Optimizing the Accuracy and Compactness of Multi-View Reconstructions

Markus Ylimäki, Juho Kannala, Janne Heikkilä

Center for Machine Vision Research,
P.O.Box 4500, 90014 University of Oulu, Finland
firstname.lastname@ee.oulu.fi

Abstract. Current evaluation metrics and benchmarks for multi-view stereo reconstruction methods mainly focus on measuring the accuracy and completeness and they do not explicitly measure the compactness, and especially the compactness-accuracy trade-off of the reconstructed models. To answer this issue, we present an evaluation method that completes and improves the existing benchmarks. The proposed method is capable of jointly evaluating the accuracy, completeness and compactness of a three-dimensional reconstruction which is represented as a triangle mesh. The evaluation enables the optimization of both the whole reconstruction pipeline from multi-view stereo data to a compact mesh and the mesh simplification. The method takes the ground truth model and the reconstruction as input and outputs an accuracy and completeness value as well as the compactness measure for the reconstructed model. The values of the evaluation measures are independent of the scale of the scene, and therefore easy to interpret.

Keywords: Multi-view stereo evaluation, compactness-accuracy trade-off, mesh optimization

1 Introduction

Multi-view stereo reconstruction methods, which create three-dimensional scene models solely from photographs, have improved a lot during the recent years [15, 17]. The focus of research has been shifting from basic algorithms to system aspects and large-scale models [1, 16, 18]. Currently there are automatic reconstruction pipelines which are able to produce compact mesh models of both outdoor and indoor environments from images [4, 19, 11, 12, 2, 3]. In addition, several companies, such as Google, Nokia HERE, and Acute3D, have shown interest and efforts towards city-scale models.

The compactness of mesh models is essential for storing and rendering large-scale reconstructions. In particular, for mobile device applications, the models should be light-weight and streamable, yet realistic and accurate. Thus, in order to advance the development of image-based modeling techniques further, there is a need for evaluation metrics and benchmarks that enable quantitative evaluation of trade-offs between the compactness and accuracy of the reconstructed models.

In fact, the recent progress of multi-view stereo has been largely driven by benchmark datasets, which have enabled quantitative comparisons of different methods. The Middlebury [14] and EPFL [15] datasets have been widely used, and recently a similar dataset with more scenes was proposed in [8]. However, the evaluation metrics used in these standard benchmarks have solely focused on measuring the reconstruction quality, i.e. accuracy and completeness, and do not explicitly measure the compactness of the models. Therefore the previous evaluation metrics can not be used for evaluating compactness-accuracy trade-off and are hence not suitable for jointly optimizing the accuracy and compactness of the results of the reconstruction pipeline.

The problem related to the lack of suitable evaluation metrics is reflected by the fact that most of the recent papers studying compactness aspects of reconstructions (e.g. [4, 11, 12, 2]) do not perform quantitative evaluations of the compactness or compactness-accuracy trade-off. In fact, in the papers [4, 2] the results are evaluated only visually.

In this paper, we address the aforementioned problem by proposing an evaluation method which is able to illustrate both the accuracy, completeness and compactness of the reconstructions with respect to the ground truth model. The method measures the accuracy and completeness jointly with the Jaccard index between the voxel representations of the ground truth model and the reconstruction. The compactness of the reconstruction is measured with a compression ratio representing the ratio of the number of vertices in the ground truth and the reconstruction. The relation between earlier evaluations and the proposed one is illustrated in Figure 1.

The proposed method is particularly suitable for evaluating the full reconstruction pipeline from images to a compact mesh but it can also be used to evaluate the following sub-tasks separately: (a) point cloud generation from a set of photographs, (b) surface mesh generation from a point cloud, and (c) surface mesh simplification. Further, the proposed evaluation metrics are not scale dependent and therefore the results are easy to interpret for different ground truth models. The proposed method is also versatile and flexible because both the ground truth model and evaluated reconstructions can be either point clouds or triangle meshes.

The rest of the paper is organized as follows. First, Section 2 presents the most essential related work. Then the evaluation method is descibed in more detail in Section 3. Evaluation results are presented and discussed in Section 4 and Section 5 concludes the paper.

