Consensus Multi-View Photometric Stereo

Mate Beljan Jens Ackermann Michael Goesele

TU Darmstadt

Abstract. We propose a multi-view photometric stereo technique that uses photometric normal consistency to jointly estimate surface position and orientation. The underlying scene representation is based on oriented points, yielding more flexibility compared to smoothly varying surfaces. We demonstrate that the often employed least squares error of the Lambertian image formation model fails for wide-baseline settings without known visibility information. We then introduce a multi-view normal consistency approach and demonstrate its efficiency on synthetic and real data. In particular, our approach is able to handle occlusion, shadows, and other sources of outliers.

1 Introduction

Recently, several multi-view photometric stereo techniques have been proposed that use some kind of global surface smoothness or similar assumptions [4, 9, 12]. In contrast, classical single-view photometric stereo is a very local technique allowing to recover even fine variations of the surface normals. In this paper, we analyze to what extent a multi-view reconstruction of surface orientation is possible if surface points are considered to be totally independent. We propose to move to a different scene representation consisting of surface points with normal information. Such a representation is especially suited for cluttered, real-world scenes with multiple, arbitrary objects.

One could argue that real-world objects typically show some kind of connectivity that should be exploited. However, the validity of any assumption on surface point interdependence is subject to the scale at which we observe the surface. For example, any approximation with linear or even curved patches only makes sense for a sufficiently low-frequent surface compared to the patch size. On the other hand, complete independence of surface points is a much more flexible paradigm. If we know that the data fulfills certain assumptions, such as smoothness, then it of course makes sense to exploit those. But if such properties are not known *a priori* it makes more sense to employ a general algorithm than to work with a method whose prerequisites are not met.

Similar arguments hold for explicit outlier removal: Modeling each error source individually, e.g., by shadow detection or special treatment of non-Lambertian points, can only address limited types of outliers. In contrast, the proposed hypotheses concept allows us to treat all outliers uniformly and to automatically select suitable subsets of observations that lead to consistent normals.

Our contribution lies in deriving a consistency measure that incorporates both position and surface orientation and does not rely on assumptions concerning surface smoothness. This new measure is based on the consistency of normal hypotheses that are generated from different subsets of all available views. In particular, we present a reconstruction approach that operates without an initial segmentation of object and background. It can handle scenes with an unknown number of objects, and is robust against shadows and occlusions by treating both uniformly as outliers.

1.1 Related Work

Multi-view stereo techniques [5, 14, 16] reconstruct a surface by comparing the texture of small patches. They typically fail on uniformly colored objects [15]. In contrast, photometric methods use shading information, e.g., based on the Lambertian image formation model to estimate the surface orientation. This is only possible if pixel correspondences in different views are known. In the single-view case these correspondences are given inherently. But for a multi-view data set this poses a fundamental challenge since knowing the pixel correspondences is equivalent to knowing the actual 3D position of the respective surface point.

As observed by Zhang *et al.* [19], most multi-view reconstruction techniques do not explicitly compute surface normals or even consider them in their matching cost. A notable exception are techniques that reconstruct specular surfaces [1, 13]. One of the first approaches was proposed by Coleman and Jain [3] who also introduced the concept of normal hypotheses. They propose to extend classical photometric stereo by taking four images from a fixed view-point under varying illumination, therefore generating four albedo and normal hypotheses per pixel. Only those hypotheses observing the surface under Lambertian conditions are correct and mutually consistent. Maki and Cipolla [10] extend this later to five images. We use the idea of hypothetical normals and extend it to the multi-view setting. This increases the number of hypotheses tremendously and introduces new challenges including computational complexity and occlusion handling.

Similar to us, Higo *et al.* [6] determine several normals at a single point from different sets of lighting directions, but for a single-view setting. They restrict the solution space to a cone spanned by these normals whereas we use them to find a maximal inlier set and then perform regular photometric stereo. Another line of research related to our work is the reconstruction of moving, diffuse objects. Maki *et al.* [11] present the theory and analyze the least squares intensity constraint. Their ideas use intensity subspaces and assume that the light source illuminates all points on the object, avoiding shadows and occlusions. Simakov *et al.* [15] also use the least squares error of the Lambertian image formation model as their consistency measure. The maximal rotation in their experiments is 50 degrees between a pair of views, but the whole set of views actually spans only a limited range of angles. Almost all surface points are seen in every view and therefore there are no outliers due to occlusion that would bias the least squares error. Lim *et al.* [9] investigate the question whether inaccurate initial pixel correspondences, e.g., from a rough, piecewise planar approximation, can

