DEPTH IMAGE BASED COMPOSITING FOR STEREO 3D

Lars Schnyder, Manuel Lang, Oliver Wang, Aljoscha Smolic

Disney Research Zurich,
Zurich, Switzerland

ABSTRACT

Compositing is an essential tool in modern film making. Established workflows exist for single-view video compositing, however new problems that demand new solutions arise when considering stereoscopic compositing. We investigate new methods for compositing live action stereo 3D, given two stereo camera systems and a corresponding depth map. We break the process up into three main steps; First, we show how to use trifocal tensors to robustly project 3D content from one video into another. Second, we analyze different image-based rendering methods for drawing objects to be composited into the new video. Finally, we describe a novel super-pixel based depth-test that increases robustness to errors in depth map accuracy. A user study was conducted to validate different steps of this process.

Index Terms — Image fusion, Stereo image processing

1. INTRODUCTION

Stereoscopic 3D is an important focus of the entertainment industry, creating new possibilities and new challenges. Despite its many advances, producing stereo content remains significantly more restrictive than traditional monoscopic content [1]. These difficulties arise due to the tight relationship between stereo views that need to be correctly maintained during editing operations. In this work, we address the task of compositing, combining objects from different stereo sources into one single stereo output. We want to present a method that makes this process more robust and therefore also cheaper in production. The lack of an existing widespread stereoscopic editing pipeline is the fundamental motivation for our work.

Stereo compositing requires projecting pixels from the *source* stereo pair into a virtual stereo view, as defined by the *target* pair. This projection requires knowledge of the geometric relation between stereo cameras, and the object depths, which we assume to be given beforehand, in the form of a depth map. While there exist numerous methods for computing depth maps, it is an under-constrained problem, and state of the art methods will contain errors. Our system is therefore designed to generate convincing results even in the presence of moderate errors in depth-maps.

Our method can be broken up into three steps. First, we transfer pixels belonging to an object from one stereo camera pair into a second camera pair, through a new application of the trifocal tensor. Second, we draw the object using these projected points to guide our rendering. Without a full 3D model, image-based approaches can only approximate this new view, and there will be a trade-off between distortion and accuracy. Therefore, we conduct a small user study to evaluate a set of different potential fitting approaches. Finally, when compositing the new footage, a novel super-pixel based segmentation method is proposed to enforce similar depth tests of local regions.

2. RELATED WORK

Stereo cinematography is a well-studied area; the challenges of current stereoscopic and 3D video post-productions was presented by Starck et al. [2]. Smolic et al. [3] gave an additional overview of the state of current 3D production research.

"Stereoscopic 3D Copy&Paste" shown in [4] represents a system that projects object billboards into new views. Another layered approach to stereo reconstruction [5] divides the depth map into multiple approximately planar layers. They are limited to billboard representations of objects, which can fail in examples with strong perspective change. We explore a more flexible image-warping approach for computing perspective deformations.

Free-viewpoint depth image based rendering is a common approach t project information into new virtual views [6, 7, 8]. Our method performs a similar task, but uses the trifocal transfer in cooperation with image-based rendering. With this, we avoid the direct computation of 3D points, which is often performed by other methods but usually inaccurate due to depth map errors.

For monoscopic images, recoloring and relighting for seamless compositing [9, 10] have been well studied. However, these methods create unpredictable results for objects with tight borders and can lead to color bleeding artifacts. We address a simpler problem where alpha mattes are available through blue- and green-screen methods well established in industry.

The composition of different masked objects was proposed by [11] and uses robust handling of layers defined by user ordering. We present an application tailored for interactive systems where the layering is not yet clear and is determined by object depth. Further, we introduce an automatic super-pixel based layering approach to enforce local consistency, without requiring a higher level object segmentation.

3. METHOD

Traditional DIBR based methods operate by projecting and re-projecting pixels to 3D world space, creating a pixel cloud. However doing so requires calibrated cameras as well as accurate depth maps and complex hole-filling strategies due to disocclusions.

