STEREOSCOPIC 3D VIEW SYNTHESIS FROM UNSYNCHRONIZED MULTI-VIEW VIDEO

Felix Klose, Kai Ruhl, Christian Lipski, Christian Linz, Markus Magnor

Computer Graphics Lab, TU Braunschweig
Muehlenpfordtstrasse 23, 38106 Braunschweig, Germany
phone: +49 531 391 2102, fax: +49 531 391 2103, email: cg@tu-bs.de
web: http://graphics.tu-bs.de

ABSTRACT

We present an alternative approach to flexible stereoscopic 3D video content creation. To accomplish a natural image look without the need for expensive hardware or time consuming manual scene modeling, we employ an image-based free-viewpoint system to synthesize the stereoscopic views. By recording the sequence in a sparse multi-view setup, we are able to maintain control over camera position and timing as well as the parameters relevant for stereoscopic content. In particular, we are able to use the system to match camera path and timing of time lapsed background footage and a live-action foreground video.

1. INTRODUCTION

Looking at todays cinema, stereoscopic 3D movies have become common place. However the added degree of artistic freedom changes the traditional production pipeline. To achieve a pleasing stereoscopic 3D impression, it is necessary to validate view parameters such as plane of convergence and baseline throughout the entire production pipeline [1]. It may for example be necessary to re-render a scene to adapt the depth in respect to the surrounding footage in order to get the best possible viewer experience. While this is relatively easy for CG animation movies, changing the recording parameters after the shoot in a live-action scenario can be extremely costly or even impossible.

Free-viewpoint video (FVV) systems try to retain some of the degrees of freedom until the post-production stage, or even until viewing time. Generally the term FVV stands for the ability to render virtual new camera views after the recording has been completed. Based on the underlying scene model, different classes of those systems exist. We demonstrate the capabilities of a purely image-based FVV system for the creation of stereoscopic 3D content. The inherent flexible choice of viewing direction creates the ability to synthesize stereoscopic renderings.

In addition, we are also able match the camera path and timing across multiple recordings, where the recording time modalities vary drastically.

We give a short overview of relevant methods and current research in Sect. 2, followed by a description of our recording setup, Sect. 3. We then discuss the preprocessing of the input data in Sect. 4 and the image formation method used, Sect. 5. Finally, in Sect. 6 we show results that demonstrate the post-production flexibility.

2. RELATED WORK

Since its inception in 1838, stereoscopy has been widely used in the photography and film making industry. Recently, it has received renewed attention because the necessary equipment has reached technical maturity, allowing both stereoscopic recording and playback within reasonable constraints. Although the basic principle of stereoscopic image acquisition seems simple, many pitfalls exist that can make editing of stereoscopic material a tedious task [2].

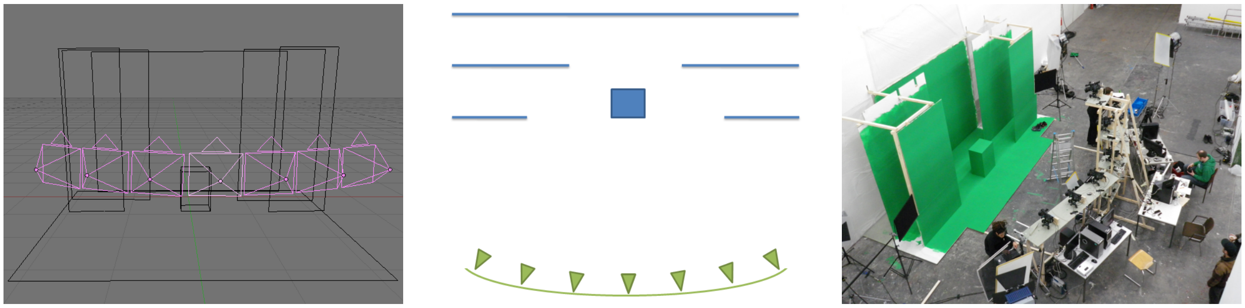
Typical stereoscopic editing tasks are image rectification, color balancing and baseline editing. In order to perform these tasks, it is desirable to keep control over as many camera parameters as possible even during post-production.

We focus on free-viewpoint video systems to provide the post-production camera control needed for stereoscopic content creation. Differentiated by the underlying models, two major groups can be identified: Those based on a geometric scene representation, and purely image-based systems.

