Our 3D Vision Data-Sets in the Making

H. Aanæs[1] K. Conradsen[1] A. Dal Corso[1] A. B. Dahl[1] A. Del Bue[2] M. Doest[1]
J. R. Frisvad[1] S. H. N. Jensen[1] J. B. Nielsen[1] J. D. Stets[1]
G. Vogiatzis[3]
[1] Technical University of Denmark [2] Istituto Italiano di Tecnologia [3] Aston University, UK

1. Introduction

Over the previous years, we have at the Section for Image Analysis and Computer Graphics at the Technical University of Denmark been working on generating high quality data sets for computer vision via our lab setup using a 6-axis industrial robot. This has provided a new data set aimed at feature matching [1, 4], and two data sets aimed at multiple view stereo [16, 18]. The resulting data sets are publicly available via `http://roboimagedata.compute.dtu.dk/`.

The evaluation of computer vision algorithms on these data sets has provided useful insights on realistic scenarios by setting a rigorous framework for evaluation. The results of these efforts have been well received by the community and the hardware and software platform associated with the robot is now well developed. We are currently in the process of making three new data sets aimed at 3D vision, with a special focus on the more challenging aspects, such as radiometry and the modelling of non-rigid objects. The construction of these data sets all leverage on our robotic setup's ability to produce ground truth camera and surface geometry, as briefly outlined in Section 2, and there is a great deal of commonality in the making of the data sets.

This abstract describes our current ongoing work on this data set construction for 3D vision. The data sets include:

1. A direct extension of our large multiple view stereo (MVS) data set [16], where we are now including transparent and semi transparent objects into the scenes, Section 3. A challenge in doing this is getting the ground truth geometry of the transparent objects.

2. A data set addressing the radiometric challenges in 3D vision as presented in Section 4 where we aim at extending our MVS data set by explicitly measure the bidirectional reflectance distribution function (BRDF) of the surfaces. This will have the additional feature to finally give a data set for evaluating photometric stereo with a ground truth.

3. An extension of our data set on feature matching to evaluate these algorithm with non-rigid objects, (Section 5) where we use actuators to make stop motion 3D data sets. This data set will also evaluate Non-rigid Structure from Motion (NRSfM) with realistic objects.

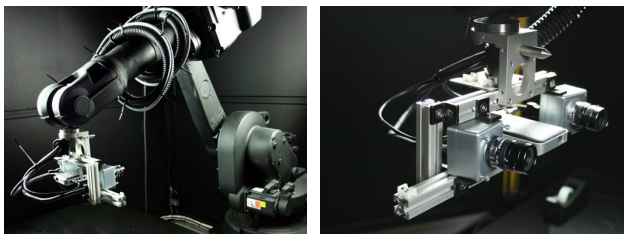

Figure 1. Photos of the 6-axis industrial robot mounted with two cameras and a projector. Cameras allow for MVS, and in conjunction with the projector SL provides ground truth point clouds.

2. Brief System Overview

Our experimental setup, cf. [1], is built around a 6-axis ABB IRB 1600 industrial robot, providing a flexible, precise, and highly repeatable camera pose. The robot is mounted with two Point Grey Grasshopper3 3376×2704 8-bit RGB cameras and a projector (for previously published datasets the cameras were 1600×1200 8-bit Point Grey Scorpion cameras). From each position ground truth surface point clouds are obtained using structured light (SL), and stereo images with a 32 cm baseline are captured with the camera pair. Five individually controlled 6500K LED tube lights allow for soft natural illumination of scenes from varying directions. Figure 1 shows the robot.

Previous evaluations of our system [16] have shown that the ground truth samples obtained through SL have good accuracy with a surface standard deviation of 0.14 mm. We expect similar or better performance in this data set. Positioning repeatability of the robot is very high, with a standard deviation of 0.0031 mm over two months.