In the following, we present our image domain re-projection method that does not require the *full* camera calibration. We first transfer a subset of stable image features from the source camera system to the target camera system. We call them "guidance points". These are then used to optimize for a dense 2D image deformation that maps the source compositing object into the target footage.

3.1. Guidance point transfer

Epipolar transfer [12] is often used to compute pixel coordinates in a novel view. However, this approach exhibits numerical instability for (common) stereo setups with parallel or near parallel cameras. Instead, we perform a trifocal point transfer from a given source stereo camera pair to one novel new view.

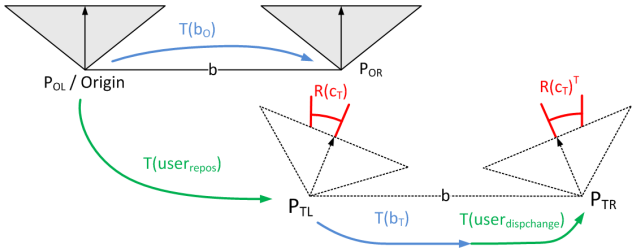


Figure 1. Visualization of all the system-related transformations for the camera matrix calculation. The user is allowed to define object positioning $user_{repos}$ and object depth $user_{dispchange}$.

Relative camera matrices We first describe the camera matrices of the system (Figure 1), assuming the origin of the coordinate systems to be set to the left camera of the source camera pair. The left source camera matrix is defined by rotation and projection, the right by translation, rotation and projection, as shown in Figure 1.

$$\begin{aligned} P_{OL} &= P(f_O)R(c_O), \\ P_{OR} &= P(f_O)R(c_O)^T T(b_O). \end{aligned} \tag{1}$$

Part of a compositing operation is a user-controlled repositioning and scaling of the different objects. All of these possible editing operations can be modeled as a movement of the target camera system pair (e.g. moving an object left is equivalent to moving the virtual camera right). Similarly, depth scaling can be achieved by changing the target baseline. The target camera matrices are then dependent on user input and can be defined as:

$$\begin{aligned} P_{TL} =& P(f_T)R(c_T)T(user_{repos}), \\ P_{TR} =& P(f_T)R(c_T)^T T(b_T)T(user_{dispchange}) \\ & T(user_{repos}). \end{aligned} \tag{2}$$

For this method, only the baseline b, the converging angle c and the focal length f have to be known, rather than the full camera calibration matrices.

Normalized Camera Matrices In [12], it was shown that, for simplicity, the point transfer should be computed in a canonical/normalized setup. For this, we transform the previously created matrices of when we project points. For the left output P_{TL} we multiply all matrices with the pseudo-inverse H_{OL} of P_{OL}. This yields $P'_{OL} = [\, I \mid 0 \,]$, $P'_{OR} = P_{OR}H_{OL}$, $P'_{TL} = P_{TL}H_{OL}$.

Computing the Trifocal Tensor As shown in [12], the trifocal tensor for the left output view is then defined as:

$$T^{(k,j,i)} = P'^{(j,i)}_{OR} * P'^{(k,3)}_{TL} - P'^{(j,3)}_{OR} * P'^{(k,i)}_{TL}. \tag{3}$$

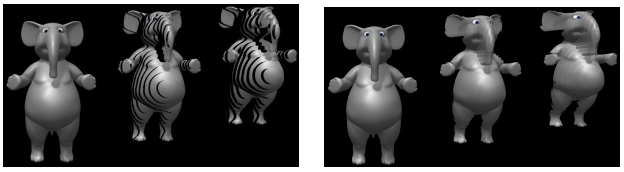
Point Transfer Finally, we can transfer the guidance points p from source to the target system. For transferring left source points into the left target view, we take the following steps:

(i) Having $P_{OL} = [\, I \mid 0 \,]$ and $P_{OR} = [\, M \mid m \,]$ we calculate the fundamental matrix with $F_{LR} = [\, m \,]_\times M$. $[\, m \,]_\times$ is the sqew-symmetric matrix calculated from m [12].

