Large Scale Multi-view Stereopsis Evaluation

Rasmus Jensen[†], Anders Dahl[†], George Vogiatzis[‡], Engin Tola[§], Henrik Aanæs[†]
[†]Technical University of Denmark, [‡]Aston University, [§]Aurvis R&D
raje@dtu.dk, abda@dtu.dk, g.vogiatzis@aston.ac.uk, tola@aurvis.com, aanes@dtu.dk

Abstract

The seminal multiple view stereo benchmark evaluations from Middlebury and by Strecha et al. have played a major role in propelling the development of multi-view stereopsis methodology. Although seminal, these benchmark datasets are limited in scope with few reference scenes. Here, we try to take these works a step further by proposing a new multi-view stereo dataset, which is an order of magnitude larger in number of scenes and with a significant increase in diversity. Specifically, we propose a dataset containing 80 scenes of large variability. Each scene consists of 49 or 64 accurate camera positions and reference structured light scans, all acquired by a 6-axis industrial robot. To apply this dataset we propose an extension of the evaluation protocol from the Middlebury evaluation, reflecting the more complex geometry of some of our scenes. The proposed dataset is used to evaluate the state of the art multi-view stereo algorithms of Tola et al., Campbell et al. and Furukawa et al. Hereby we demonstrate the usability of the dataset as well as gain insight into the workings and challenges of multi-view stereopsis. Through these experiments we empirically validate some of the central hypotheses of multi-view stereopsis, as well as determining and reaffirming some of the central challenges.

1. Introduction

Stereopsis including both two and multiple views (MVS) is one of the central problems in computer vision, allowing us easy capture of our environment such that appealing 3D models can be made. This has many applications in entertainment, augmented reality, robotics, as well as industrial inspection and aireal cartography. During the last decade the advances in MVS have been driven by the seminal benchmark MVS datasets. Popular benchmark data include the Middlebury Multi-View Stereo data [17] and the building dataset by Strecha et al. [18]. Although these datasets have been tremendously useful they also have their limitations due to their relatively small sizes – Middlebury contains two scenes and Stretcha et al. contains twelve. To continue the important advancement of MVS, the basis for empirical development comparison and evaluation has to advance along with the methodology.

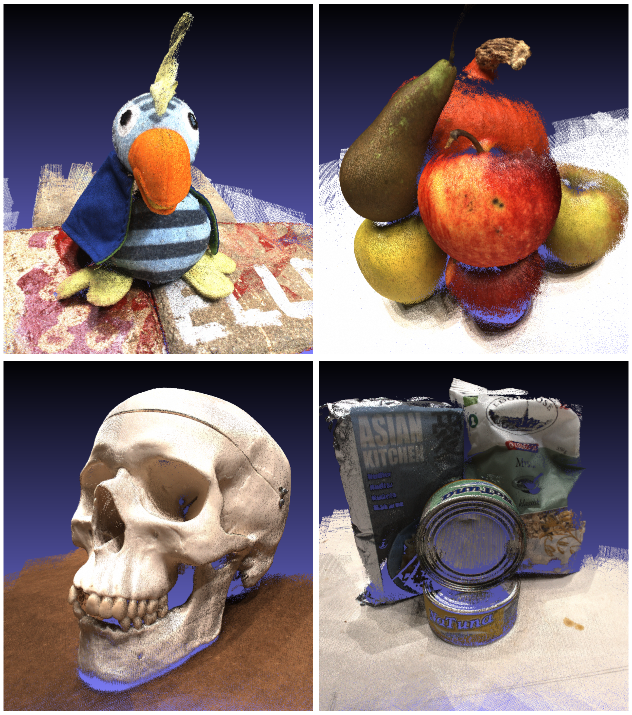


Figure 1. Subset of point clouds in our reference dataset. The images shows point reconstruction in scenes with a variability in geometry, reflectance, and texture. These images are grouped in our analysis into categories like groceries and vegetables. The point reconstructions are shown before pruning.

To address this issue we propose a new dataset aimed at MVS, consisting of 80 different scenes. This makes this dataset almost an order of magnitude larger than the current state of the art. Examples of point clouds from the proposed dataset are shown in Fig. 1. Apart from an increase in the size of the dataset, the variability in surface reflectance properties and geometric complexity have also been increased. The previous sets mostly focused on Lambertian surfaces of relatively simple geometry – e.g. Strecha et al. [18] consists of twelve outdoor scenes of historical buildings, with few specular surfaces such as steel and glass.

The dataset proposed here is compiled using a 6-axis industrial robot, with the evaluation reference achieved via a structured light scanner. We have chosen the term *reference data* instead of ground truth, to emphasize that these are also physical measurements. As described in Section 3, this setup enables us to make data of high quality and quantity. After proposing a new evaluation protocol in Section 4, which is an update from the one in [17] to address the high increase in geometric complexity, an evaluation of MVS methods on this dataset is presented in Section 5. This evaluation validates the usability of the dataset as well as providing insight into issues affecting MVS performance.