2 Related Work

The first widely used benchmark dataset for evaluating MVS algorithms was the Middlebury Multi-View Stereo Data [14]. The data consists of two different scenes, both having three sets with a varying number of low-resolution images. The evaluation is available in the Internet[1] where anyone can submit their own

[1] http://vision.middlebury.edu/mview/eval/

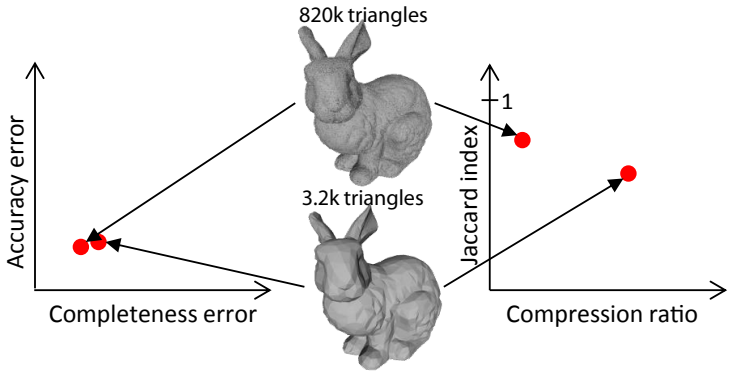

Fig. 1. Two mesh evaluations plotted with the current evaluation metric [8] (left) and with the proposed one. The Jaccard index joins the accuracy and completeness of the current metric and that enables the compression ratio, illustrating the compactness, to be shown in the same graph. According to the current evaluation metric the meshes are almost equal even though the latter mesh has over 250 times less triangles.

result for evaluation and compare the performance of their method with dozens of other MVS algorithms. Later, Strecha et al published the EPFL evaluation benchmark [15] consisting of more scenes with higher resolution[2]. The evaluation is no longer available, but the laser scanned ground truth models for two scenes are still downloadable on the web page. The recently published DTU dataset [8] further improved the existing benchmarks with totally 80 datasets covering a wider range of 3D scenes.

The current evaluation benchmarks evaluate the accuracy and completeness of the reconstructions. In [8], accuracy is measured as the distance from the reconstruction to the ground truth, and the completeness is measured from the ground truth to the reconstruction. Therefore, changes in the compactness of the reconstruction cannot be explicitly observed. The proposed method focuses on the compactness evaluation but still measures the accuracy and completeness jointly with the Jaccard index. As far as we know, this evaluation is the first of its kind, and thus, brings a new aspect for reconstruction evaluation in the future research challenges.

The main parts of a typical reconstruction pipeline from MVS data to a compact mesh, that need to be evaluated, are the point cloud creation and the meshing. In the pipeline, the consistency data of photographs is first converted into a three-dimensional point cloud using e.g. PMVS [6]. Then the point cloud is transformed into a surface mesh using methods like Poisson Surface Reconstruction (PSR) [9] or energy minimization approach [10]. Thus, both phases affect the compactness of the final reconstruction and can be optimized separately by the proposed evaluation metric.

The meshing process, if not already optimised, could be followed by mesh simplification [5] which tries to optimize the mesh by converting several trian-

[2] http://cvlabwww.epfl.ch/data/multiview/

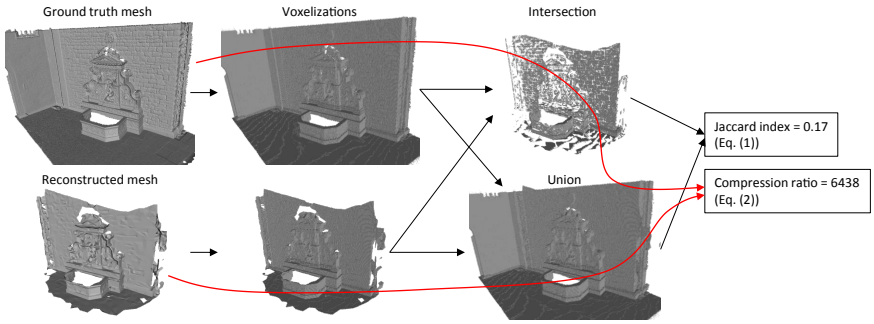

Fig. 2. An overview of the proposed evaluation pipeline with an example data. The ground truth and the reconstruction are converted to voxel representations. The Jaccard index is the ratio of intersection and union of the voxelizations. The compression ratio represents the ratio of the number of vertices in the ground truth and the reconstruction.

