CONSENSUS-BASED MULTIVIEW TEXTURING AND DEPTH-MAP COMPLETION

Kai Ide, Ivo Keller, Thomas Sikora

Communication Systems Group
Technische Universität Berlin
10587 Berlin, Germany

ABSTRACT

We describe an active depth imaging system based on phase measuring triangulation. Typically depth-maps generated with such 3D scanning systems suffer from occlusions and imperfections, especially in the vicinity of depth discontinuities. Applying multiple color images, captured with a camera array, for view synthesis from the erroneous depth-maps can result in severe texturing artifacts. Our consensus-based approach greatly reduces these artifacts by comparing the similarity of the multiview texture images during the blending process to detect outliers in the form of foreground texture projected on background surfaces and specular ambiguity. Additionally, the approach is applied to dramatically improve the depth-maps by generating multiple depth-map hypotheses and selecting the areas of each that have the highest consensus with the set of multiview texture images. Our approach yields accurate and occlusion-free depth-maps in real-time.

Index Terms— 3D Scanning, Structured Light, Free Viewpoint Video, Multiview, View Synthesis

1. INTRODUCTION

The television and movie industry has always striven to create a more realistic and more immersive experience by continuously improving the quality of video and sound. A few years ago digital capturing and digital display technologies have made high quality stereoscopic 3D possible at home and in the theaters. The next logical step are autostereoscopic multiview displays that do away with 3D glasses and avoid visual discomfort [1]. These displays require not just two but five to thirty input views, all with the highest degree of image alignment with respect to parallax and colorimetric properties. Next generation free viewpoint displays and holographic displays will even call for a number of input images in the range of thousands. These displays will create the impression of viewing a scene *through* an actual window and not *on* a flat display. While it is to a certain extend possible to directly capture a number of input views with a synchronized and properly aligned multi-camera array [2], holographic displays would require light field cameras [3] with a wide maximal camera baseline in order to allow viewers to immerse in a scene from a perspective of their choosing. Recording such imagery directly, quickly becomes infeasible.

This work demonstrates capturing entire 3D models of a scene at video frame rates. Sufficient quality provided, the 3D models, along with several synchronously captured color images, can serve as mediators to render any number of high quality views for arbitrary viewpoints, as suggested in [4]. For this reason, we have addressed two key challenges. The first being the removal of errors in the depth-maps while additionally filling all remaining occlusions within the depth-maps as illustrated in Fig. 3. Even after one or multiple depth-maps have been filtered, completed and combined to a geometric model, the resulting 3D data will most likely contain a multitude of small and large errors. Subsequent texturing with multiview color imagery will reflect these errors in the form of texture-artifacts. Our consensus-based texturing and depth-map completion method addresses both these problems.

2. RELATED WORK

The underlying capturing and 3D reconstruction system described in this paper is based on phase measuring triangulation (PMT). PMT falls within the category of structured light, an active stereo image matching technique that measures the deformation of a known light pattern after it is reflected from the scene and captured with a camera [5] [6] [7]. A concise summary of structured light patterns is given by Salvi et al. [8]. In contrast, passive stereo image matching techniques apart from Structure from Motion [9] or Depth from Defocus [10] approaches, require two or more cameras and rely solely on ambient light. Passive techniques suitable for real-time reconstruction at video frame rates include i.e. guided image filtering [11] and hybrid recursive matching [12]. A concise overview is given in [13] and most notably in the Middlebury dataset [14]. Development towards multiview reconstruction is accounted for in [15].

3D reconstruction via PMT has been demonstrated in [16]. An extension of the method to allow for motion compensation and the scanning of discontinuous objects has been demonstrated by Weise et al. [17]. A four pattern phase shift with an integrated binary coding scheme is applied by Wissmann et al. in [18].