The evaluation of Section 5, specifically applies the methods of Tola et al. [19], Furukawa and Ponce [3] and Campbell et al. [2] to the proposed dataset computing 3D point reconstructions as well as dense triangular surface reconstructions based on these. Here we investigate the properties of the meshing. Among others, we find that there is a tradeoff between how accurate the method is and its completeness, so the most complete method is the least accurate and vice versa. Meshing is generally poor for complex geometries, despite quite accurately reconstructed surface points. So, there is a large research potential in improving meshing methods. The dataset consists of images, camera calibration (internal and external), reference scans and observability masks, as well as code for evaluation[1].

2. Related work

The first work that attempted to benchmark MVS algorithms was [17] where the performance of six algorithms was measured across six different scenes. The authors subsequently invited submissions of reconstruction results from dozens of different algorithms that were publicly ranked against each other. The somewhat artificial, low-resolution setup of [17] was subsequently improved in the evaluation effort of [18] that consisted of high-resolution images of outdoor scenes. Both [17] and [18] made an invaluable contribution to the advancement of MVS technologies by providing a solid platform on which improvement to existing state of the art can be measured and recorded.

Our work contributes to the evaluation of MVS, albeit with a different focus. In [17, 18] the evaluators' basic question was "which MVS algorithm works best for *this scene*?" In our work we ask the question "what scene types works best for this MVS algorithm and what scene features make MVS reconstruction fail?" Posing the question this way facilitates more detailed understanding of current state of the art MVS and several future research challenges for MVS.

The evaluations of [17, 18] consider a small number of 3D scenes that are thought to be representative of real-world application domains for MVS. In practice, they choose well-textured diffuse-reflectance 3D objects on which MVS algorithms tend to perform quite well. They then apply several algorithms in order to create a performance ranking for each scene. Our approach is to consider the widest possible range of 3D scenes one might encounter in real applications, and then consider how particular types of MVS algorithms perform on each type of scene. This approach sheds light on the performance of MVS technology as a whole and its overall suitability for particular applications.

Most successful MVS algorithms can be divided into two main categories: point-cloud based (e.g. [3, 5, 7, 19, 20]) and volume-based methods (e.g. [6, 11, 12]). Volume-based methods aggregate photo-consistency data in a 3D volume and compute a 3D surface within that volume using surface optimisation. On the other hand, point-cloud based methods convert photo-consistency data into a 3D point-cloud, which is then converted into a 3D surface using standard meshing techniques such as Poisson reconstruction [9], Graph-cuts [20] or signed distance functions [13]. In this work we focus on point-cloud based methods because we can easily isolate the point-cloud stage from the surface extraction stage and all the filtering and regularisation this entails.

Within point-cloud based methods we can distinguish two different paradigms. Feature expansion [3] and depth-map fusion [2, 5, 7, 19, 20]. Under the feature expansion paradigm the algorithm starts from a set of 3D features in the scene, which then expand into nearby 3D points while outliers are filtered using occlusion reasoning. Depth-map fusion works by computing independent depth-maps for each image using neighbouring images. These depth-maps are then merged into a single point-cloud. We chose Furukawa and Ponce [3], Campbell et al. [2], and Tola et al. [19] as representative algorithms from the feature expansion and depth-map fusion families. It must be stressed again that our aim is not to directly compare the three methods or the three families of algorithms. Rather, by running these methods on a large selection of datasets we highlight the effect on performance of different types of 3D scenes.

Perhaps closer in spirit to the present work are some previous attempts at investigating in detail different aspects of MVS performance. In [10] there is a theoretical analysis of the impact of scene geometry on feature-expansion MVS methods. A serious evaluation of MVS algorithms based on depth map fusion is presented in [8]. Our work can be seen as an empirical analysis of both families of MVS algorithms.

A recent trend in MVS research has been to automate all aspects of the MVS pipeline, including viewpoint selection and image capture. For example, in [1, 4] MVS is applied to photographs of famous landmarks, harvested from online photo-collections. Similarly, the authors of [21] pro-

[1] Available from http://roboimagedata.imm.dtu.dk/

pose using MVS with sequences of images obtained by a remote controlled model helicopter for the purposes of automatic 3D mapping. These examples highlight a detailed understanding of the performance of MVS algorithms under different conditions, which is the purpose of the proposed dataset.

3. Data

Evaluation of calibrated MVS reconstruction requires data with knowledge about the camera position as well as the spatial position of the depicted surface. To get this we have chosen to construct a controlled environment for image acquisition similar to Seitz et al. [17]. In our setup, the camera is mounted on a 6 axis industrial robot as illustrated in Fig. 2. This provides flexible and precise camera pose. At each position we obtain a surface point cloud using structured light. Further we control the illumination by using a set of 18 light emitting diodes (LEDs) mounted above the scene.