(ii) For every point p_{OL}, compute the line l' which goes through p_{OR} and is perpendicular to $l'_e = F_{LR}p_{OL}$. If $l'_e = (l_1, l_2, l_3)$ and $x' = (x_1, x_2, 1)$ then $l' = (l_2, -l_1, -x_1l_2 + x_2l_1)$. One could also directly use $l' = (1, 0, -x_1)$ for nearly parallel systems. This l' is a vertical line through point x'.

(iii) Get the transferred point by $p^k_{TL} = \sum_{ij} p^i_{OL} l'_j T^{jk}_i$.

This gives us a sparse set of *guidance points* in the target view, which can be used to drive an interpolation of all other pixels.



(a) DIBR point *(b) DIBR mesh*

Figure 2. (DIBR) image transfers done with the point-based method (a) and the mesh method (b). As expected, we result in disoccluded regions without image information. These points represent the guidance points.

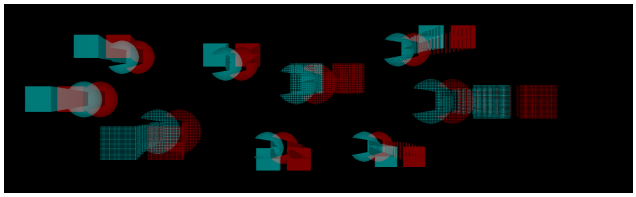
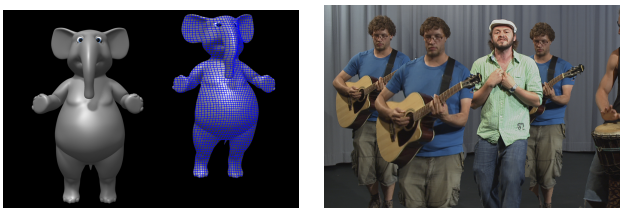


Figure 3. In this image we see the effects of perspective change on the original pixel cloud of the left source object. As a result, we see point overlaps and disocclusions at the target location.



(a) Elephant with grid *(b) Warped band*

Figure 4. Applying the vision-aware warping method to the elephant example (a) and two copies of the guitar player (b). Blurriness comes from interpolation due to magnification of the object (parts).

3.2. Rendering

To accomplish this interpolation, we explore multiple image-based approaches: point- and mesh-based rendering, warping and fitting matrix transformations [10].

Direct Pixel Transfer A simple naive pixel-rendering approach with (Figure 2(b)) and without (Figure 2(a)) an interpolation mesh, provides accurate point locations, but holes or distortion at disocclusions. Subsequent methods analyze the trade-off between the correct projective result (Figure 3) and these local distortions.

Fitting By fitting the initial image to these points with a reduced degree of freedom, we can ensure that the warped objects appear realistic, and reduce the effect of noise in the projected points. We analyzed deformation models with various degrees of freedom; uniform and non-uniform translation in x- and y-direction plus scaling, affine and finally perspective transformations. This fitting is computed as a least square solution x, with $\|x\| = 1$, of a homogeneous system similar to $Mx = 0$ by using the SVD ($U\Sigma V^T$) of M. We therefore use RANSAC to find the best of multiple solutions from random point subsets.

Warping In addition to low-degree of freedom models, we can compute a mesh-warping approach to fit the image to the guidance points, using visual saliency to hide distortions [13, 14]. The warp is a mesh deformation densely mapping input points to a new output destination. It is computed by minimizing smoothness constraints that reduces overall distortion and overlaps, while en-

(a) Depth map inaccuracies

(b) Band

Figure 5. Pixel-based compositing errors due to depth maps are shown in image (a). Figure (b) shows the application of SLIC super-pixel segmentation to one of our test images.