3. CONSENSUS-BASED TEXTURING

Once a preliminary depth-map has been calculated (Fig. 1a) the resulting 3D model $\mathbf{M}$ can be textured. Texture images are uploaded to the GPU in a compressed raw Bayer eight bit per pixel format and are then converted to RGB on the GPU for fast processing [19]. The OpenGL / GLSL environment on the GPU allows for efficient shadow mapping (Fig. 1d), given $\mathbf{M}$, for each c-th texturing camera $c = [1, \ldots, N]$. Without shadow maps $\mathbf{s}$, direct projective texturing will fail and result in double or triple texturing, as indicated in Fig. 1c.

$$\mathbf{s} \in R^N \;\text{ with }\; s(c) = \begin{cases} 0: & \text{shadowed} \\ 1: & \text{visible} \end{cases} \tag{1}$$

Subsequently, shadowed regions are textured with one (Fig. 1e) or all remaining color images but erroneous depth discontinuities (Fig. 1b) result in texturing artifacts (Fig. 1f). Direct averaging of all overlapping textures can only conceal these errors partially (Fig. 1g). Thus, we calculate a consensus vector $\mathbf{a_c}$ to detect which texture is causing the error in a small window with $k = 2$ around each fragment

$$\mathbf{a}_c = s_c \sum_{\Delta x=-k}^{k} \sum_{\Delta y=-k}^{k} \mathbf{\Gamma}_c \left(\frac{x_c}{w_c} + \Delta x, \frac{y_c}{w_c} + \Delta y \right) \tag{2}$$

and an error metric b_c formulated by Eq. 3 and Eq. 4

$$b_c = s_c \left| \frac{1}{N} \sum_{c=1}^{N} (\mathbf{a}_c) - \sum_{\Delta x=-k}^{k} \sum_{\Delta y=-k}^{k} \mathbf{\Gamma}_c \right|, \tag{3}$$

$$\mathbf{\Gamma}_c(x, y) = [L(x, y), a(x, y), b(x, y), \mathbf{m}'(x, y)]', \tag{4}$$

where L, a, b refer to the CIE 1976 (L^*, a^*, b^*) color space transform of the RGB textures. $\mathbf{m}(x, y)$ denotes the 3D coordinate seen at the respective pixel locations of the c-th camera. b_c is low if colorimetric features $\mathbf{\Gamma}_{1:3}$ and the geometric features $\mathbf{\Gamma}_{4:6}$ align with the mean consensus in Eq. 2. The last three elements $\mathbf{\Gamma}_{4:6}$ assert that texels at depth discontinuities will be selected as outliers if the majority of texels target a geometrically smooth surface. The pixel-coordinates for each fragment in the c-th camera are given by

$$\begin{bmatrix} x_c \\ y_c \\ w_c \end{bmatrix} = \mathbf{P}_c \mathbf{m} \quad \forall\, \mathbf{m} \underbrace{\subset}_{\text{visible}} \mathbf{M}. \tag{5}$$

Here $\mathbf{P}_c$ is the c-th camera's projection matrix, with $\mathbf{P}_c = \mathbf{K}_c [\mathbf{R}_c | \mathbf{t}_c]$. $\mathbf{K}_c$ represents the respective camera's intrinsic matrix, $\mathbf{R}_c$ the rotation matrix and $\mathbf{t}_c$ the translation vector, while w_c is applied for coordinate normalization in Eq. 2. Geometrically, the cameras and projectors in the array are calibrated fully automatically in a common world coordinate system with the method we have described in [20].

Texels with a maximum $b_{c,max}$ are selected outliers, as shown in Fig. 1h, if $b_{c,max} > 2\sigma_b$, where σ_b is the standard deviation over all b_c. Texels of the remaining subset of cameras are then weighted based on the distances of their respective camera centers $\mathbf{C_c} = -\mathbf{R_c}'\mathbf{t_c}$ from the 3D coordinate $\mathbf{m}$ assigned to the fragment, yielding a far better blending result as shown in Fig. 1i.