The robot provides very precise camera positioning due to its very high position repeatability[2]. We obtain actual positions by a set of predefined path locations, and this path is calibrated using a fixed checkerboard pattern. The encoded path has subsequently been used for acquiring images of the 80 scenes in our dataset. The introduction of the robot arm allows for a free design of our experiment. In some of the scenes we let the robot move to camera positions on concentric spheres, something that would not be possible with a static setup.

The 80 scenes contain different number of camera positions. 59 scenes contain 49 camera positions and 21 scenes contain 64 camera positions. Example data is shown in Fig. 1. The camera positions of the smaller sets are placed on a sphere with the radius of 50 cm, i.e. around 35 cm from the scene surfaces. The larger sets contain an additional 15 positions on a concentric sphere at a distance of 65 cm from the scene centres as shown in Fig. 3. The content of the scene is chosen with varying reflectance, texture, and geometric properties, and include fabric, print, groceries, fruit, a bunny sculpture, and more, see Fig. 1.

The dataset has variation in light obtained by strobing the LEDs in groups to generate directional lighting variations. It should, however, be noted that for the experiments in this paper only the uniform illumination images are used, generated by using all 18 LEDs at once[3]. The image resolution is 1200 × 1600 pixels in 8 bit RGB colour, with practically all the scene being in the depth of field (long exposure, small aperture).

The reference points, obtained from structured light scans, are based on binary stripe encoding, which is recommended as being one of the most precise structured light methods [14, 15, 16].

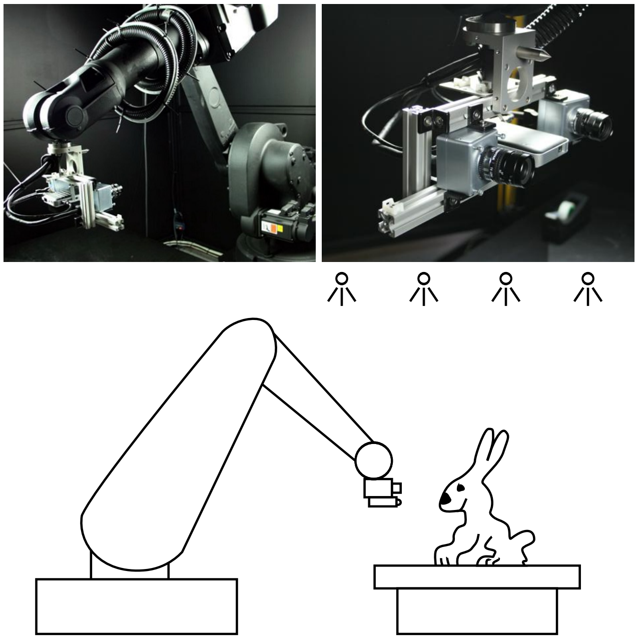


Figure 2. Top shows photos of the industrial robot mounted with the two cameras and the projector. Bottom is a schematic illustration of the setup, consisting of the industrial robot, LEDs in the ceiling, and the scene placed on a table.

Our experiments are dependent on the accuracy of structured light scans, and we have therefore measured the scan precision using an object with known geometry. We chose a bowling ball, because it is a spherical object of suitable size with simple and known geometry. A reference scan was obtained from each camera position, and all these scans were combined to make up the total reference data for each scene. For each scan we estimated the centre position and the radius of the sphere form the surface points using linear least squares. This also enabled us to estimate the deviation of the individual points from the sphere's surface. We obtained a standard deviation of 0.17 mm on the centre position estimates, and an average standard deviation on the surface points of 0.14 mm corresponding roughly to 0.6 pixels. Positioning repeatability of the robot turned out to be very high. Over the two months of data acquisition period we performed 10 calibrations, and the average standard deviation of the camera positions were 0.0031 mm. The reprojection error here was 0.067 pixels.

The reference scans are not complete. The main cause is that we only cover the front of the objects, but still there are areas seen by the cameras that have not been covered. This e.g. occurs because of object self occlusion and small holes where the structured light images have been severely underexposed. Despite these minor incompleteness the scans are

[2] The robot is coded with a predefined path, and all positions on that path are calibrated photogrammetrically.

[3] In few of the extreme positions the robot shaded a few of the LED's.