Additional instruments used for generating the data include a CT (Computed Tomography) scanner for ground truth geometry of transparent objects (described in Section 3) and an illumination arch for controlled directional light-

Figure 2. Preliminary images from our data set. In the first row, three glass objects (sphere, bowl, teapot) with markers placed on. On the second row, three calibration and rendering tools part of the pipeline: a black and white checkerboard (coordinate estimation), an X-Rite ColorChecker® (color balance compensation) and a chrome sphere (environment light evaluation).

ing (described in Section 4).

3. Transparent Objects

Our goal is to extend our original MVS dataset, [16], to account for transparent objects where the focus is on reconstruction of geometry and appearance. Usually, the radiometric behavior of the objects used in 3D reconstructions is assumed diffuse and opaque. This leads to a number of simplifications that we cannot apply to transparent objects. In the case of transparent objects, refraction and reflection cause distortion effects that complicate reconstruction.

Previous methods acquire data sets useful for image-based rendering of a transparent object [20, 12]. However, these methods do not produce an actual triangle mesh and require special rendering techniques for reconstruction of the appearance of the transparent object. A survey on methods that do provide a triangle mesh is available [14]. In this survey, they note that CT scanning of refractive objects like glass is costly but straight forward. Thus, we use CT scanning to obtain ground truth geometry. Another way is to acquire shape and pose of a transparent object from motion [3]. In any case, there seems to be no data set, like the one we propose, which is useful for multiple view reconstruction of transparent objects.

3.1. Data

Our data set contains a set of multiple view HDR images of three glass objects with different radiometric properties (top row of Figure 2). We currently[1] use a solid sphere, a bowl with lid (composed of two parts) and a teapot with multiple thin glass layers (composed of three parts). The walls of the bowl and the teapot have different thickness. A diffuse backdrop is provided for the objects. We have made this as a gradient checkerboard, so that one half of the squares varies in color from left to right, and the other half varies in color from top to bottom. In this way, we can see how light reflects, refracts and scatters through the objects. The refractive index of the glass objects will be estimated directly from the scanned images, or, if this is unsuccessful, by the use of a refractometer. We marked the objects with small black plastic spheres, in order to easily determine their position relative to the scene. In our data set, we also provide high-resolution triangle meshes generated from CT scans. We use these scans as ground truth for either geometrical reconstruction algorithms or physically based rendering algorithms for appearance modelling.

Our current data set creation procedure is as follows. First, we choose a sequence of camera positions and orientations for our industrial robot. The robot enables us to reproduce a given set of positions and orientations with a very high precision. Then, we capture a first set of images placing a black and white checkerboard in the scene. This is done to obtain the camera positions relative to the scanned objects and calculate camera parameters for the setup. Secondly, we scan a commercial color checker, which allows us to compensate for color channel alterations in the final images. Finally, we scan a chrome sphere to get an HDR environment map of the surroundings. We use the resulting map as a light source in our rendering algorithms [5], so we can simulate the resulting scene with high precision. After these three calibration steps, we can finally scan the glass objects using the same pre-defined path used for the calibration images.

Once compiled, we are planning to use this data set to verify that the radiometric models [9] properly describe the radiometric properties of the scene. To do this we plan to feed the ground truth of our data into a custom-built renderer based on the NVIDIA OptiX library [23], and see how well it reproduces the images. If successful, we have a validated computational model, which in principle we 'just' have to invert to do 3D reconstruction of transparent objects. Following this we plan at applying state of the art 3D reconstruction algorithms, c.f. e.g. [11, 15, 21],and quantify how far the state of the art has come toward solving this central 3D vision reconstruction problem.

4. BRDF measurements and Photometric Stereo

The radiometric behaviour of an object plays a crucial role in MVS. Often this behaviour has been ignored or at most assumed Lambertian. This allows for acceptable reconstructions of geometry, but often poor recovery of the reflectance. For more accurate MVS and reflectance capture, the BRDF of an object should be taken into account

[1] The object set will be significantly expanded.

and this is a problem that receives a growing amount of attention [17, 27]. Within the field of photometric stereo, the reflectance of an object is the key element in recovering surface normals and thereby indirectly the object's geometry. Also here, assumptions about reflectance are made, these include e.g. Lambertian behaviour [30] or isotropic BRDFs [13].