The first category relies on a geometric reconstruction of the scene. Although stereoscopic rendering is straightforward if the scene geometry is known, they suffer from typical drawbacks of geometry-based systems: Zitnick et al. [3] presented a system for view interpolation from synchronously captured multi-view video data. Unfortunately, time interpolation is not possible with their approach and cameras have to be densely spaced. De Aguiar et al. [4] presented a high-quality performance-capturing that requires the exact knowledge of the 3D geometry of the performing actor. Zhang et al. [5] use a moving camera as an equivalent to a multi-view setup to infer depth maps. Their focus is on synthesizing new views close to the original camera path, rather than a wide set of novel views. Zhou et al. [6] recover depth maps from unsynchronized views by first synchronizing the images and then applying a stereo algorithm. Eisemann et al. [7] showed that misaligned reprojected textures can be corrected on-the-fly with image-based techniques; however they assume that the overall object silhouette is faithfully preserved. To improve the interpolation results for dynamic scenes, scene flow algorithms try to reconstruct the object movement [8]. Klose et al. [9] designed a scene flow reconstruction that is able to cope with unsynchronized multi-view recordings. However, their estimation provides only quasi-dense information and does not recover a valid model in poorly textured regions.

The second category is entirely image-based, avoiding explicit geometry and instead focusing on dense correspondence fields and image morphing between two or more images.

Traditionally, the correspondence fields are created in a user-assisted workflow [10] or are derived from other data,



(a) (b) (c)

Figure 1: Recording setup. (a) Camera array seen from behind (b) Floor plan: 3-layer background wall and desk cube, seen from above (c) Physical recording setup.

such as depth maps [11, 12]. The automatic estimation of these fields can be accomplished with optical flow algorithms. A survey on recent optical flow algorithm was composed by Baker et al. [13]. Special optical flow approaches have been proposed which are tailored for the task of image morphing. Stich et al. [14] designed a perceptually inspired optical flow algorithm for view interpolation. Lipski et al. [15] introduced representative SIFT descriptors for high-resolution image correspondence estimation and Linz et al. [16] combined the two latter approaches with a gradient-domain based rendering [17].

The correspondence estimation is followed by image morphing, a technique that accepts at least two input images and lets the user create a visually plausible in-between representation, using the dense pixel correspondences.

Seitz and Dyer [18] extended the original forward-warping and blending technique to produce geometrically plausible results. We employ their proposed image reprojection to align our input data and to produce the desired output, i.e., parallel or converging stereoscopic views.

Several more image-based free-viewpoint approaches exist: Germann et al. [19] represent soccer players as a collection of articulated billboards. Their approach is restricted to scenes with background that have known geometry and color distributions, e.g., soccer stadiums. Ballan et al. [20] presented an image-based view interpolation that uses billboards to represent a moving actor. Although they produce good results for a single rendered viewport, their billboard-based technique is not suitable for stereoscopic rendering since the foreground billboard would be identical for the stereoscopic image pair. Further more their approach focuses on plausible transitions between cameras and not on rendering virtual viewpoints. Lipski et al. proposed an image-based free-viewpoint system [21] that is based upon multi-image-morphing. They accept unsynchronized multi-view recordings as their input and are able to control both spatial and temporal camera parameters.

3. EXPERIMENTAL SETUP

We use a multi-view setup to record separate foreground and background scenes. First, the constructed set serves as a green-screen for the performance of the actors. These live-action elements are later time-resampled and then feature both slow-motion as well as fast-forward effects. Afterwards, for artistic reasons, the background is painted with different graffiti motifs while being recorded in stop motion over the course of several days (Fig. 2, 3). As a result, we have back- and foreground material with varying time sampling. Even in such a scenario, the free-viewpoint video approach allows us to composite both layers with a common virtual camera path in post production; this would not be feasible with a single moving camera. The seven cameras used for recording are Canon XHA1 set up in a horizontal arc as shown in Fig. 1(a), spaced approx. 1m and 10° apart. Since the consumer-grade camera does not feature synchronization, the image interpolation later on takes place both in the spatial and temporal domain at once.

Up to 4 cameras are connected to one PC via Firewire. The entire setup can be triggered to record a continuous video or a single snapshot. To facilitate the trigger mechanism we fitted a customized version of the Firewire recording software dvgrab [22] with network-control capabilities.

Unfortunately, the cameras' Firewire protocol does not support frame accurate start of recordings. To obtain an approximate temporal alignment, which is later refined during post production, the internal PC clocks are synchronized with local NTP and the video streams are marked with timestamps. To follow normal actor movements, this accuracy is sufficient.