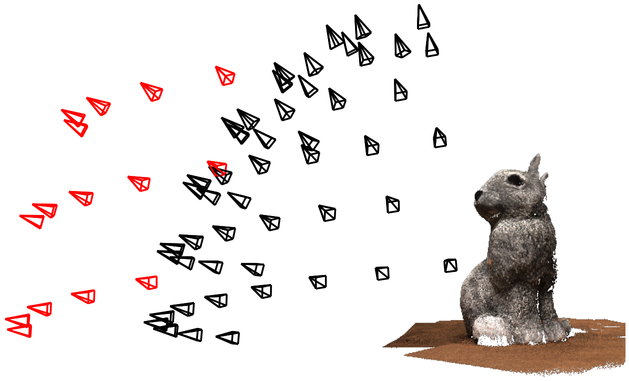


Figure 3. Camera positions at 50 cm distance (black) and 65 cm distance (red).

very dense, each containing 13.4 million points on average.

Only the scene objects are used in the evaluation. This is done by removing the part of the reconstruction containing the supporting table, simply by discarding points below a manually placed plane.

4. Evaluation method

The evaluation method must be chosen according to the nature of the reference dataset. With our choice of scenes, where textural and geometric complexity varies, we must consider how to handle issues like missing data, how to quantify the distance between two irregularly sampled surface reconstructions, and how to handle outliers. With these considerations in mind, we have strived at an evaluation method that is unbiased, but still sensitive to performance differences of the considered MVS algorithms.

4.1. Missing Data

To obtain a fair comparison we must handle missing data. In the Middlebury Multi-View Stereo benchmark [17], the issue of missing data was addressed by fitting a closed surface to the reconstructed structured light points. Surface points were then added in areas with no reference data, by placing points on the reconstructed surface with the same density as the rest of the scanned surface. In the evaluation, points closest to the inserted data were then removed. In effect, this creates an implicit 3D observability mask that allows an evaluation only of the points within the mask.

In our evaluation method we have chosen a similar approach. We explicitly compute an observability mask, and only evaluate stereo reconstructed points located within it. This is because of the high complexity of some scenes as well as the partial coverage, where the scenes are only seen from one side, it is hard to obtain a good surface reconstruction of the missing parts. The observability mask is obtained as the union of the individual visibility mask estimates of the 49 or 64 structured light scans. We employ a voxel grid with a voxel size of 1 mm.

The reasoning behind the observability mask is that no surface should be present between the reconstructed 3D points and the cameras. A mask is hereby computed by making a voxel grid around the object in question, and casting rays from the camera to the reconstructed points. These rays are extended an extra 10 mm and all voxels along that ray are marked as observed. The 10 mm depth assumption is needed to include stereo points reconstructed immediately behind the structured light reference points. The threshold of 10 mm was chosen as a tradeoff between including wrongly reconstructed MVS points in areas with reference data, and excluding correct MVS points in areas with no reference points.

4.2. Distances Between Reconstructions

As in the Middlebury evaluation [17], we also evaluate based on the *accuracy* and the *completeness*. Accuracy is measured as the distance from the MVS reconstruction to the structured light reference, and the completeness is measured from the reference to the MVS reconstruction. Both MVS reconstructions and structured light references are represented as point clouds, and for each point in one, we calculate the closest distance to the other. For the multiple view stereo points we, however, only use the points within the observability mask discussed above.

MVS methods, especially depthmap-fusion based ones, typically generate more 3D points around strongly textured surface regions. This can potentially cause a bias in the evaluation where we ideally would like the error measure uniformly sampled on the reconstructed surface. To avoid this problem in our evaluation, we decimate the MVS point clouds such that no two points are closer than 0.2 mm by visiting the points randomly and removing nearby points residing in a 0.2 mm neighborhood. This 0.2 mm sampling threshold is chosen to match the estimated resolution of our reference reconstruction. This decimation process ensures an unbiased evaluation across the whole reconstruction by keeping points in lower density areas, thus including outliers intact and reducing the effect of dense regions on the overall reconstruction accuracy.

Our evaluation include both MVS reconstructed 3D points as well as meshed surfaces from these points. To handle the surfaces meshes in our evaluation, we convert them back into point clouds. This is done by first supersampling the faces with a lower point density than 0.2 mm and then subsequently reducing the high point densities as described above. This gives an equal comparison of the effect of meshing, and the regularization of the result it implies.

For a given reconstruction, i.e. scene and method, the distances of each 3D point are condensed into comparable statistics by computing the mean and median for the the *accuracy* and *completeness*. This is, however, first done by

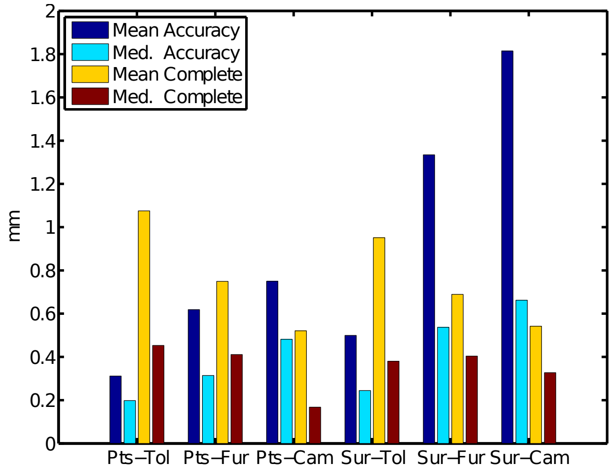


Figure 4. Performance over all 80 scenes of accuracy and completeness of reconstructed points (Pts) and meshed surfaces (Sur). The error is measured both as mean and median. Tol is Tola et al. [19], Fur is Furukawa and Ponce [3], and Cam is Campbell et al. [2].