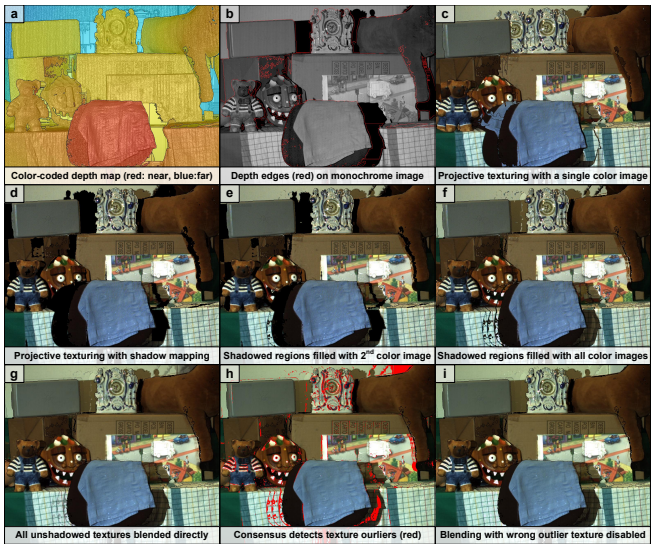

Fig. 1. Consensus-based texturing with multiview images and an imperfect depth-map.

4. DEPTH ESTIMATION AND DEPTH-MAP COMPLETION

Our 3D scanning system is based on phase measuring triangulation which allows for a good balance between geometric accuracy and data density. We sequentially project two phase shifted sinusoidal triplets – one with just one period and one with $N = 32$ periods. The later is converted to a modulo 2π wrapped phase Φ'_h. Absolute phase recovery is performed by utilizing the low frequency patterns for guided phase unwrapping with the low frequency phase Φ'_l. In practice it is beneficial to apply a denoising filter to Φ'_l to remove outliers. This is achieved by bilateral filtering of Φ'_l, as described by Shi and Tomasi in [21], and by removing pixels with large gradients in x or y. Epipolar constrained triangulation yields the final 3D reconstruction, while four color cameras provide additional texture as summarized in Fig. 2. For more details regarding the 3D reconstruction process, we refer to our previous work [22].

Once all camera pixels that capture light reflected from the scene by one of the projectors are filtered and converted to a 3D point cloud, the point cloud is converted to a mesh in the form of a continuous triangle strip in which three neighboring camera pixels containing vertices create a triangle if the longest edge of that given triangle is below a certain threshold, depending on the resolution and placement of the cam-

eras and the projectors. In our case this threshold is usually a few millimeters. The resulting mesh is re-converted to a depth-map (Fig. 3a) for each c-th camera with depth values $d_c(x, y)$ given by the distance of the visible 3D point $\mathbf{m}(x, y)$ from the respective camera center $\mathbf{C_c}$

$$d_c(x, y) = |\mathbf{C_c} - \mathbf{m}(x, y)|. \tag{6}$$

In the hole filling process we first approximate the correct depth discontinuities by slightly dilating foreground objects in a small window with $k = 1$, so that and $x', y' \in \{-k, 0, k\}$.

$$d_{c,min}\,(x', y') = \min(d_c(x + x', y + y')). \tag{7}$$

In order to remove spherical distortions of the dilated depth-map, we calculate two vectors for the fragments at (x, y) and the pixel location of $d_{c,min}$

$$\begin{aligned} \mathbf{a}^* &= {\mathbf{R_c}'}{\mathbf{K_c}}^{-1}[x, y, 1]' \\ \mathbf{b}^* &= {\mathbf{R_c}'}{\mathbf{K_c}}^{-1}[x + x', y + y', 1]'. \end{aligned} \tag{8}$$

These vectors are then normalized to homogeneous directional vectors, pointing from the origin of the world coordinate system to the $z = 1$ plane in the rotated but untranslated camera, with $\mathbf{a} = {}^{\mathbf{a}^*}/_{a_z^*}$ and $\mathbf{b} = {}^{\mathbf{b}^*}/_{b_z^*}$. The relationship of length between the two vectors provides the required factor by which the found minimum depth value $d_{c,min}$ has to be multiplied, resulting in the new depth

$$d_c(x, y) = \frac{|\mathbf{a}|}{|\mathbf{b}|} d_{c,min}\,(x', y')\,. \tag{9}$$

In a similar fashion the remaining unknown regions are considered as *background* by find the *maximum* depth value within a larger neighborhood with $k = 100$. Formulating this as a separable problem in horizontal and vertical direction we yield a linear computational complexity of $O(2n)$ as compared to $O(n^2)$. Under the assumption of missing regions in the depth-map being no larger than $2k$ pixels in horizontal or vertical direction we achieve convergence for a $[1280 \times 1024]$