gles into one which follow the original surface as well as possible. One simple approach, presented in [13], clusters the vertices of the triangle mesh and then triangulates the cluster centres to form a new mesh with fewer faces. On the other hand, one of the top performing decimation methods, presented in [7], simplify the surface mesh by iterative contraction of vertex pairs (edges) so that the geometric error approximation, represented using quadric matrices, is maintained. However, although widely used, the method in [7] is almost twenty years old. With the proposed method, the mesh decimation algorithms can be quantitatively evaluated, which facilitates further developments.

3 Evaluation method

3.1 Overview

The proposed evaluation method takes the ground truth and the reconstruction as input and outputs two evaluation values: the Jaccard index J and the compression ratio R. The evaluation pipeline with an example data is presented in Figure 2. The input data can be meshes or point clouds. The Jaccard index illustrates the accuracy and completeness of the reconstruction, indicating the proportion of the ground truth mesh which is covered by the reconstruction within a certain threshold. The index is calculated using the voxel representations of the ground truth and the reconstruction and the threshold is the width of a voxel. The compression ratio illustrates the compactness of the reconstruction representing the ratio of the number of vertices in the ground truth and the reconstruction.

The proposed evaluation method consists of three phases: (1) the initialization of the voxel grids of the ground truth and the reconstruction, (2) transforming the ground truth and the reconstruction into voxel representations and (3) the actual calculation of the evaluation values. The following sections give more detailed descriptions of the phases.

3.2 Voxel grid initialization

The resolution of the voxel grids is defined by the width of a voxel and the size of the bounding box covering both the reconstruction and the ground truth. The width of a voxel is defined by the average distance between the ground truth vertices. That is, the width is twice as long as the median distance between a point and its k:th nearest neighbor. The value of k defines the sensitivity of the evaluation. Thus, too large voxel width smooths the details of the reconstruction and too small width causes holes in the voxelization of the ground truth. However, we found a value that can be kept as a default width for datasets for which the ground truth has a uniform vertex/point density. In all our experiments, the value of k was fixed to 10.

The bounding box is turned into a grid of voxels by dividing its dimensions with the defined voxel width. The grid is presented in an integer coordinate frame where every voxel has integer index coordinates. This grid is used in the voxel representations of the ground truth and the reconstruction.

3.3 Converting a mesh to a voxel representation

At first, both the ground truth and the reconstruction are converted to point clouds. The vertices of the ground truth mesh form the ground truth point cloud which is assumed to be dense, so that the average distance between points is below the voxel width. The mesh reconstruction is converted to a point cloud by sampling points on the triangles so that the density of the points matches the density of the ground truth vertices. Then, the point clouds are mapped into the integer coordinate frame of the voxel grid. The voxels are labelled as occupied if at least one point is inside the voxel or as unoccupied otherwise.

3.4 Calculation of evaluation values

The Jaccard index is calculated by comparing the voxel representations of the ground truth and the reconstruction. Lets denote the voxel grids of the ground truth and the reconstruction with $\mathbf{V_g}$ and $\mathbf{V_r}$, respectively. Now, the Jaccard index J is defined with the equation:

$$J = \frac{|\mathbf{V_g} \cap \mathbf{V_r}|}{|\mathbf{V_g} \cup \mathbf{V_r}|}, \tag{1}$$

where $|\cdot|$ means the number of voxels. Thus, the value of $|\mathbf{V_g} \cap \mathbf{V_r}|$ is the number of voxels which are occupied both in $\mathbf{V_g}$ and $\mathbf{V_r}$ and $|\mathbf{V_g} \cup \mathbf{V_r}|$ is the total number of occupied voxels in both grids. The Jaccard index is in the interval [0,1].