The set is lit with studio spots and fluorescent lights all running with the power grid frequency of 50 Hz. Since the foreground action has fast moving objects such as flying hair, and sharp images are preferred for automated image correspondence estimation, we use a shutter time of 1/400s. Although we have a constant lighting situation, the combination of short shutter times, the power grid frequency being a multiple of the 25 fps camera framerate, and manufacturing tolerances leads to visible brightness fluctuations within a video. These have to be considered and corrected in a post processing step. For background snapshots, the scene is static and movement is not an issue, therefore we use a 1/25s shutter to integrate over the radiance variation.

All cameras are white balanced equally and the camera recording options such as focus and gain control are set to manual. Still, due to hardware manufacturing tolerances, the resulting color alignment between cameras is unsatisfactory and makes inter-camera color correction necessary.

4. DATA PREPROCESSING

Our data preprocessing consists of camera alignment, color correction, and dense correspondence estimation for all fore- and background material.

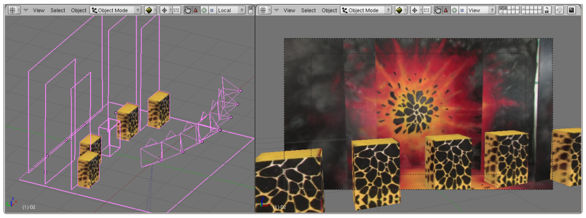


Figure 2: Camera position and sparse geometry are imported to 3D modeling tools (Blender). Using standard modeling tools, simple geometries can be reconstructed (background planes). Additional objects can be inserted (e.g., four clones of central box) and composited with free-viewpoint footage (right).

4.1 Camera Alignment

An offline processing estimates the in- and extrinsic camera parameters including acquisition time.

Camera positions are estimated with sparse bundle adjustment [23]. Image rectification, e.g. to correct keystoning, is postponed to the rendering stage in order to minimize the number of intermediary image resamplings.

To be able to use the unsynchronized camera streams for viewpoint interpolation, it is necessary to determine the sub frame accurate acquisition time. Starting from the approximate video timestamps Meyer et al. [24] present a image based method for temporal alignment.

The extrinsic camera parameters in conjunction with the image acquisition time form a high dimensional space. Our choice for an embedding into a parameter space is a spherical parameterization of the camera setup.

We define a navigation space that represents spatial camera coordinates as well as the temporal dimension. Cameras are placed on the surface of a virtual sphere, their orientations are defined by azimuth φ and elevation θ. For the concrete camera setup presented here, where only an arc of cameras is present, we further reduce the camera position to a two dimensional function of φ and t. A novel view $I(\varphi,t)$ is a function of those two parameters.

To map the euclidean camera positions to the two parameter representation, the spherical coordinates of the camera position are determined, where the sphere center $\mathbf{p}_S$ and the radius of the sphere r_S are computed from the extrinsic camera parameters $\mathbf{R}$ and $\mathbf{p}$ in a least-squares sense. The third parameter from the spherical parametrization is fixed for the camera arc setup, leaving the two final dimensions φ and t.

4.2 Color Correction

The recording setup and lighting conditions make color correction for the multi-view recordings necessary. Both intra-camera correction in time and inter-camera correction in space have to be performed. We approach both at the same time to achieve globally consistent color characteristics.

Simple adjustment of brightness over all cameras, as well as transfer of color statistics in $l\alpha\beta$ color space as proposed by Reinhard et al. [25], did not yield satisfactory results. In particular, visible color shifts between the cameras remained.

With the dense correspondence fields for our multi-view setup known, we can determine which image positions in two images should have similar colors. We select a target camera to which all colors should be adjusted. Given the color vector $\mathbf{x} = (r,g,b,1)^T$ in camera i and the color of the same point in the target camera $\mathbf{x}'$, we determine a linear color space transformation matrix $\mathbf{M}_i \in \mathbb{R}^{3\times4}$ such that

$$\mathbf{M}_i \cdot \mathbf{x} = \mathbf{x}' \tag{1}$$

for all color vectors $\mathbf{x}$ of the source image. The equation system is solved for all positions in a least-squares sense and the resulting transformation is applied to the source image.

The results are visually convincing. However in some cases, when the input images have too few colors, the degeneration of the equation system leads to some remaining artifacts. To avoid these problems, we chose motifs with rich colors for calibration. Further improvements can be achieved by using a variational approach that adaptively balances differences in color characteristics with differences in the image gradient, as proposed by Pitié et al. [26].

