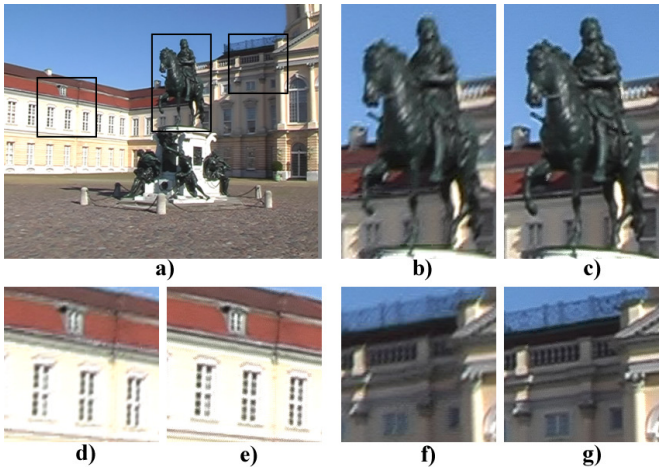

Fig. 7. Super-resolution stereo view synthesis of the "Statue"-sequence (1080x864 pixels). a) super-resolution stereo view. b), d), f) close-up of the up-sampled and c), e), g) close-up of the super-resolution stereo frame

For each pixel in a stereo frame we determine the position in surrounding views as described in section III. 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 Fig. 6 an example of this process is given. Assume 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.

B. 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

$$I_{med}(x,y) = \underset{\forall i}{\textbf{median}}\ I_i(x,y), \tag{12}$$

where I is the color value of the pixel in each frame i. Pixel values with a high absolute deviation from the median (i.e. statistically unreliable) 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.

A further enhancement of the super-resolution module might be achieved if bi-cubic interpolation is applied for pixel warping instead of bilinear interpolation. However, a significant performance gain is not expected since the interpolation effects are heavily reduced by choosing the best pixel locations.

V. Simulation Results

Two example figures show the performance of the super-resolution mode of our approach. In Fig. 7, a super-resolution virtual stereo frame (size 1080x864 pixels) of the "Statue"-sequence is presented. Three close-ups should stress the difference between the super-resolution frame and an up-sampled frame using Lanczos-filtering (original size was 720x576 pixels). Fig. 7d shows some typical artifacts when dealing with interlaced PAL video and up-sampling: Sawtooth pattern can be noticed along edges resulting from de-interlacing. 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 Fig. 7).

Fig. 8 shows 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. Four close-ups illustrate the reduction of the previous mentioned artifacts on sequence "Dome". Especially, the sawtooth pattern was significantly reduced in the super-resolution mode (Fig. 8c,g,i). Furthermore, aliasing artifacts become more visible in the up-sampled frame, which can be seen on the top of the right arc in Fig. 8e.

VI. Psycho-visual Experiments

Two subjective quality tests were carried out to stress the performance of RSVS against standard 2D/3D conversion approaches like DIBR and time-shift. The test conditions are standard conform to the ITU recommendations [47]. In the first session of the experiment, 15 subjects were asked to rate the quality impression of 25 real test sequences. Since no ground truth data was available (the test sequences were originally captured with a single camera), a *single stimulus* (SS) method was applied [48].

In a second session, again 15 subjects were asked to rate the quality impression of three synthetic test sequences rendered