The compression ratio R is defined with the equation:

$$R = \frac{N_{GT}}{N_{REC}}, \tag{2}$$

where N_{GT} is the number of vertices/points in the ground truth model and N_{REC} is the number of vertices/points in the evaluated point cloud or mesh. Thus, the compression ratio illustrates the ratio of memory usage of the compared models.

4 Experiments

4.1 Overview

The experiments were carried out in three phases. First, the proposed evaluation method was tested with the range scanned data from the Stanford 3D Scanning Repository[3]. Then the results of the proposed method were compared with those of the DTU benchmark dataset in [8] and finally, a couple of evaluations were made with the EPFL dataset in [15]. The experiments are described in the following sections.

4.2 Stanford range scan dataset

In the first phase, the proposed method was tested with the range scanned data of the Stanford bunny. Notice that this kind of data does not contain the computer vision aspect but can still be used for benchmarking meshing and mesh simplification methods. The bunny data consists of ten scans which were transformed into the same coordinate frame to form a single point cloud. Then, the point cloud was turned into a triangular mesh using the Poisson Surface Reconstruction [9] (PSR) and our implementation of [10] (LAB). PSR was used with the default parameters except the Octree Depth which was set to 14. The parameters for LAB were $\alpha_{vis} = 32$, $\lambda_{qual} = 5$ and $\sigma = 0.001$. The meshes were then gradually decimated in Meshlab[4] using the Quadric Edge Collapse Decimation [7] (QECD) by halving the amount of triangles in every step. Meshes were also decimated with the Clustering Decimation [13] (CD) so that the step sizes were roughly the same as in QECD. All the meshes were evaluated with the proposed method. The results are presented in Figure 3. The x-axis illustrates the compactness as the compression ratio which represents the ratio of the number of vertices in the ground truth and the reconstruction (see Eq.2). Notice the logarithmic scale on the x-axis. The y-axis is the Jaccard index calculated with Equation 1. In addition to the meshes, also the point cloud, from which the meshes were created, was evaluated. That is presented as a single dot (PC) in the figure.

The figure clearly shows the difference between the decimation methods. That is, with CD the Jaccard indices of the meshes drops much earlier than with QECD. Also, when looking at a few left most meshes, i.e. the most complex ones, the LAB meshes have somewhat better Jaccard index and compression ratio in comparison with the PSR meshes.

The main reason for the Jaccard index difference between the evaluated point cloud and the mesh reconstruction is the fact that the evaluated point cloud is the ground truth model. In addition, the ground truth has some holes which do not appear in the reconstructions, and therefore, they also drop the index a bit. PSR tends to round the sharp edges of the reconstruction, and therefore, the LAB meshes have a somewhat better Jaccard index.

[3] http://graphics.stanford.edu/data/3Dscanrep/
[4] http://meshlab.sourceforge.net/

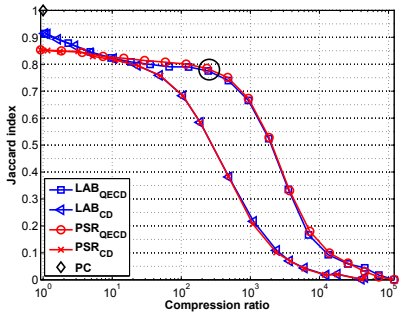

Fig. 3. Evaluation result of the Stanford Bunny. LAB and PSR refer to the mesh creation methods, i.e. [10] and [9], respectively. Subscripts QECD and CD refer to the used decimation methods, that is [7] and [13], respectively. PC is the point cloud which was used to create the meshes. The black circle indicate the decimated mesh presented in Figure 4.

Figure 4 shows the ground truth point cloud, the voxel presentation of it, the voxel presentation of the LAB mesh, the PSR and LAB meshes and a simplified version of the LAB mesh decimated with QECD. The voxel presentations of the ground truth and the reconstruction look very similar, because the reconstruction was created from the ground truth data. The voxelization of the ground truth is dense enough to preserve the details of the model but still sparse enough to make the voxelization uniform regardless of the minor misalignment issues of the scans. The decimated mesh has lost some details but has still relatively high Jaccard index as indicated with the black circle in Figure 3.