To account for the brightness fluctuation in time caused by the fluorescent lighting, a simple linear brightness adjustment within a camera stream is sufficient. We select a background region and constrain the brightness to be constant over time within this region. The rest of the frame is then adjusted accordingly.

4.3 Background Correspondence Fields

Our set background consists of a back wall, four chipboard walls in front and a desk-like cube in the center, see Fig. 1. In the course of several days, different graffiti motifs have been painted onto the entire set. The timelapsed recordings are then time-resampled in a non-linear fashion.

Along with the estimated camera parameters, structure-from-motion calibration [23] yields a sparse set of reconstructed scene points. After importing the resulting data into a modeling software (e.g. Blender), we can easily align the scene geometry, see. Fig. 2. This is done by manually fitting planes to the sparse point cloud. We use the proxy geometry to generate dense correspondence maps. Additionally, when the camera calibration is imported into the modeling tool, new geometry can easily be added to the scene. E.g. in Fig. 2, we apply projective texturing to the box in the center and duplicate it several times.

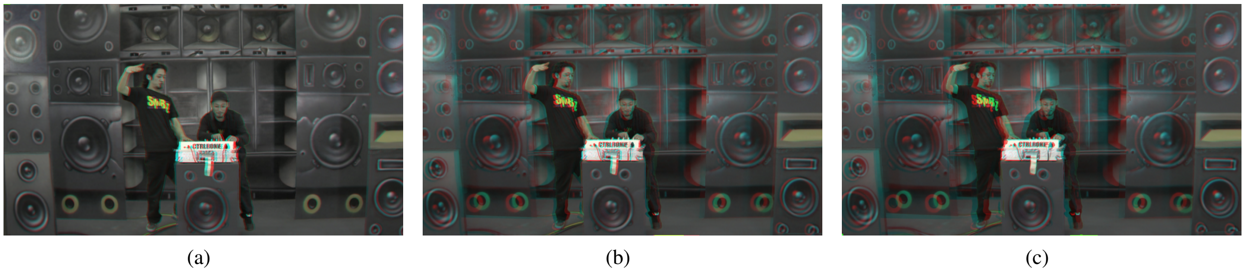
For more complex backgrounds it is possible to automatically approximate the geometry with planes [27] or use a multi-view reconstruction algorithm.

4.4 Foreground Correspondence Fields

The foreground recordings are dynamic in nature and contain fine details like hair, which are traditionally difficult for image interpolation. Additionally, the distance between corresponding pixels in image-space poses a challenge. Therefore, we opt for the belief propagation approach by Lipski et al. [15].

Occlusions and disocclusions are estimated with a symmetric term in the optical flow, which considers correspondence fields $\mathbf{w}_{ij}$ and $\mathbf{w}_{ji}$ at the same time, concluding that undetermined correspondences are most likely caused by disocclusions.

The horizontal component in the correspondence field can also be used to estimate the maximum tolerable stereoscopic ocular baseline for a given pair of images.


(a)

(b)

(c)

Figure 3: Anaglyphs with different baselines. (a) Composited image with 1.0 degrees baseline and zero parallax (b) Background with 2.0, foreground with 3.0 degrees baseline, parallax -16px (c) Background with 4.0, foreground with 3.0 degrees baseline, parallax -22px

5. STEREOSCOPIC VIRTUAL VIEW SYNTHESIS

In this section, we recapitulate image-based virtual view synthesis and show how to use them to synthesize a stereoscopic image pair.

Following Lipski et al. [21], we synthesize the camera view $I(\varphi,t)$ for every point inside the recording hull by multi-image interpolation:

$$I(\varphi,t)=\sum_{i=1}^{3}\mu_i\tilde{I}_i, \tag{2}$$

where μ is the relative weight of each input image and

$$\tilde{I}_i\left(\Pi_i\mathbf{x}+\sum_{j=1,\dots,3,j\neq i}\mu_j(\Pi_j(\mathbf{x}+\mathbf{w}_{ij}(\mathbf{x}))-\Pi_i\mathbf{x})\right)=I_i(\mathbf{x}) \tag{3}$$

are the forward-warped images [28]. The set of re-projection matrices $\{\Pi_i\}$ map each image I_i onto the image plane of $I(\varphi,t)$, as proposed by Seitz and Dyer [18]. Those matrices can be easily derived from camera calibration. Since the virtual image $I(\varphi,t)$ is always oriented towards the center of the scene, this re-projection corrects the skew of optical axes potentially introduced by our loose camera setup. Image reprojection is done on the GPU without image data resampling.