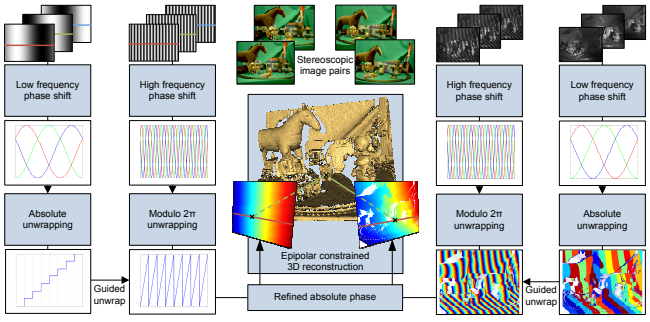

Fig. 2. Our phase unwrapping procedure for active 3D scene reconstruction, showing corresponding projected images (left) and captured images (right). The resulting epipolar constrained 3D reconstruction is shown in the center. Four color images provide additional texture for view synthesis.

depth-map in a single iteration, by searching along the horizontal direction first and then filling the depth-map vertically. Since the second vertical pass utilizes but overwrites filled depth values estimated in the previous pass, we consider the vertical direction as the dominant one in this case (Fig. 3b).

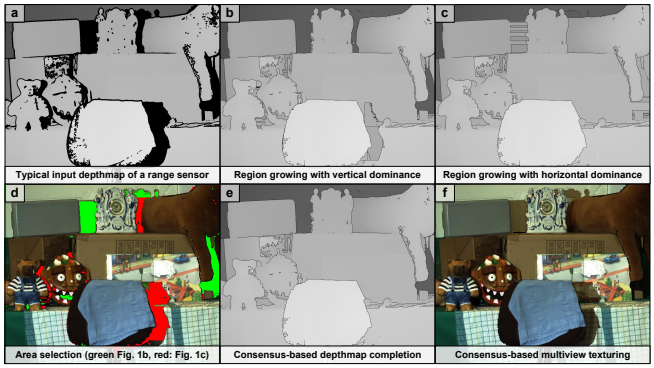

Fig. 3. Consensus-based area selection for depth-map completion (one camera, one projector) and subsequent multiview texturing with all available color information.

A second pass of the occlusion filling stage, with orthogonal directional dominance compared to the previous run, allows for an alternative version of the completed depth-map (Fig. 3c). Both depth-maps then have multiple different erroneously and correctly estimated regions. Applying the techniques of consensus-based texturing, we can evaluated for which part of which depth-map the participating texture images align better (Fig. 3d). The binary absolute difference Δ between the two depth-maps with $\Delta = \lfloor |\Psi_x - \Psi_y| \rfloor$ is efficiently segmented into connected components S_i [23] after which the individual fractions of each depth-map can be allocated to the final result in Fig. 3e, so that each fragment $\mathbf{f}$ is given by:

$$\Psi = \begin{cases} \Psi_x & : \quad N_y > N_x \\ \Psi_y & : \quad \text{else} \end{cases} \tag{10}$$

with N_x and N_y given as $N_{x,y} = \sum_{\forall \mathbf{f} \in S_i \subset \Phi_{x,y}} (\mathbf{f}_o + \mathbf{f}_s)$, while $\mathbf{f}_o$ and $\mathbf{f}_s$ denote the presence of an outlier texture or a shadowed fragment in the consensus-based texturing stage, respectively. Likewise, the selection provided by Eq. 10 is applied to create the improved textured view in Fig. 3f without the computational effort of a subsequent rendering pass.

5. DATA CAPTURE OVERVIEW

The system that we use for data capture is comprised of two Viewsonic PJD6241 120 Hz DLP projectors and two Basler A504k high speed cameras. Four Basler Scout scA1300-32gc color cameras provide texture information. Two of these, along with one projector and one high speed camera form a scan unit, as depicted in Fig. 4.