For both of the above areas, a multi-view data set having ground-truth reflectance behaviour would be of great value, and does, to our knowledge, not currently exist. We are therefore now working on a MVS data set where not only the ground-truth geometry is given, but also a densely sampled BRDF ground-truth for all materials in the scene. In the following, we will elaborate on the details of how this data set will be acquired and what it will include.

4.1. Concept

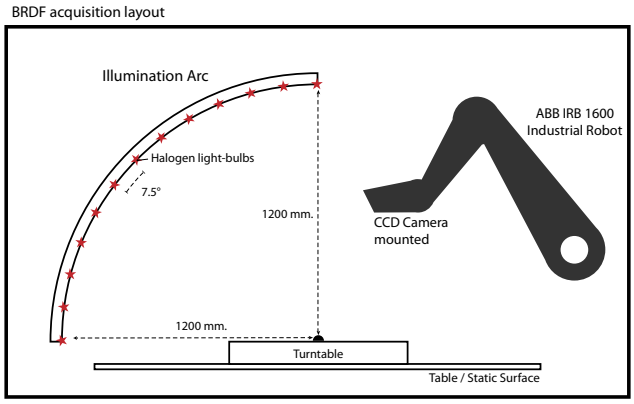

Figure 3. Schematic of BRDF capturing setup. Setup includes a 6-axis industrial robot holding a CCD (stereo) camera for view, and an arc in conjunction with a turntable for illumination.

Capturing the reflectance of a material generally requires four degrees of freedom: polar and azimuthal angle of illumination-direction, and polar and azimuthal angle of view-direction, $\rho(\omega_i, \phi_i, \omega_v, \phi_v)$. Utilizing our lab-facility's 6-axis industrial robot, mounted with a stereo-camera setup, all view directions (ω_v, ϕ_v) can effectively be captured. For illumination directions (ω_i, ϕ_i), we utilize an illumination arc and a rotation-table. The arc holds a range of halogen light-bulbs and is capable of covering the polar angle ϕ_i in 7.5° intervals. The rotation-table turns the target sample with a resolution of $< 1^\circ$, thus densely covering θ_i. Figure 3 shows a schematic of the BRDF capturing setup, and Figure 4 is a photo of an actual acquisition scene.

Using the above described setup, we intend to densely sample the BRDFs of a collection of objects whose surfaces consist of one or a few, isotropic, BRDFs. The BRDFs of each material will be stored in the 3-dimensional Rusinkiewicz frame for isotropic BRDFs [24], as also done

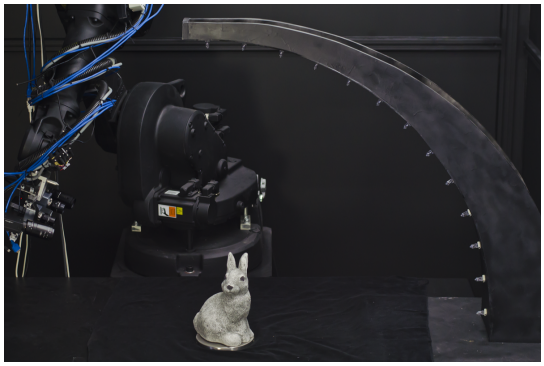

Figure 4. Capturing the BRDF of an object with known geometry. All illumination directions and view-directions are covered for each type material present on the object.

in the MERL database[19], although with a coarser resolution of 7.5° in each dimension. In conjunction with the densely sampled BRDFs, stereo images of scenes containing the sampled objects will be acquired for a wide range of directions. Objects will be of relatively low geometric complexity, and scenes will consist of one or more of the objects.

5. Non-Rigid Structure from Motion

Evaluating Non-rigid feature matching and NRSfM algorithms[2] in a quantitative manner has in the literature proven to be problematic. Deformations are inherently a dynamic process and subject to the physical properties of the objects in consideration. Thus, evaluating deformation modelling algorithms require a reasonable number of different objects and set of motions. Also, given the dynamic deformation objects might change their topology (e.g. stretching and tearing) and easily self-occluded some parts of the shape. For this reason, many approaches have provided several models that fit specific types of deformation, but that cannot comprise all of them. For this reason understanding the real performance of methods on realistic deformations is necessary to push forward advancements in this field.