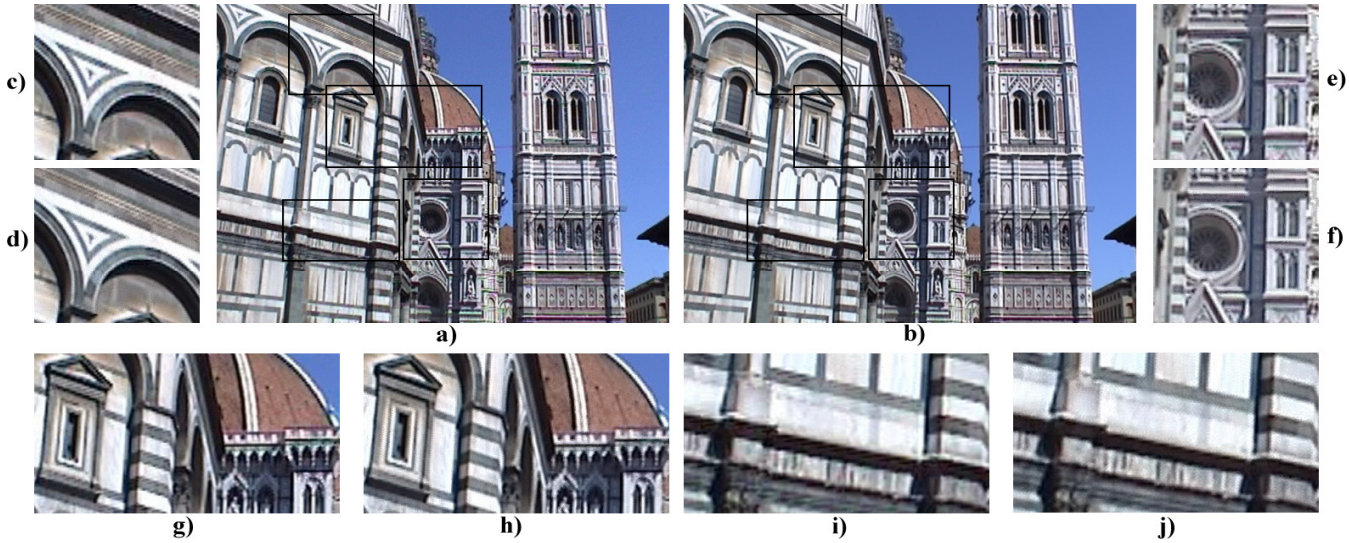

Fig. 8. Super-resolution stereo view synthesis of the "Dome"-sequence (1080x864 pixels). a) virtual stereo view up-sampled and b) super-resolution stereo view. c), e), g), i) close-up of the up-sampled and d), f), h), j) close-up of the super-resolution frame

from a 3D model with a parallel stereo camera rig. Since ground truth data could always be used as reference, a *double stimulus continuous quality scale* (DSCQS) method was applied [48].

A. *Subjective Quality Test (SS method)*

In the first session of the experiment, 15 subjects were asked to rate the quality impression of 25 test sequences. 18 of the sequences were taken from the BBC documentation "Planet Earth" and seven were captured with a handheld camera (Canon XLs 1, progressive scan, DV coded). Two example pictures are illustrated in Fig. 9. Whereas the BBC sequences have a very smooth and linear camera path, the camera motion of the handheld sequences was very unsteady. All test sequences have similar (almost static) content.

Each of the sequences was displayed in three different modes: 2D (the same sequence is presented to the left and right eye), 3D converted with RSVS and 3D using a time-shift (i.e. a time-delay of 4 to 10 frames was applied to the left or right eye sequence). The presentation period of each test sequence was about 10 seconds followed by the voting period of 10 seconds, where the assessors had to rate the overall quality of the previous displayed test sequence.

The evaluation results, in terms of the average *mean opinion scores* (MOS) and the standard deviations (S.D.), are presented in TABLE I. In Fig. 11 a diagram of the average mean opinion scores with 95% confidence intervals for all test sequences is displayed.