removing all distances over 20 mm. We remove points to avoid biasing by outliers. In addition, spurious, closed surfaces are obtained by meshing the stereo reconstructions to remove such outliers.

5. Evaluation

The MVS methods of Campbell et al. [2], Furukawa and Ponce [3], and Tola et al. [19] have been evaluated by comparing both the raw point clouds and the mesh surfaces computed from these. A summary of the overall performance is shown in Fig. 4. The error measure from the MVS reconstructions to the reference data shows the accuracy of the method, whereas the opposite measures the completeness. This figure shows clearly that there is a tradeoff between completeness and accuracy with [19] being the most accurate and [2] being the most complete. This tradeoff manifests itself in the obtained detail at the expense of more errors, most notably outliers. In this sense there is no clear winner of the three methods. It should be noted that we did not optimize the method parameters for a tradeoff between the accuracy and completeness to make the methods more comparable because this would take them away from their original formulation. Furthermore, the method of Tola et al. [19] is *designed* for much higher resolution images than the ones used here, which in turn translates into a high accuracy and low completeness on these images.

With the vast dataset presented here we have observed some general trends for the investigated MVS methods. Some findings are in accordance with our expectations and others are surprising. Firstly, we found that largest source of poor performance is by far the lack of texture, as seen in Fig. 5. In many cases the meshing closes holes which compensates for this lack of texture. The success of this method, however, depends on the noise and the complexity of the surface. The box sequence shown in Fig. 5 for example is improved by meshing where the surface meshing fills holes that closely follow the reference surface points. For more complicated geometries the meshing does, however, not improve performance, but will often corrupt finer details.

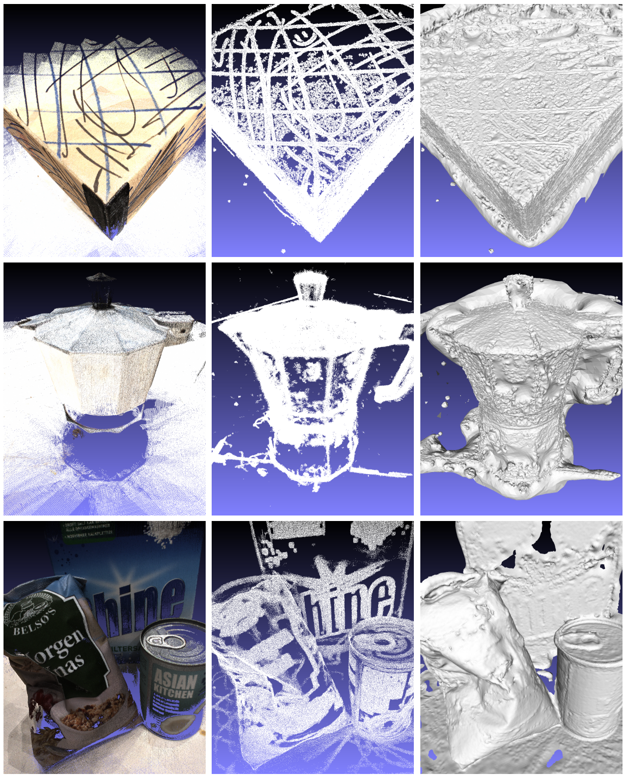
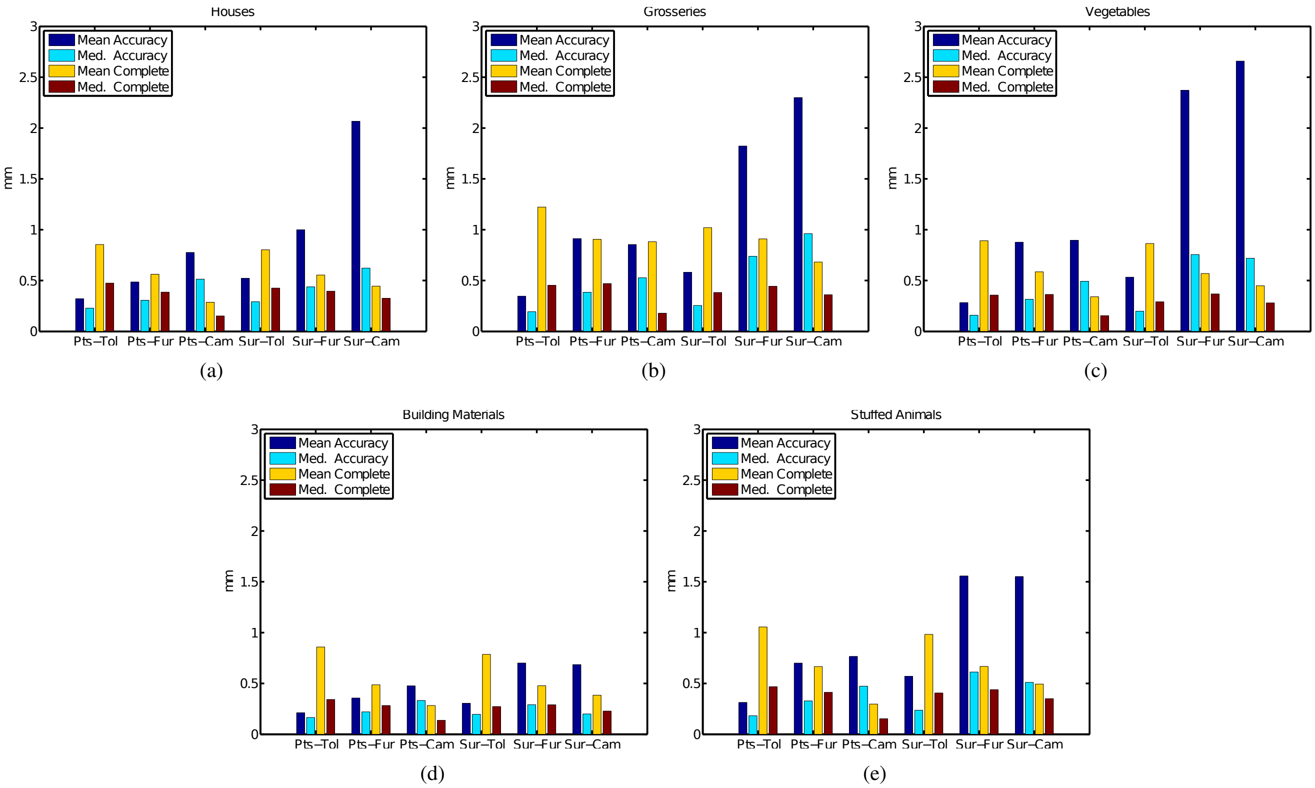