4.3 DTU multi-view dataset

In the second experiment, we illustrate the difference between the results of the proposed evaluation method and those of DTU benchmark [8]. We took the point cloud and mesh reconstructions (created with PMVS [6] and PSR [9], respectively) from the DTU package (House, scan no. 025) and did the decimations for the mesh with QECD [7] and CD [13], like with the range data in Section 4.2, and evaluated the meshes both with the proposed method and DTU method. The meshes were evaluated against the structured light reference (STL) provided by DTU. The results are presented in Figure 5. In DTU evaluations the accuracy and completeness are illustrated with mean distances in millimetres from the reconstruction to STL and from STL to the reconstruction, respectively.

As the results show, both evaluations are able to illustrate the difference between the decimation methods, but DTU does not explicitly show the compactness of the reconstructions. Also notice that DTU values are in millimeters, and thus, scale dependent and more difficult to interpret. In addition, due to the distance differences between the values, the interpretation of DTU result is not possible without a closer look of the first values.

Both methods give better accuracy or Jaccard index for the point cloud (PC) than the meshes. That happens because the mesh reconstructions contain both

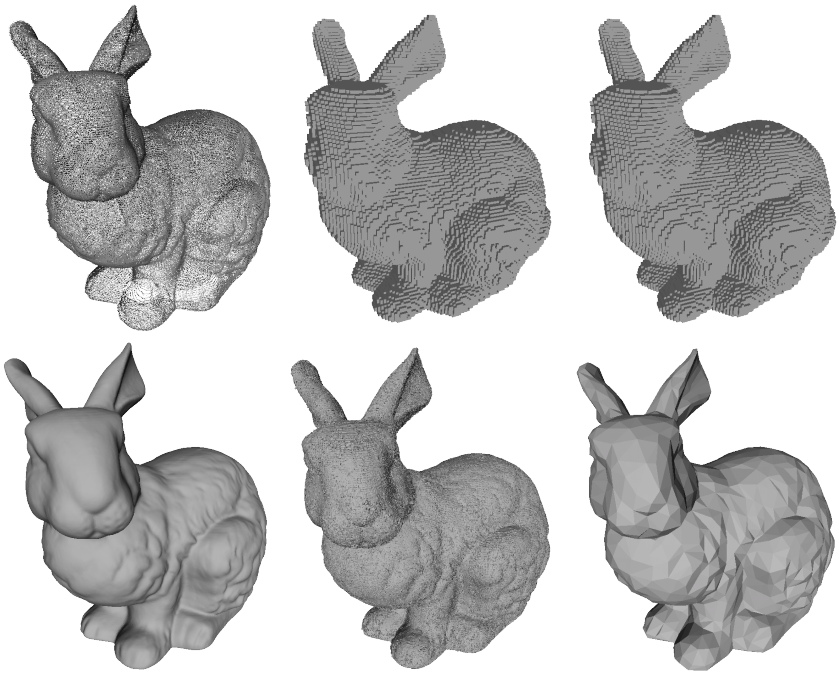

Fig. 4. Bunny voxelizations and mesh reconstructions. Top: the ground truth point cloud (left) and voxel representations of the ground truth and the LAB mesh reconstruction. Bottom: The triangle mesh of the reconstruction created with PSR [9] (left) and LAB [10] and the triangle mesh decimated from the LAB mesh by QECD.

correct and incorrect surfaces that exist neither in the structured light reference nor the point cloud, as illustrated in Figure 6. In other words, the ground truth scans are incomplete. Due to PSR meshing, the incorrectly reconstructed areas are mainly located at the outer boudaries of the reconstruction (black ellipses). However, PSR can also fill holes correctly (white ellipses).

The metric used in DTU benchmark is not able to detect pure compactness change. For example, regardless of the number of triangles (if at least two), the accuracy and the completeness of a rectangular wall do not change. That is, the more planar surfaces in the scene the less the compactness affects the evaluation results of DTU benchmark.