The central problem of producing reference ground truth has been approached from many different angles. Several works compare their methods using synthetically generated images, as the true 3D geometry is readily available[29, 25, 22, 10]. Another popular approach is using MOCAP data, mainly human motion, for generating both test video sequence with 3D reference points [2, 7, 10, 28, 29]. Both falls short, as the former often lacks the complexity found in real-life scenes and the latter provides only a sparse set of reference points that are likely not to be possible to detect from images because of occlusions. As stated in [8, 25], there is a lack of and a need for a real-life NRSfM sequence

[2] A review on NRSfM methods, updated to 2010, can be found here: [26]

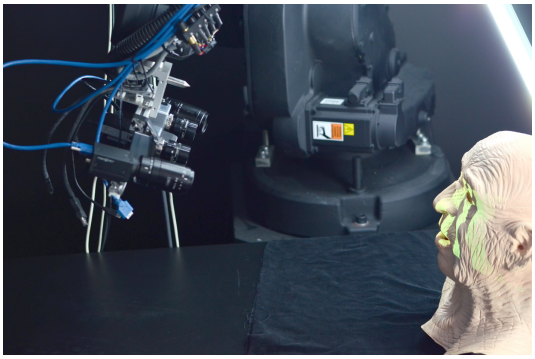

Figure 5. Robot arm carrying cameras for capturing stop motion frame and structured light data. A Gray code pattern is currently being projected onto the object.

with a dense 3D reference.
We seek to remedy this situation by providing a video recording of real objects with dense 3D ground truth for each frame. It will be accomplished using a stop motion like animation technique and structured light 3D scanning, combined in our unique recording setup.

5.1. Concept

We wish to simulate motion in a manner similar to stop motion animated films. Here a rigid object is moved into a certain pose, an image is taken, the object is slightly changed with a deformation, another image is taken etc. The result is a sequence that, when played at an interactive frame rate, provides the illusion of motion. We will apply the same principle here, in generating a benchmarking data set for NRSfM with ground truth.
Now one may ask, why not just record the motion using ordinary video format? After all, stop motion techniques does not properly reproduce motion blur artifacts that are present in standard recorded video sequences. Our approach has several significant advantages that greatly outweigh the loss of motion blur. Most importantly, we can obtain a 3D ground truth for each frame. After adjusting the object into its current frame position and acquiring an image for the stop motion sequence, we will perform a 3D scan using structured light. Utilizing gray code patterns we obtain a dense ground truth so obtaining both the image frame and a 3D reference for benchmarking and validation.
Another advantage is that we can obtain data from multiple views by acquiring images at different angles thus providing data for evaluating multi-view NRSfM (e.g. [6]). Furthermore, this procedure provides a great degree of control over both camera movement and object pose. As each frame is recorded independently, time in between becomes a non-issue.

Figure 6. Actuators for manipulating the geometry of the mask. The image of the mask has been superimposed on an image of the actuators, illustrating their functionality.

5.2. Implementation

Such data could be acquired by pure manual effort, however that would be extremely time consuming and error-prone. As such, a robotics solution is currently being developed with a the data acquisition procedure that is predictably and reproducibly implemented. In detail, a robotic arm move the camera and the projector needed for data acquisition and structured light scan. From this the view position can be determined with high precision and reproducibility. Figure 5 illustrates this setup.
Additionally, object deformation will also be automated and Figure 6 shows an example with an object where a mask resembling a human face is put on top of two actuators. Manipulating the actuators deforms the mask geometry, simulating facial movement. Similar results can be obtained with cloth, paper and other deformable materials.

6. Concluding Remarks

We have here presented our ongoing work on making high quality data sets for evaluating and developing methods for 3D vision. A motivation for doing this is that we see a need for this, especially with respect to making data sets that are large enough, so that it is possible to reasonably determine if differences in performance are a statistical fluke, or are in fact statistically significant.