The overall ratings indicate that 3D converted with RSVS has the best quality impression with a MOS of 7.485 on all assessors. Due to a smooth horizontal and linear camera motion, the ratings for the BBC sequences are very similar. Here, a time-shift yields almost the same results as RSVS, although a small vertical parallax could be noticed for some of the time-shifted sequences. For the handheld sequences, RSVS outperforms time-shift significantly. Here, the vertical parallax in the time-shifted sequences, caused by unsteady camera motion, was quite annoying to all test subjects. Thus, the test

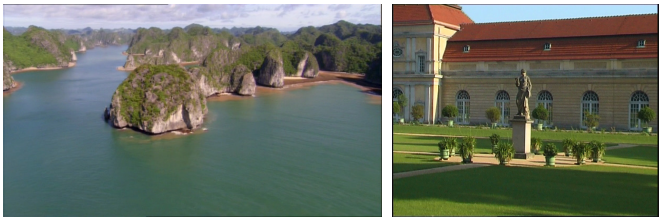

Fig. 9. Example pictures of the test sequences. Left: "Caves09" from the BBC-documentation "Planet Earth", series "Caves" (720x405 pixels). Right: handheld sequence "Charlottenburg02" (720x576 pixels)

Fig. 10. Example pictures of the three synthetic sequences. Left: "TUBroom1", middle: "TUBroom2", right: "TUBroom3"

demonstrates the flexibility of RSVS against simple time-shift, which is heavily restricted to horizontal camera movements.

Although a 3D presentation of the test sequences was always preferred by the subjects, the 2D mode got similar results as RSVS. One reason is the appearance of *cross talk* in the 3D sequences resulting from imperfect image separation [49], i.e. the left-eye view leaks through to the right-eye view and vice versa. This effect is naturally not visible in the 2D mode. Thus, the stereoscopic depth perception has both a positive contribution to the overall image quality and a negative effect caused by cross talk.

TABLE I
SS METHOD: AVERAGE MEAN OPINION SCORES AND STANDARD DEVIATIONS

Data set	2D		3D /RSVS		3D / time-shift	
	MOS	S.D.	MOS	S.D.	MOS	S.D.
BBC	7.707	1.663	7.830	1.703	7.819	1.792
handheld	6.390	1.550	6.600	1.858	2.971	1.632
overall	7.339	1.631	7.485	1.746	6.461	1.747

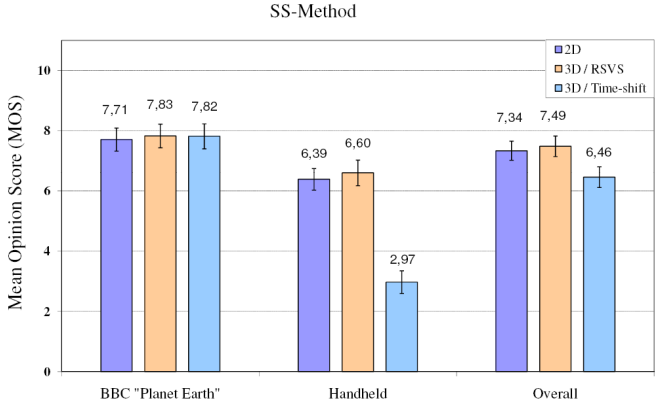

Fig. 11. Average results of the SS-method. Average perceived subjective quality impression (MOS) with 95% confidence intervals

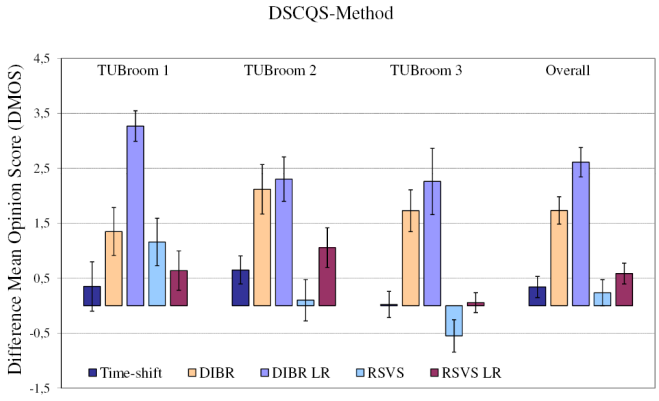

Fig. 12. Average results of the DSCQS-method. Average perceived subjective quality impression (DMOS) with 95% confidence intervals

B. *Subjective Quality Test (DSCQS method)*

In the second session of the experiment, 15 subjects were asked to rate the quality impression of three synthetic test sequences (see Fig. 10) rendered from the same 3D model with different perspectives and camera motions, which have different critical influences on the conversion algorithms and even on the reference sequences. Each of the sequences was displayed in five different 3D modes converted with: time-shift, standard DIBR [15] (one view remains original), DIBR LR[1] (left and right view are generated from the original view), standard RSVS (one view remains original) and RSVS LR[1] (left and right view are generated from surrounding original views).