4.4 EPFL multi-view dataset

In the third phase, we performed evaluations using the two publicly available models from EPFL dataset [15]; Fountain-P11 and Herz-Jesu-P8. Point cloud reconstructions were first created using the PMVS program [6]. The number of points in the point clouds were then reduced to about 500k points in order to run our implementation of [10] in reasonable time. The same reduced point cloud was used in all mesh constructions and evaluations. Now, the evaluation results were obtained like with the range data in Section 4.2. The parameters for PSR

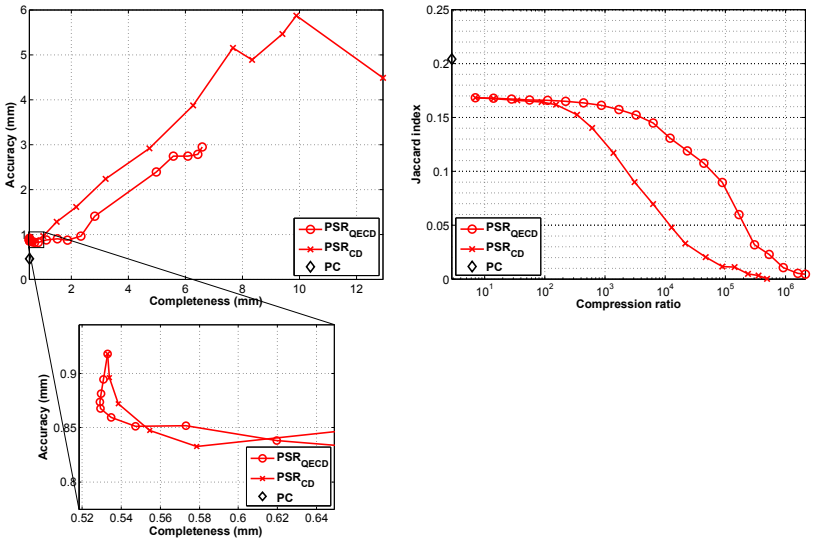

Fig. 5. Comparison of DTU evaluation (left) [8] and the proposed one (right). The original mesh provided in DTU dataset is created with [9] (PSR) and the decimations were made with [7] (QECD) and [13] (CD). PC is the point cloud which was used to generate the original mesh.

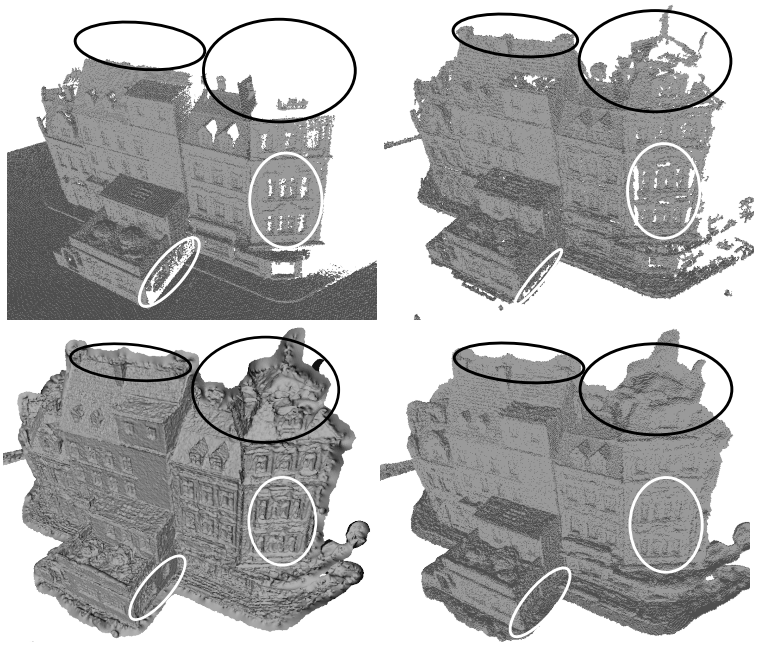

Fig. 6. Voxel presentations of the House. Top: Voxelization of the ground truth (left) and the PMVS point cloud. Bottom: the PSR triangle mesh of the reconstruction (left) and its voxelization. Ellipses highlight the areas which contain surfaces that are incorrectly (black) or correctly (white) reconstructed in the mesh (bottom) but do not appear either in the ground truth or the point cloud (top).