By presenting our ongoing work in this forum, we hope to get valuable and constructive feedback on how these data sets in the making could be adapted to serve the needs of the computer vision communities as best possible.

References

[1] H. Aanæs, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, 97(1):18–35, 2012.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(7):1442–1456, 2011.

[3] M. Ben-Ezra and S. Nayar. What does motion reveal about transparency? In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 1025–1032, 2003.

[4] A. Dahl, H. Aanæs, and K. Pedersen. Finding the best feature detector-descriptor combination. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325, 2011.

[5] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Proceedings of ACM SIGGRAPH 98*, pages 189–198, 1998.

[6] A. Del Bue and L. Agapito. Stereo non-rigid factorization. *International Journal of Computer Vision*, 66(2):193–207, February 2006.

[7] J. Fayad, L. Agapito, and A. Del Bue. Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, volume 6314 of *Lecture Notes in Computer Science*, pages 297–310. Springer, 2010.

[8] K. Fragkiadaki, M. Salas, P. Arbelaez, and J. Malik. Grouping-based low-rank trajectory completion and 3D reconstruction. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 55–63. Curran Associates, Inc., 2014.

[9] A. S. Glassner. Surface physics for ray tracing. In A. S. Glassner, editor, *An Introduction to Ray Tracing*, chapter 4, pages 121–160. Academic Press Ltd., London, UK, 1989.

[10] P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 802–809. IEEE, 2011.

[11] S. W. Hasinoff and K. N. Kutulakos. Photo-consistent reconstruction of semitransparent scenes by density-sheet decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):870–885, 2007.

[12] T. Hawkins, P. Einarsson, and P. E. Debevec. A dual light stage. *Rendering Techniques 2005 (Proceedings of EGSR 2005)*, pages 91–98, 2005.

[13] M. Holroyd, J. Lawrence, G. Humphreys, and T. Zickler. A photometric approach for estimating normals and tangents. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2008)*, 27(5):133, 2008.

[14] I. Ihrke, K. N. Kutulakos, H. Lensch, M. Magnor, and W. Heidrich. Transparent and specular object reconstruction. *Computer Graphics Forum*, 29(8):2400–2426, 2010.

[15] I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich. State of the art in transparent and specular object reconstruction, 2008.

[16] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 406–413, 2014.

[17] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo beyond Lambert. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I:171–178. IEEE, 2003.

[18] S. Kim, H. Aanæs, A. Dahl, K. Conradsen, R. Jensen, and S. Kim. Multiple view stereo by reflectance modeling. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, 2012.

[19] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2003)*, 22(3):759–769, 2003.

[20] W. Matusik, H. Pfister, R. Ziegler, A. Ngan, and L. McMillan. Acquisition and rendering of transparent and refractive objects. pages 267–278, 2002.

[21] N. Morris and K. Kutulakos. Reconstructing the surface of inhomogeneous transparent scenes by scatter-trace photography. In *ICCV 2007*, pages 1–8, Oct 2007.

[22] S. I. Olsen and A. Bartoli. Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision*, 31(2-3):233–244, 2008.

[23] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich. OptiX: a general purpose ray tracing engine. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2010)*, 29(4):66:1–66:13, July 2010.

[24] S. Rusinkiewicz. A new change of variables for efficient BRDF representation. In *Rendering Techniques (Proceedings of EGWR 1998)*, June 1998.

[25] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *Proceedings of 3DIMPVT 2012*, pages 509–516. IEEE, 2012.

[26] M. Salzmann and P. Fua. Deformable surface 3d reconstruction from monocular images. *Synthesis Lectures on Computer Vision*, 2(1):1–113, 2010.

[27] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 519–528. IEEE, 2006.

[28] L. Tao and B. J. Matuszewski. Non-rigid structure from motion with diffusion maps prior. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 1530–1537. IEEE, 2013.

[29] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 2761–2768, 2010.

[30] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):191139, 1980.