Using this notation, the stereoscopic image pair can be synthesized by offsetting the camera position along the φ-axis. The offset Δ between the views for the left $I^L(\varphi-\frac{\Delta}{2},t)$ and right $I^R(\varphi+\frac{\Delta}{2},t)$ eye is the camera baseline.

A common rule for stereoscopic capture is that the maximum angle of divergence between the stereo camera axes should not exceed 1.5 degrees. Beyond that, the eyes are forced to diverge to bring distant objects in alignment, which usually causes discomfort. In our approach, we choose to render converging stereo pairs with varying baselines between 0.5 and 4.0 degrees, and compare the results with an initial baseline of 1.0 degrees, which is an estimate for the most pleasing stereoscopic setting.

6. RESULTS

The results in this section are presented in red (left) - cyan (right) anaglyph images, as shown in Fig. 3. The scene shows composited foreground and background. Using the original images as stereo pair is not feasible, since the baseline (ocular disparity) would be far too wide. In contrast, the final render results for the left and right stereo image are quite close. As outlined before, our recording setup and the construction of the navigation space sets the point of convergence to the scene center.

An initial stereoscopic setting is shown in Fig. 3(a). Using a 1.0 degrees baseline, this is a pleasant setting leading the attention to the center of the image plane.

In addition to the usual adjustments to baseline and parallax, we can now experiment with different settings for background and foreground. Fig. 3(b) shows a composition where the foreground has a greater baseline than the background (3.0 vs. 2.0 degrees). In effect, the actors seem to float to the front, although the horizontal window violation at the bottom hampers a part of the effect.

Conversely, Fig. 3(c) depicts a situation where the background baseline is greater than the foreground's (4.0 vs. 3.0 degrees). Knowing that the actors are subjected to higher disparity due to being in front of the background, we can increase the background baseline further without suffering adverse effects. As intended, the actors still appear viewable, and the background exhibits much improved depth.

While in general stereoscopy enhances the authenticity of the scene, there are some limitations to our work. First, when interpolating in time, we are restricted to time steps that are small enough to approximate non-linear motion if there is any (e.g. a person on the highest point of a jumping motion). This effect tends to be less noticeable in the monocular case. Second, we are limited in the angular placement of the cameras. Increasing the inter-camera spacing beyond 15 degrees dilates the results considerably. This is similar to geometric approaches, where occlusions and large variations in appearance make automatic reconstruction infeasible. Finally, vertical disparity errors are more noticeable in stereoscopic mode. Our optical flow considers all vector directions equally, and does not account for this binocular property.

7. CONCLUSIONS AND FUTURE WORK

We presented an approach for stereoscopic free-viewpoint rendering that circumvents the need for explicit 3D reconstruction. It enables the flexible creation of stereoscopic content of complex natural scenes, where parameters as baseline, viewpoint and scene time can easily be modified in post production. In a single workflow, image alignment, free-viewpoint video and baseline editing can be performed.

Our approach can cope with asynchronously captured material and loosely calibrated camera setups, greatly reducing the hardware requirements needed for stereoscopic 3D recording. A small subset of special effects was demonstrated, focusing on diverging stereoscopic settings for background and foreground footage, and many further possible effects that integrate seamlessly are conceivable.

One future research direction could be to conceive an image-based editing framework, where the original footage could be manipulated with classic 2D tools, and the changes then propagated back to adjacent camera views.

8. ACKNOWLEDGEMENTS

Funding by the European Research Council ERC under contract No. 256941 "Reality CG" and by the German Science Foundation DFG MA 2555/1-3 is gratefully acknowledged.

REFERENCES

[1] Autodesk, "The Business and Technology of Stereoscopic Filmmaking," *Stereoscopic Filmmaking Whitepaper*, 2008.

[2] L. Wilkes, "The role of ocula in stereo post production," *The Foundry, Whitepaper*, 2009.

[3] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-Quality Video View Interpolation Using a Layered Representation," *ACM Trans. on Graphics*, vol. 23, no. 3, pp. 600–608, 2004.

[4] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance Capture from Sparse Multi-View Video," *ACM Trans. on Graphics*, vol. 27, no. 3, pp. 1–10, 2008.

[5] G. Zhang, Z. Dong, J. Jia, L. Wan, T.-T. Wong, and H. Bao, "Refilming with depth-inferred videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 828–840, 2009.

[6] C. Zhou and H. Tao, "Dynamic depth recovery from unsynchronized video streams," in *CVPR (2)*, pp. 351–358, 2003.