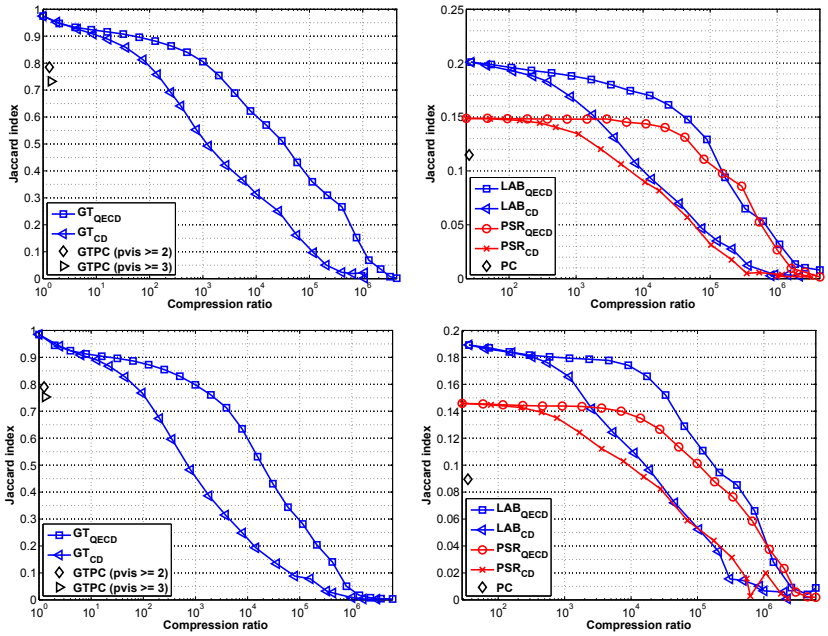

Fig. 7. Evaluations of Fountain-P11 (top) and Herz-Jesu-P8 (bottom) datasets. Left: The evaluations of the ground truth meshes (GT) decimated with [7] (QECD) and [13] (CD) and the evaluations of the ground truth vertices lying in the field of view of at least two or three cameras. Right: The evaluations of the LAB [10] and PSR [9] reconstructions decimated with QECD and CD and the evaluations of the point clouds which were used to generate the corresponding meshes.

and LAB were the same as in Section 4.2 except σ in LAB which was now fixed to 0.01. The point cloud (PC) which was used to create the meshes was also evaluated. The results are presented in Figure 7 (right).

In addition to the reconstruction evaluations, we performed the same decimations and evaluations for the ground truth meshes. Also, the point clouds containing only those vertices of the ground truths which are in the field of view of at least two or three cameras were evaluated. These point clouds illustrate the theoretical maximum part of the ground truth model which could be reconstructed using MVS methods. The results are presented in Figure 7 (left). Like the range data results, the results clearly show the difference between the decimation methods (QECD vs CD) as well as mesh creation methods (LAB vs PSR). The difference between the mesh and point cloud reconstructions is explained mainly by the sparsity of the point cloud. In addition to the sparsity, the difference between the point cloud reconstruction and the ground truth point cloud where points are in the field of view of three cameras (GTPC, pvis>=3), is explained by the noise, holes and missing regions in the PMVS point cloud (bottom left vs bottom right in Figure 8). The holes and missing regions result from self occlusions and certain textureless areas. Also, a possible misalignment between the ground truth model and images may cause errors in the PMVS point cloud and thus also in the mesh reconstruction.