In the first sequence the camera moved partly in a forward direction, which is critical for the synthesis with RSVS. The second sequence has a slightly vertical motion, so that some vertical parallax is expected for time-shift. Finally, the third sequence has a large foreground object passing the image boundaries, which is even critical for the reference parallel stereo rig. Here, a break of depth cues is expected since a parallel camera rig produces only negative parallax, with the objects coming out of the screen. When passing the image boundaries, these objects should be located behind the screen (positive parallax) with respect to the monocular depth cues.

[1] The parallax value from left to right view is identical to the standard approach

TABLE II
DSCQS METHOD: AVERAGE DIFFERENCE MEAN OPINION SCORES AND STANDARD DEVIATIONS [DMOS (S.D.)]

Data set	Time-shift	DIBR	DIBR LR	RSVS	RSVS LR
TUBroom1	0.35 (1.77)	1.35 (1.72)	3.27 (1.10)	1.16 (1.71)	0.64 (1.42)
TUBroom2	0.65 (1.01)	2.12 (1.78)	2.30 (1.59)	0.10 (1.49)	1.06 (1.43)
TUBroom3	0.02 (0.94)	1.73 (1.50)	2.26 (2.38)	-0.55 (1.16)	0.06 (0.72)
overall	0.34 (1.32)	1.73 (1.70)	2.61 (1.83)	0.24 (1.63)	0.58 (1.30)

According to the DSCQS test conditions, the subjects were presented with a series of pairs of stereoscopic sequences. Each converted sequence (test) was related twice successively to its error free version (reference). The time slot for each stereoscopic sequence was 9 seconds and the time between two sequences was 2 seconds. The persons recorded their assessment of the quality of both sequences (reference and test) on two continuous graphical scales for each test period. A measurement of length makes the subjective score available, which is within a range of 0 to 10.

The evaluation results, i.e. the average *difference mean opinion scores* (DMOS) and the standard deviations (S.D.), are presented in TABLE II. In Fig. 12 a diagram of the average mean opinion scores with 95% confidence intervals for all test sequences is displayed.

The subjective quality test shows that our proposed approaches yield excellent results. Starting with the overall ratings, the sequences converted with standard RSVS has the best quality impression with a DMOS of 0.24. Even time-shift achieved good results for all test sequences, although, as mentioned in the previous subsection, vertical parallax and *shear distortion* [49] could be noticed. For the "TUBroom 1"-sequence time-shift even outperforms all other conversion methods. The good performance of time-shift results from two significant factors. First, all test sequences have almost a horizontal and linear camera movement, which is necessary for a simple delay in time. Second, time-shift is the only conversion method that uses two original frames of the sequence to generate a stereoscopic depth perception. Some of the subjects noticed a blurring effect on the sequences, which is always present in the virtual views of the other conversion methods (resulting from the bilinear interpolation). This can also be seen in the ratings of the DIBR LR and RSVS LR approaches where both stereoscopic views are virtually synthesized. Only for the "TUBroom 1"-sequence RSVS LR outperforms standard RSVS. Here, the camera motion was quite critical for RSVS (see section VII for details on the limitations). Namely, the camera moved temporarily in a forward direction. Thus, the baseline increases between a virtual view and a neighboring original view resulting in larger transformation errors.

Finally, RSVS has a negative DMOS for the "TUBroom 3"-sequence, i.e. most of the test subjects rated the test sequence better than the reference. Here, the reference sequence had a break in the depth cues: The cube in the foreground (see Fig.