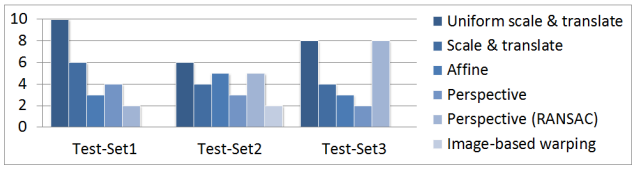


Figure 6. The three test sets evaluated by the 25 participants. Brighter colors represent methods with more degrees of freedom. The x-axis represents the three test sets and y-axis the amount of participants which voted for this method. Higher scores are better.

forcing a data constraint that maps guidance points to their target location ((Figure 4). Please see [13] for an extended description.

3.3. Compositing

Finally, we composite the transferred image in the virtual camera with the target content. To do this, we need to update the disparity map to match the virtual view, before we can apply a depth z-test.

For the specific guidance points, we know exactly how the disparity is changing during transfer, and for in between points, we can interpolate disparities based on the fitting approach used. This is done by updating the source disparity map with target disparities, followed by the application of the transformation already used on the image content. In this process, disparities are automatically interpolated between the target pixel positions.

Super-Pixel Z-Test Real world disparity maps are often inaccurate and their direct use for z-test can introduce notable artifacts (Figure 5(a)). We therefore present a super-pixel approach to decrease the compositing's sensitivity to noise by enforcing neighborhood similarity similar to cross-bilateral filtering. In images, object borders are not discrete and usually blur over a few pixels. Additionally, they do not always correspond to object edges in the depth map. Using a mixture of image and depth edges, we extract super-pixels [15] (Figure 5(b)) for the object and target image. In contrast to other filtering, we then assign a super-pixel its underlying median depth and require that an entire super-pixel is classified as either "over" or "under" the target stereo footage during the z-test. In addition, we enforce corresponding super-pixels in the left and right image have the same "over" or "under" property.

3.4. User Study

We conducted a user study to analyze the perceptual trade-off between image distortion artifacts and adherence to the perspective transformation. In particular, we were interested which image domain deformation for fitting the guidance points gives best compositing impression for viewers. This user study was composed of 25 individuals, 5 of whom were very familiar with 3D content, and 20 of whom had some experience with 3D. The first group evaluated the results on an interlaced 3D display, while the second group used anaglyph glasses on corresponding anaglyph images. One participant was removed due to lack of 3D vision.

In the user study, three test sets were used (Figure 6). In each set, participants were able to choose their best compositing solution of all the methods while not having any reference to the source object for comparison. Amongst other things, this is why participants were generally less influenced by missing perspective changes in the projection. Although, the composited objects in test sets did not show very strong distortions, already small unnatural distortion artifacts caused warping to only score two votes.

4. RESULTS

Figure 7 shows the application of different fitting methods on objects. In these visualizations, the distortion introduced due to a strong perspective change of the object is clearly shown. The point- and mesh-based DIBR transfers provide an approximation of the optimal distortion, but with holes. Only the warping delivers a result similar to the mesh-based DIBR rendering.

Disparity noise influences the performance of the point and image transfer method as well as the compositing. The unprocessed results show that super-pixels can counter a fair amount of noise in image regions and borders (Figure 8).

5. CONCLUSIONS AND FUTURE WORK

We have presented a pipeline for depth image based stereoscopic compositing. We introduced a novel use of the trifocal tensor for robust object mapping, analyzed a set of rendering approaches, and proposed a final robust compositing technique. All of these together form a fully functional stereoscopic compositing workflow. However, we found that while disparity noise can be countered to some extent, occasionally depth prediction causes large regions to have incorrect disparity. These cases remain problems even with our techniques. Furthermore, if the perspective change is very strong or changes the silhouette of the object the warping approach can create strong unwanted artifacts.