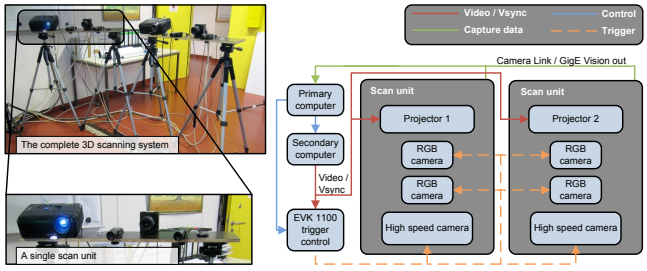

Fig. 4. The complete 3D scanning system (top-left), comprised of two identical scan units (bottom-left). The block diagram to the right illustrates the system's signaling flow path.

The periodic, time-multiplexed 12 frame structured light sequence is continuously analyzed by an external Atmel EVK1100 AVR32 microcontroller, in order to maintain trigger synchronization with the camera array. The currently used primary workstation for performing 3D reconstruction and view synthesis is based on a quad 3.2 GHz Core i7 and a Nvidia GTX680 graphics card. Configuring the cameras to capture structured light from both projectors, results in two additional wide base line camera-projector pairs. This yields a total of four pairs which capture simultaneously, each contributing to the 3D reconstruction as illustrated in Fig. 5.

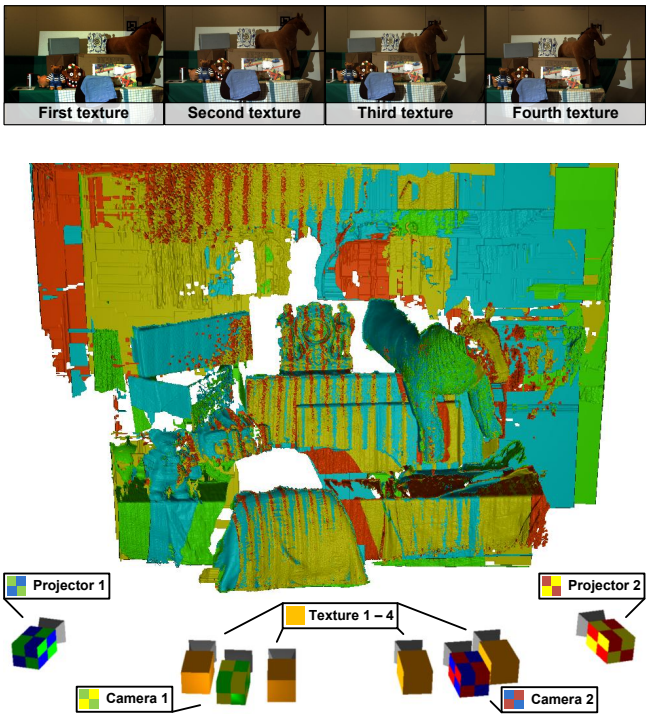

Fig. 5. Color-coded multiview reconstruction illustrating the resulting gain in terms of reconstruction completeness with up to four camera-projectors pairs. The four texture images of this setup are shown above.

6. TIMING OVERVIEW

Averaged per frame timing diagrams for complete reconstruction and view synthesis with the presented system are illustrated in Fig. 6. The resolution of the depth-maps is 1280×1024, the resolution of the four images applied for texturing in the consensus-stage is 1294×964. An OpenGL / GLSL based GPU implementation allows the method to be real-time capable. Currently, a complete reconstruction and rendering cycle accumulates to 79.5 ms. Image capturing from the multi-camera array is performed by the CPU in parallel. The time necessary for 3D reconstruction (7 ms) could be neglected with a dedicated depth-imaging device, such as a Time of Flight camera or a Microsoft Kinect.

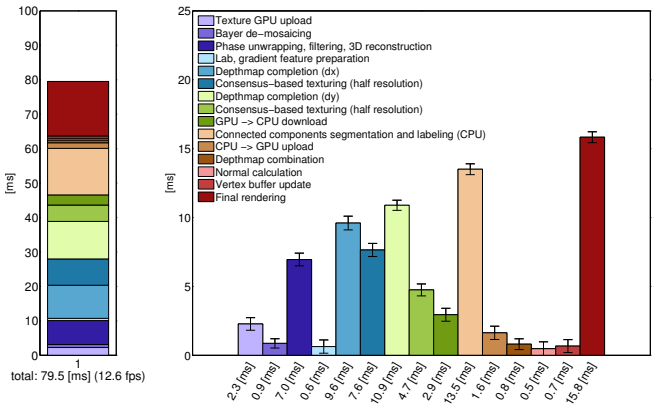

Fig. 6. Averaged per frame timing diagrams for complete 3D reconstruction, showing the accumulative sum over all processing steps (left) and their individual timing details (right).