10) went out of the image boundary, meaning, the cube must be located behind the image screen (monocular depth cue). Since a parallel camera setup was used for the reference sequence resulting in a negative parallax, the cube is located in front of the screen (stereoscopic depth cue). This break of depth cue is also present in the test sequence, but its effect was reduced by a smaller negative parallax and an additional positive parallax.

Although the depth range between fore- and background is quite high, the planar transformation errors of RSVS are still small and not visible within the sequence.

VII. Limitations

The RSVS approach presented 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 motion. 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 Fig. 1 on the bottom left. Hence, a planar transformation might not be valid any longer if the scene consists of fore- and background objects with different depths, i.e. the larger the depth range of the scene is, the more transformation errors can occur if the baseline length between virtual an original views increases.

VIII. 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 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. The simulation results show the remarkable performance of the conversion process. Especially the super-resolution mode reduced significantly de-interlacing-, aliasing-, ghosting- and blurring artifacts.

Furthermore, we evaluated standard RSVS with two subjective quality tests. In the first test RSVS was compared with a 2D presentation mode and a 3D presentation mode utilizing time-shift. Although a slightly vertical parallax was noticed in the time-shift mode, which naturally causes eye-strain, the results compared with RSVS were very similar. Since the duration of typical video is much longer, a further study of this stereoscopic impairment has to be carried out.

In a second subjective test, five different conversion methods were applied to three synthetic test sequences. Although ground truth depth maps for DIBR were available, RSVS significantly outperforms DIBR. Usually, existent video material to be converted into stereoscopic 3D has no depth maps available. Hence, dense depth estimation for each frame has to be applied, which is, as stated in Section I, still an error prone task and computationally very expensive. Nevertheless, if depth maps are available, DIBR has no restrictions concerning video content or camera motion. Furthermore, the synthesis of virtual views is less complex than using RSVS which heavily depends on the camera motion. Only for horizontal panning with almost no camera rotations, where RSVS just needs one original view to generate the desired virtual view, DIBR and RSVS have the same complexity.

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 post-production is still done using semi-automatic software systems. Here the presented automatic algorithm heavily reduces the manual workload for many sequences.

Acknowledgment

The authors would like to thank Aljoscha Smolic, Peter Kauff, Ingo Feldmann and Christoph Fehn from FhG-HHI for the fruitful discussions on 2D/3D-conversion issues.

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] R. Szeliski and S. B. Kang, "Recovering 3D shape and motion from image streams using non-linear least-squares", Journal of Visual Communication and Image Representation, 5(1):10–28, March 1994

[4] R. Szeliski, "Scene reconstruction from multiple cameras", in IEEE Signal Processing Society, Int. Conf. on Image Processing, Vancouver, Canada, September 2000

[5] P. Beardsley, P.H.S. Torr, and A. Zisserman, "3D model acquisition from extended image sequences", In Proc. of European Conf. on Computer Vision (ECCV), pp. 683-695, 1996

[6] 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

[7] D. Nister, "Automatic passive recovery of 3D from images and video", in Proc. of Int. Symposium on 3D Processing, Visualization and Transmission (3DPVT), pp. 438–445, Thessaloniki, Greece, September 2004

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

[9] P. Sturm, B. Triggs, "A factorization based algorithm for multi-image projective structure and motion", in Proc. of the European Conference on Computer Vision (ECCV), pp. 709–720, Cambridge, UK, 1996

[10] 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

[11] J.-F. Evers-Senne, A. Niemann, and R. Koch, "Visual reconstruction using geometry guided photo consistency", Int. Workshop on Vision, Modeling, and Visualization (VMV), Aachen, Germany, 2006

[12] 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

[13] 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

[14] 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

[15] 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

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

[17] P. Harman, J. Flack, S. Fox, M. Dowley, "Rapid 2D to 3D conversion", in Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems X Vol. 4660, pp. 78-86, 2002

[18] Y. Matsumoto, H. Terasaki, K. Sugimoto, and T. Arakawa, "Conversion system of monocular image sequence to stereo using motion parallax", in Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems IV, Vol. 3012, pp. 108-112, May 1997

[19] B. J. Garcia, "Approaches to stereoscopic video based on spatio-temporal interpolation", in Proc. of SPIE: Stereoscopic Displays and Virtual Reality Systems IV, Vol. 2653, pp. 85-95, Apr. 1996

[20] J. Ross, "Stereopsis by binocular delay", *Nature* 248, Vol 2, 363-364, 1974

[21] 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

[22] 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.