In order to improve quality, additional lighting and color corrections as well as image-based shadows should be considered. Some artifacts in the rendering could be additionally reduced by using a discontinuous warping method [16]. Finally, we analyzed images but a video approach would be a logical next extension. Most work in this area goes into stabilizing the object's target position and fitting in the presence of errors.

6. REFERENCES

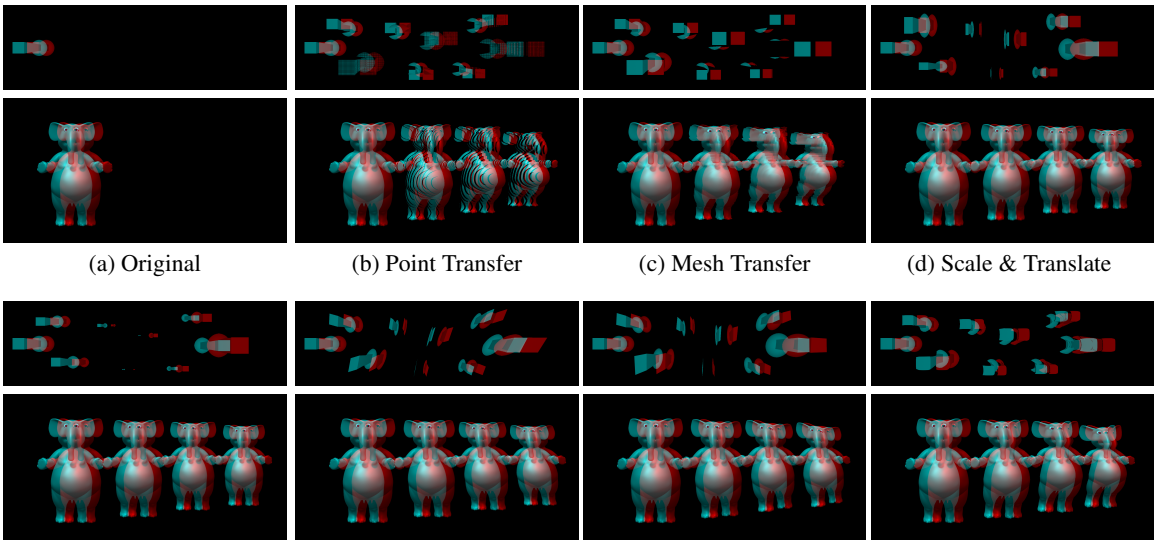
[1] F. Zilly, J. Kluger, and P. Kauff, "Production Rules for Stereo Acquisition," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 590–606, Apr. 2011.

[2] Jonathan Starck, "Stereo in post-production," in *Proceedings of the 1st international workshop on 3D video processing*. 2010, 3DVP '10, pp. 37–38, ACM.

[3] A. Smolic, S. Poulakos, S. Heinzle, P. Greisen, M. Lang, A. Hornung, M. Farre, N. Stefanoski, O. Wang, L. Schnyder, R. Monroy, and M. Gross, "Disparity-aware stereo 3d production tools," in *Visual Media Production (CVMP), 2011 Conference for*, nov. 2011, pp. 165 –173.

[4] WY. Lo, J. van Baar, C. Knaus, M. Zwicker, and M. Gross, "Stereoscopic 3d copy & paste," in *ACM SIGGRAPH Asia 2010 papers*. 2010, SIGGRAPH ASIA '10, pp. 147:1–147:10, ACM.