Figure 5. The top row shows an example of an object with missing texture resulting in reconstructions with holes. The simple geometry of the box did however recover the holes well. From left to right: the reference data points, the [3] point reconstruction of and the surface reconstruction of these points [9]. The middle row shows a highly specular espresso coffee pot, which is just about as difficult as it gets. From left to right: the reference data points, the point reconstruction [2] and the surface reconstruction of these points [9]. The bottom row shows a scene with both specularities and lack of texture. From left to right: the reference data points, the point and surface reconstructions of [19].

More surprisingly, we found that many other factors, which we expected to seriously corrupt the results, were not as problematic. As an example, the geometric complexity of the scenes did not influence the results to the extent we had expected. This was especially true for the point reconstructions. Specular surfaces did, in a similar manner, not influence the reconstructions as negatively as expected, which is shown in Fig. 5. Testing this particular scene of the espresso can with two view stereo, we did find a large



(a) (b) (c)

(d) (e)

Figure 6. Performance for different scene types. (a) is model houses, (b) is groceries, (c) is vegetables, (d) is building material, and (e) is stuffed animals.

degradation of the result. Our hypothesis is that with MVS there will almost always be an image pair, viewing a particular surface patch, which is uncorrupted by reflectance – note that reflectance has a high visual effect, but only in limited directions.

To enable a more refined analysis we grouped a part of our dataset into scene type categories, aiming at isolating more subtle effects. This also allows us to do an analysis by factoring the result out into these categories, as shown in Fig. 6. Here it is seen that the types of scenes typically used such as (model) houses and diffuse square building materials do well, whereas less traditional objects such as texture poor and specular objects found in a grocery store are more challenging.

5.1. Points vs. Surfaces

The state of the art in MVS has to some degree converged upon an approach where 3D points are reconstructed and then formed into a dense surface, typically via Poisson reconstruction [9] – referred to as meshing. This is also the case for the three state of the art methods presented here. We evaluate both the 3D point reconstructions and the meshed aggregates in order to investigate the properties of the meshing, but also because there is a debate as to which result is the correct to report.

As seen in Fig. 4, the point reconstructions in general perform best, which expresses a very clear trend looking at the individual reconstructions. As a general point, the cases where the meshed results are best, are as the box in Fig. 5, where there are large texture poor regions where no points are estimated *and* the geometry is simple enough for the implicit smoothing prior of the meshing to smooth noise and fill holes. Typically this applies to flat or spherical surfaces.