[23] S. Knorr, A. Smolic, and T.Sikora, "From 2D- to stereo- to multi-view video", 3DTV-CON, Kos, Greece, 2007

[24] S. Knorr, M. Kunter, and T. Sikora, "Super-resolution stereo- and multi-view synthesis from monocular video sequences", 3-D Digital Imaging and Modeling (3DIM), Montréal, Canada, 2007.

[25] M. Op de Beeck and A. Redert, "Three dimensional video for the home", in Proc. of EUROIMAGE ICAV3D 2001, International Conference on Augmented, Virtual Environments and Three-Dimensional Imaging , pp. 188-191, 2001

[26] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, L. van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimised approach on 3D-TV", International Broadcasting Convention (IBC), 2002

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

[28] Q.-T. Luong and O. Faugeras, "The fundamental matrix: theory, algorithms, and stability analysis", Int. Journal of Computer Vision, 17(1):43-76, 1996

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

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

[31] B. Triggs, and P. McLauchlan, R. Hartley, A. Fitzgibbon, "Bundle adjustment - a modern synthesis", in "Vision Algorithms: Theory & Practice", LNCS Vol.1883, pp.298-372 Springer-Verlag, 2000

[32] C. Harris and M. Stephens, "A combined corner and edge detector", Fourth Alvey Vision Conference, pp.147-151, 1988

[33] J. Shi and C. Tomasi, "Good features to track", IEEE Int. Conf. on Computer Vision and Pattern Recognition, Seattle, June 1994

[34] C. Tomasi, and T. Kanade, "Detection and tracking of point features", Technical Report CMU-CS-91-132, Carnegie Mellon University Technical, 1991

[35] J.-Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker", Intel Corporation, Microprocessor Research Labs, 2000 <http://www.intel.com/research/mrl/research/opencv/>

[36] P.H.S. Torr, A.W. Fitzgibbon, and A. Zisserman, "Maintaining multiple motion model hypotheses over many views to recover matching and structure", Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), pp. 485-491, 1998

[37] P.H.S. Torr, A.W. Fitzgibbon and A. Zisserman, "The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences", International Journal of Computer Vision, 32(1):27-44, August 1999

[38] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, pp. 381-385, 1981

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

[40] 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.

[41] Q.-T. Luong and O. Faugeras, "Self calibration of a moving camera from point correspondences and fundamental matrices", Int. Journal of Computer Vision, vol.22-3, 1997

[42] M. Pollefeys, "Self-calibration and metric 3D reconstruction from uncalibrated image sequences", PhD thesis, K.U.Leuven, 1999

[43] 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

[44] M. Irani and S. Peleg, "Super resolution from image sequences", In Proc. of International Conference on Pattern Recognition (ICPR), Atlantic City, NJ, 1990

[45] 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

[46] S. Borman and R. L. Stevenson "Super-Resolution from image sequences - a review," Midwest Symposium on Systems and Circuits, pp. 374-378, 1998

[47] ITU, "Methodology for the subjective assessment of the quality of television pictures", Recommendation BT.500-10 , 2000

[48] F. Pereira and T. Ebrahimi, *The MPEG-4 Book*, Prentice Hall, 2002

[49] L. M. J Meesters, W. A. IJsselsteijn, and P. J. H. Seuntiëns, "A survey of perceptual evaluations and requierements of three-dimensional TV", IEEE Trans. On Circuits and Systems for Video Technology, Vol. 14, No. 3, pp. 381-391, March 2004