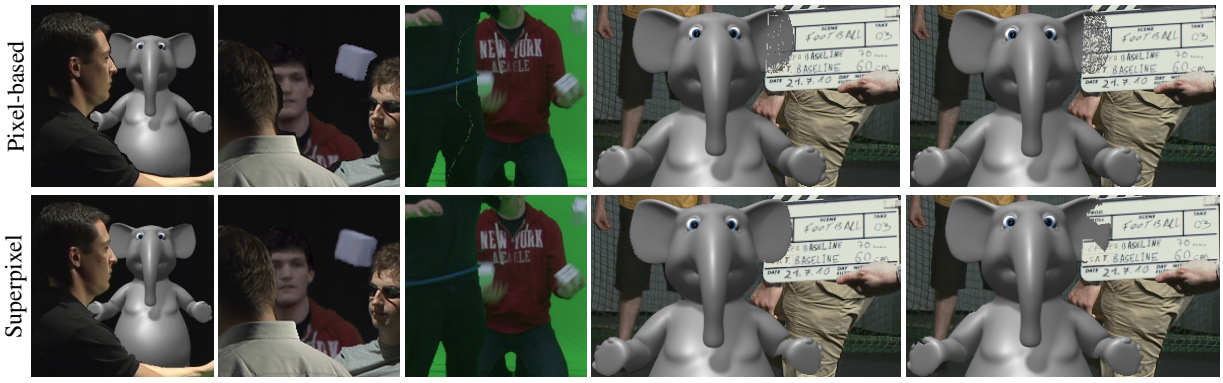
[5] S. Baker, R. Szeliski, and P. Anandan, "A layered approach to stereo reconstruction," *CVPR 1998*, pp. 434–441, 1998.



(a) Original (b) Point Transfer (c) Mesh Transfer (d) Scale & Translate

(e) Uniform Scale & Translate (f) Affine (g) Perspective (h) Warping

Figure 7. All fitting methods under the effect of strong perspective change of the object are shown in images (a) to (h). We chose artificially created objects with a large depth volume. This intentionally increases the effect and improves the visibility of each method's advantages and disadvantages. The images of (b) and (c) can be used as ground truth approximations for the image transfers of (d) to (h).



a b c d e

Figure 8. Super-pixels are used to enforce spatial similarity in the presence of noisy depth maps. (a) (b) and (c) show common depth map errors around object borders that lead to incorrect depth tests. These are fixed in the superpixel approach. (d) shows an example of a noisy depth map that can be fixed by using a spatial similarity assumption, while (e) shows an example where the depth map is too incorrect to be corrected.

[6] C. Fehn, "Depth-Image-Based Rendering (DIBR), Compression and Transmission for a New Approach on 3D-TV," in *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI*, 2004, pp. 93–104.

[7] S. Zinger, L. Do, and P. H. N. de With, "Free-viewpoint depth image based rendering," *J. Vis. Comun. Image Represent.*, vol. 21, no. 5-6, pp. 533–541, July 2010.

[8] L. Do, S. Zinger, and P. H. N. de With, "Quality improving techniques for free-viewpoint dibr," in *IST / SPIE Electronic Imaging*, 2010, p. 10 pages.

[9] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*. 2003, SIGGRAPH '03, pp. 313–318, ACM.

[10] Z. Farbman, G. Hoffer, Y. Lipman, D. Cohen-Or, and D. Lischinski, "Coordinates for instant image cloning," in *ACM SIGGRAPH 2009 papers*. 2009, SIGGRAPH '09, pp. 67:1–67:9, ACM.

[11] J. McCann and Pollard N, "Local layering," in *ACM SIGGRAPH 2009 papers*. 2009, SIGGRAPH '09, pp. 84:1–84:7, ACM.

[12] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[13] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3d," in *ACM SIGGRAPH 2010 papers*. 2010, SIGGRAPH '10, pp. 75:1–75:10, ACM.

[14] A. Goldstein and R. Fattal, "Video stabilization using epipolar geometry," *ACM Trans. Graph.*, vol. 32, no. 5, 2012.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels," Tech. Rep., 2010.

[16] O. Wang, M. Lang, M. Frei, A. Hornung, A. Smolic, and M. Gross, "Stereobrush: interactive 2d to 3d conversion using discontinuous warps," 2011, SBIM '11, pp. 47–54, ACM.