Examples of surface meshing are shown in Fig. 7, which illustrates how fine surface details are preserved by the method of [2] where many surface points are reconstructed, whereas many of these details are smoothed away in [19]. Complex geometry as seen in the middle front part of the house images are, however, severely corrupted by the surface meshing. This is one of the scenes where the meshing performed worst relative to the 3D point reconstructions. Firstly, it is seen that the meshing has problems with finer details. Such fine details are *inconsistent* with the implicit smoothing prior of the meshing algorithm. Secondly it is seen that more fine detail are captured in [2], but also more gross errors. This relates back to the accuracy/completeness tradeoff discussed above, in that more complete 3D point data gives more data to constrain the meshing. On the other hand the meshing process is relatively sensitive to outliers, which are increased by poorer accuracy. Sometime these outliers also seem to result in large surface portions being hallucinated.

Overall, our investigation shows that the three state of the art surface reconstruction algorithms investigated here have high precision in reconstructing surface points. Depending on the number of generated points a more or less detailed set of surface points can be obtained. Even small features like a small antenna of a thickness of around 1 mm on a model house was covered by precisely reconstructed surface points. Extending these surface points to a triangular mesh, however, is not easily done and many of these fine details are often lost. This is not surprising, because it can be hard to distinguish points on small surface details from groups of falsely detected points. Meshing the surfaces is, however, an important task for bringing MVS to the use for applications in e.g. entertainment, robotics, industrial inspection or aireal cartography. We see this as a great challenge and hope that the provided dataset can aid in this development as well as many other investigations within MVS or other computer vision problems.

6. Conclusion

In this paper we have presented a dataset and an evaluation procedure for MVS and performed an evaluation on the three state of the art methods by Campbell et al. [2], Furukawa and Ponce [3], and Tola et al. [19]. These methods have been evaluated in relation to estimated surface points and meshed surfaces. Our evaluation is based on a large collection of calibrated images and accurate 3D reference points together with an evaluation procedure. We evaluate both in relation to completeness and accuracy and we see that there is a tradeoff between accuracy and completeness in the three methods, such that the method of [19] has highest accuracy whereas Campbell et al. [2] obtained the highest completeness. This tradeoff can be caused by how discriminative towards reconstructed points the methods are. A high discrimination gives good accuracy but less completeness, whereas the opposite is seen with less discrimination. Surface meshing has a smoothing effect which is beneficial for simple geometries, because it tends to fill out holes. In general the effect of meshing is however not improving performance, because small details will be corrupted. This demonstrates the need for improvements of MVS meshing.

References

[1] C. Bailer, M. Finckh, and H. P. A. Lensch, "Scale robust multi view stereo," in *ECCV*. Springer-Verlag, 2012, pp. 398–411. 2

[2] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *ECCV*, 2008, pp. 766–779. 2, 5, 6, 7, 8

[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *TPAMI*, vol. 32, no. 8, pp. 1362–1376, 2010. 2, 5, 7, 8

[4] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *CVPR*, 2010, pp. 1434–1441. 2

[5] M. Goesele, B. Curless, and S. M. Seitz, "Multi-view stereo revisited," in *CVPR*, 2006, pp. 2402–2409. 2

[6] C. Hernández, G. Vogiatzis, and R. Cipolla, "Probabilistic visibility for multi-view stereo," in *CVPR*, 2007, pp. 1–8. 2

[7] V. H. Hiep, R. Keriven, P. Labatut, and J.-P. Pons, "Towards high-resolution large-scale multi-view stereo," in *TPAMI*, 2009, pp. 1430–1437. 2

[8] X. Hu and P. Mordohai, "Evaluation of stereo confidence indoors and outdoors," in *CVPR*, 2010, pp. 1466–1473. 2

[9] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006, pp. 61–70. 2, 5, 6

[10] R. Klowsky, A. Kuijper, and M. Goesele, "Modulation transfer function of patch-based stereo systems," in *CVPR*, 2012, pp. 1386–1393. 2

[11] K. Kolev, T. Brox, and D. Cremers, "Fast joint estimation of silhouettes and dense 3D geometry from multiple images," *TPAMI*, vol. 34, no. 3, pp. 493–505, 2012. 2

[12] S. Liu and D. B. Cooper, "A complete statistical inverse ray tracing approach to multi-view stereo," in *CVPR*, 2011, pp. 913–920. 2

[13] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *ICCV*, 2011, pp. 2320–2327. 2

[14] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004. 3

[15] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern recognition*, vol. 43, no. 8, pp. 2666–2680, 2010. 3

[16] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *CVPR*, vol. 1, 2003, pp. I–195. 3

[17] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, vol. 1, 2006, pp. 519–528. 1, 2, 3, 4

[18] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *CVPR*, 2008, pp. 1–8. 1, 2