be improved by photometric stereo. They assume an evolving surface that arises from an integrable normal field and minimizes the least squares error of the Lambertian model. Similarly, Joshi and Kriegman [7] use a cost function based on the error of a rank three approximation of the observed image intensities to get an initial, smoothed surface. The main difference that separates our work from these approaches is that they all cannot handle occlusions, shadows, or other sources of outliers. This becomes, however, a necessity if we consider wide baseline scenarios with complex scenes.

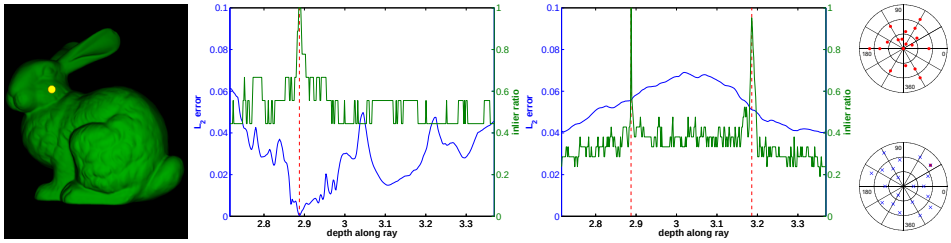
Several methods try to circumvent the problem of unknown occlusions or visibility with a proxy object that is refined iteratively. Weber *et al.* [17] apply a voxel-based approach. It relies on object silhouettes and iteratively carves away voxels outside the *consistency hull*. Due to visibility information from the evolving surface, their technique can also handle wide baseline scenarios. Similarly, Hernandez *et al.* [4] use a triangle mesh as proxy geometry and iteratively refine vertex positions and face normals. Again, the initialization and the illumination estimation rely on the visual hull being extracted from object silhouettes. Moses and Shimshoni [12] reconstruct a smooth 3D model from several calibrated images of a featureless Lambertian object under known point light illumination. Starting at an initial depth, the algorithm estimates neighbouring positions guided by the recovered normal at that point and then grows the surface by iterating this procedure. Yoshiyasu and Yamazaki [18] also use silhouettes and a mesh deformation approach. Their iterative optimization alternates between a mesh-based representation and oriented points to handle topology changes. In our work, we directly recover oriented points and do not assume an evolving surface. This is especially challenging since it implies that visibility information is not available, but allows us to reconstruct scenes even with multiple objects. Another distinguishing feature is that we are independent of object silhouettes and visual hulls which are difficult to recover in some scenes.

2 Multi-view Photometric Stereo

Outlier-free settings with a narrow baseline between views have been studied repeatedly and with good results [7, 15]. We assume that we have N images taken from known camera positions $\mathbf{C_i}$ under a point light source with constant intensity d and changing but known light position $\mathbf{L_i}$. In this case, for any voxel $\mathbf{p}$, we can find a normal $\tilde{\mathbf{n}}$ which minimizes the least squares error

$$e(\mathbf{p}) = \min_{\tilde{\mathbf{n}}} \frac{1}{N} \sum_{i=1}^{N} \left(I_{p,i} - d_{p,i}\mathbf{L}_{p,i} \cdot \tilde{\mathbf{n}}\right)^2 \tag{1}$$

defined by the corresponding pixel intensities $I_{p,i}$, light directions $\mathbf{L}_{p,i} = (\mathbf{L}_i - \mathbf{p})/\|\mathbf{L}_i - \mathbf{p}\|$ and light intensities $d_{p,i} = d/\|\mathbf{L}_i - \mathbf{p}\|^2$ in each of the N images. For large $e(\mathbf{p})$ we can be sure that the voxel is not on the surface S, since for $\mathbf{p} \in S$ the corresponding normal would explain all observations with a small error. For small $e(\mathbf{p})$, however, we cannot conclude the inverse since the error can be small for certain $\mathbf{p} \notin S$. This occurs, e.g., if the true surface points observed through



(a) (b) (c) (d)

Fig. 1: Error measures along a viewing ray (*blue:* L_2-error, *green:* inlier ratio for normal consistency). (a) The ray marked in the input image. (b) For narrow baseline (9 views) both measures indicate the correct depth (ground truth in *red*). (c) For 62 views (wide baseline) the inlier ratio detects even the intersection at the back whereas the L_2-error is not discriminant. (d) Upper and lower hemisphere with selected views for the front (*red*) and back (*blue*) intersections.