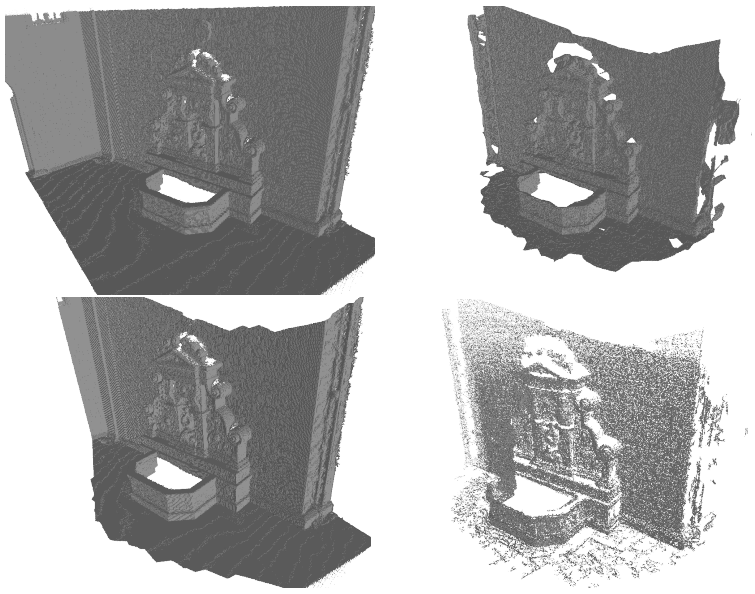

Fig. 8. Voxel presentations of the Fountain-P11. Top: Voxelization of the ground truth (left) and the reconstruction. Bottom: Voxelization of the ground truth vertices which are in the field of view of at least three cameras (left) and the voxelization of the PMVS point cloud. Notice the sparsity, noise, holes and missing regions on the right voxelizations compared with voxelizations on the left.

5 Conclusion

In this paper, we presented a method for the evaluation of multi-view stereo algorithms and triangle mesh decimations. The method enables the evaluation of the compactness-accuracy trade-off of the reconstructed models and thus completes and improves the existing evaluation benchmarks from Middlebury [14], by Strecha et al. [15] and Jensen et al. [8]. The proposed method facilitates optimization of both the reconstruction pipeline from MVS data to a compact triangle mesh and the mesh simplification. The method takes the ground truth model and the reconstruction as input and outputs the accuracy and the completeness of the model, presented with the Jaccard index, and the compactness measure. As presented in the experiments, the method can clearly illustrate the accuracy and compactness differences of certain meshes created with different meshing and decimation algorithms. In addition, the values of evaluation measures are independent of the scale of scene and can be used for any dataset with a ground truth model.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building Rome in a day. Communications of the ACM (2011)

2. Bodis-Szomoru, A., Riemenschneider, H., Van Gool, L.: Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
3. Cabral, R., Furukawa, Y.: Piecewise planar and compact floorplan reconstruction from images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 628–635 (June 2014)
4. Chauve, A.L., Labatut, P., Pons, J.P.: Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
5. Cignoni, P., Montani, C., Scopigno, R.: A comparison of mesh simplification algorithms. Computers & Graphics 22, 37–54 (1997)
6. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 32(8), 1362–1376 (2010)
7. Garland, M., Heckbert, P.S.: Surface simplification using quadric error metrics. In: 24th Annual Conference on Computer Graphics and Interactive Techniques (1997)
8. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
9. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Eurographics Symposium on Geometry Processing (2006)
10. Labatut, P., Pons, J.P., Keriven, R.: Robust and efficient surface reconstruction from range data. Computer Graphics Forum (CGF) 28(8), 2275–2290 (2009)
11. Lafarge, F., Alliez, P.: Surface Reconstruction through Point Set Structuring. Research Report RR-8174, INRIA (Dec 2012)
12. Lafarge, F., Keriven, R., Bredif, M., Vu, H.H.: A hybrid multi-view stereo algorithm for modeling urban scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 35(1), 5–17 (2013)
13. Rossignac, J., Borrel, P.: Multi-resolution 3d approximations for rendering complex scenes. In: Falcidieno, B., Kunii, T. (eds.) Modeling in Computer Graphics, pp. 455–465. IFIP Series on Computer Graphics, Springer Berlin Heidelberg (1993)
14. Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 519–528 (2006)
15. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
16. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications 23(5), 903–920 (2012)
17. Vu, H.H., Labatut, P., Pons, J.P., Keriven, R.: High accuracy and visibility-consistent dense multiview stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 34(5), 889–901 (May 2012)
18. Wu, C.: Towards linear-time incremental structure from motion. In: International Conference on 3D Vision (3DV). pp. 127–134 (June 2013)
19. Wu, C., Agarwal, S., Curless, B., Seitz, S.: Schematic surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1498–1505 (June 2012)