[19] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Machine Vision and Applications*, vol. 23, no. 5, pp. 903–920, 2012. 2, 5, 6, 7, 8

[20] G. Vogiatzis, C. Hernández, P. H. S. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *TPAMI*, vol. 29, no. 12, pp. 2241–2246, 2007. 2

[21] A. Wendel, M. Maurer, G. Graber, T. Pock, and H. Bischof, "Dense reconstruction on-the-fly," in *CVPR*, 2012, pp. 1450–1457. 2

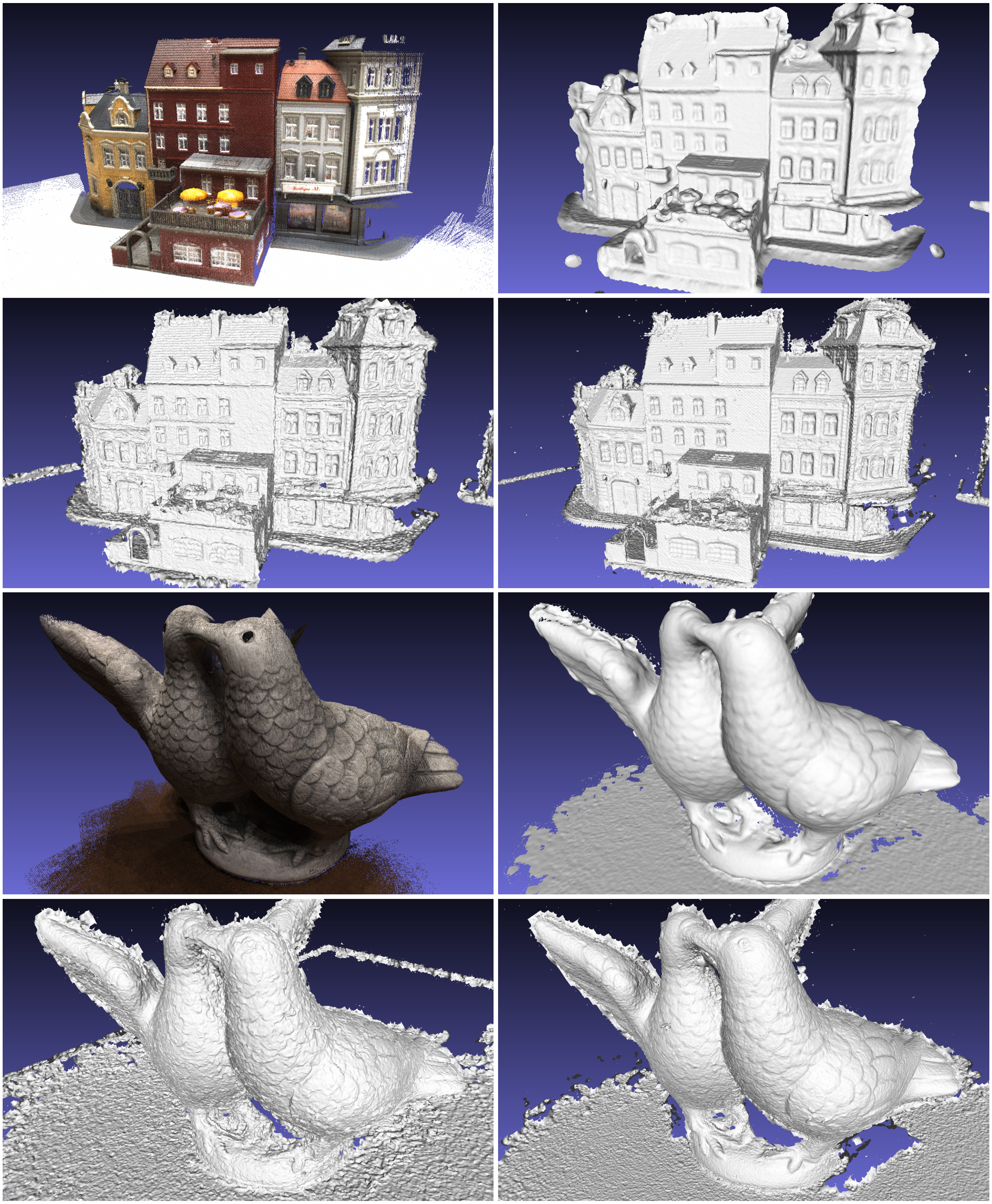


Figure 7. Reference points (upper left) and meshed surfaces of building where details are corrupted by the smoothing introduced by surface meshing. Upper right is [19], lower left is [3], and lower right is [2]. The statuette of doves is reconstructed following the same order. As with the building a slight corruption of detail is the result of surface reconstruction. In both scenes the artifacts around the edges are results of the surface reconstruction step and is not present in the point reconstruction.