[7] M. Eisemann, B. D. Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, "Floating Textures," *Computer Graphics Forum (Proc. Eurographics EG'08)*, vol. 27, pp. 409–418, 4 2008.

[8] S. Vedula, S. Baker, and T. Kanade, "Image Based Spatio-Temporal Modeling and View Interpolation of Dynamic Events," *ACM Trans. on Graphics*, vol. 24, no. 2, pp. 240–261, 2005.

[9] F. Klose, C. Lipski, and M. Magnor, "Reconstructing shape and motion from asynchronous cameras," in *Proc. Vision, Modeling and Visualization (VMV) 2010*, pp. 171–177, 2010.

[10] G. Wolberg, "Image morphing: a survey," *The Visual Computer*, vol. 14, no. 8, pp. 360–372, 1998.

[11] S. E. Chen and L. Williams, "View interpolation for image synthesis," in *Proc. of ACM SIGGRAPH'93*, pp. 279–288, ACM Press/ACM SIGGRAPH, 1993.

[12] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel, "Adaptive image-space stereo view synthesis," in *Vision, Modeling and Visualization Workshop*, (Siegen, Germany), pp. 299–306, 2010.

[13] S. Baker, S. Roth, D. Scharstein, M. J. Black, J. Lewis, and R. Szeliski, "A database and evaluation methodology for optical flow," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, IEEE Computer Society, 2007.

[14] T. Stich, C. Linz, C. Wallraven, D. Cunningham, and M. Magnor, "Perception-motivated Interpolation of Image Sequences," *ACM Transactions on Applied Perception (TAP)*, 2010. to appear.

[15] C. Lipski, C. Linz, T. Neumann, and M. Magnor, "High resolution image correspondences for video Post-Production," in *CVMP 2010*, (London), pp. 33–39, 2010.

[16] C. Linz, C. Lipski, and M. Magnor, "Multi-image Interpolation based on Graph-Cuts and Symmetric Optic Flow," in *15th International Workshop on Vision, Modeling and Visualization (VMV)* (C. R.-S. Reinhard Koch, Andreas Kolb, ed.), pp. 115–122, Eurographics, Eurographics Association, November 2010.

[17] D. Mahajan, F. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur, "Moving Gradients: A Path-Based Method for Plausible Image Interpolation," *ACM Trans. on Graphics*, vol. 28, no. 3, pp. 42:1–42:11, 2009.

[18] S. M. Seitz and C. R. Dyer, "View Morphing," in *Proc. of ACM SIGGRAPH'96*, pp. 21–30, ACM Press/ACM SIGGRAPH, 1996.

[19] M. Germann, A. Hornung, R. Keiser, R. Ziegler, S. Würmlin, and M. Gross, "Articulated billboards for video-based rendering," *Comput. Graphics Forum (Proc. Eurographics)*, vol. 29, no. 2, p. 585, 2010.

[20] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys, "Unstructured video-based rendering: Interactive exploration of casually captured videos," *ACM Trans. on Graphics (Proc. SIGGRAPH)*, vol. 29, July 2010.

[21] C. Lipski, C. Linz, K. Berger, A. Sellent, and M. Magnor, "Virtual video camera: Image-based viewpoint navigation through space and time," *Computer Graphics Forum*, vol. 29, no. 8, pp. 2555–2568, 2010.

[22] A. Schirmacher, D. Dennedy, and D. Streetman, "dvgrab." http://www.kinodv.org/, 2007.

[23] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3d," *ACM Trans. on Graphics*, vol. 25, no. 3, pp. 835–846, 2006.

[24] B. Meyer, T. Stich, M. Magnor, and M. Pollefeys, "Subframe Temporal Alignment of Non-Stationary Cameras," in *Proc. British Machine Vision Conference BMVC '08*, 2008.

[25] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *Computer Graphics and Applications, IEEE*, vol. 21, no. 5, pp. 34 –41, 2001.

[26] F. Pitié, A. C. Kokaram, and R. Dahyot, "Automated colour grading using colour distribution transfer," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 123 – 137, 2007. Special issue on color image processing.

[27] C. Schwartz, R. Schnabel, P. Degener, and R. Klein, "Photopath: Single image path depictions from multiple photographs," *Journal of WSCG*, vol. 18, Feb. 2010.

[28] W. Mark, L. McMillan, and G. Bishop, "Post-Rendering 3D Warping," in *Proc. of Symposium on Interactive 3D Graphics*, pp. 7–16, 1997.