7. CONCLUSION

We have presented work on the completion of depth-maps and subsequent consensus-based multiview texturing of the geometric 3D models built from the depth-maps. Our work is applicable to general depth imaging devices, such as Microsoft's Kinect sensor, not just to our own 3D scanning structured light array. As has been shown, the consensus-based approach for texturing 3D models with images captured from multiple viewpoints provides good visual results. Additionally, the approach can be applied for further refinement and for occlusion filling of the depth-maps in a way that lets the textures on the resulting 3D models align properly. Our method is real-time capable and thus allows for interactive systems and applications. We believe that our approach of creating intermediate textured 3D models of a scene on a frame by frame basis and then rendering from these models to virtual cameras, or an entire virtual light field, will become a powerful way to realize the extreme high number of input views necessary for next generation full parallax and holographic displays.

8. REFERENCES

[1] S. Knorr, K. Ide, M. Kunter, and T. Sikora, "The avoidance of visual discomfort and basic rules for producing good 3d pictures," *SMPTE Motion Imaging Journal*, pp. 72–79, Oct. 2012.

[2] W. Matusik and H. Pfister, "3d tv: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," in *ACM SIGGRAPH 2004 Papers*. ACM, 2004, pp. 814–824.

[3] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 1996, pp. 31–42.

[4] K. Ide and T. Sikora, "Adaptive parallax for 3d television," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*. IEEE, 2010, pp. 1–4.

[5] J.L. Posdamer and M.D. Altschuler, "Surface measurement by space-encoded projected beam systems," *Computer graphics and image processing*, vol. 18, no. 1, pp. 1–17, 1982.

[6] F. Blais, "Review of 20 years of range sensor development," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 231, 2004.

[7] E. Stoykova, AA Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, et al., "3-d time-varying scene capture technologies – a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1568–1586, 2007.

[8] J. Salvi, J. Pagès, and J Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, pp. 827–849, 2004.

[9] E. İmre, S. Knorr, B. Özkalaycı, U. Topay, A. Aydın Alatan, and T. Sikora, "Towards 3-d scene reconstruction from broadcast video," *Signal Processing: Image Communication*, vol. 22, no. 2, pp. 108–126, 2007.

[10] J. Kim and T. Sikora, "Confocal disparity estimation and recovery of pinhole image for real-aperture stereo camera systems," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*. IEEE, 2007, vol. 5, pp. V–229.

[11] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3017–3024.

[12] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo analysis by hybrid recursive matching for real-time immersive video conferencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 321–334, 2004.

[13] M.Z. Brown, D. Burschka, and G.D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 993–1008, 2003.

[14] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[15] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 519–528.

[16] S. Zhang and P Huang, "High-resolution, real-time 3d shape acquisition," *Computer Vision and Pattern Recognition Workshop, 2004*, pp. 28–28, 2004.

[17] T. Weise, B. Leibe, and L. Van Gool, "Fast 3d scanning with automatic motion compensation," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, 2007, pp. 1–8.

[18] P. Wissmann, R. Schmitt, and F. Forster, "Fast and accurate 3d scanning using coded phase shifting and high speed pattern projection," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, may 2011, pp. 108 –115.

[19] M. McGuire, "A fast, small-radius gpu median filter," in *ShaderX6*, February 2008.

[20] K. Ide, S. Siering, and T. Sikora, "Automating multicamera self-calibration," in *Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009.

[21] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.

[22] K. Ide and T. Sikora, "Real-time active multiview 3d reconstruction," in *Computer Vision in Remote Sensing, International Conference on, Xiamen, China*, 2012.

[23] K. Wu, E. Otoo, and K. Suzuki, "Optimizing two-pass connected-component labeling algorithms," *Pattern Analysis & Applications*, vol. 12, no. 2, pp. 117–135, 2009.