$\mathbf{p}$ have the same normal (e.g., for a planar target). The blue curve in Fig. 1(b) shows this L_2-error for voxels along a ray from the camera center of the reference view in Fig. 1(a). The ground truth depth of the surface point is indicated by the red line and we observe that the L_2-error is a useful indicator of the surface. This example was computed from nine synthetic images of the *bunny* (see Section 3) with camera positions from the same octant.

The illustration in Fig. 1(c), however, clearly shows that this error measure is not suitable for wide-baseline datasets. In this example, the same object is rendered from 62 views that are uniformly distributed on a sphere. There is no minimum at the ground truth location and the curve gives no indication of the true surface. In contrast our proposed technique (in green) which we will discuss in detail in the next section finds both front and back surface points reliably.

2.1 Consistency of Normal Hypotheses

The problem with the least squares error is that it treats all occurring errors equally and is thus not robust against outliers. In real scenarios with multiple views and wide baselines, an observed intensity might be an outlier due to occlusion, shadows, or for various other reasons. Since we have no visibility information available, it is important to robustly select a subset of suitable observations. We therefore introduce the concept of normal hypotheses that allows for a more fine-grained selection of reasonable observations and fits well into a RANSAC framework.

For a single voxel $\mathbf{p}$ and each possible triplet of intensities $\mathbf{I}$ arising from linearly independent light directions $\mathbf{L}_{p,i}$ we can solve the illumination equation

$$\mathbf{I} = \begin{pmatrix} d_{p,1}\mathbf{L}_{p,1}{}^T \\ d_{p,2}\mathbf{L}_{p,2}{}^T \\ d_{p,3}\mathbf{L}_{p,3}{}^T \end{pmatrix} \cdot \mathbf{n} \tag{2}$$

for $\mathbf{n}$. This yields a hypothesis for the albedo $\beta = \|\mathbf{n}\|$ and for the normal $\tilde{\mathbf{n}} = \mathbf{n}/\beta$. The case $N = 4$ with a fixed camera has been discussed by Coleman *et al.* [3] in the context of specularities. We consider arbitrary N and make use of the fact that the concept is independent of the fixed view-point assumption. Because we do not allow repetitions this results in $\binom{N}{3}/3!$ distinct possibilities. The similarity of these hypotheses can now provide an important clue. For $\mathbf{p} \notin S$ the actually observed surface points differ and thus the computed normal hypotheses will differ considerably. For points on the surface, most hypotheses will coincide except those obtained from false observations.

We propose to look at the set of supporting observations for each hypothesis. That means we build an inlier set $\mathcal{I}$ of views that are consistent with the normal $\tilde{\mathbf{n}}$ we estimated from views i_1, i_2 and i_3 by applying an angular threshold t_{ang}:

$$\left| \cos^{-1}\left(\frac{I_i}{\beta d_{p,i}}\right) - \cos^{-1}(\mathbf{L}_{p,i} \cdot \tilde{\mathbf{n}}) \right| < t_{\text{ang}}, \quad \forall i \in \mathcal{I}. \tag{3}$$

Instead of evaluating all triplets which becomes infeasible with an increasing number of images we use a RANSAC algorithm at each $\mathbf{p}$ to select the best supported hypothesis. We found, however, the above criterion alone to be ambiguous in some cases. If view i has actually observed the occluding surface point $\overline{\mathbf{p}}$ with normal $\overline{\mathbf{n}}$ and this normal is by chance in the cone defined by $\mathbf{L}_{p,i}$ and the hypothesis $\tilde{\mathbf{n}}$, then i will be included as inlier. For this reason we introduce an additional check for views in $\mathcal{I}$. We require all hypotheses $\tilde{\mathbf{h}} \in \{\tilde{\mathbf{n}}_{i_1,i_2,i}, \tilde{\mathbf{n}}_{i_1,i,i_3}, \tilde{\mathbf{n}}_{i,i_2,i_3}\}$ to deviate at most t_{ang} degrees from the candidate $\tilde{\mathbf{n}}$ before including i in $\mathcal{I}$:

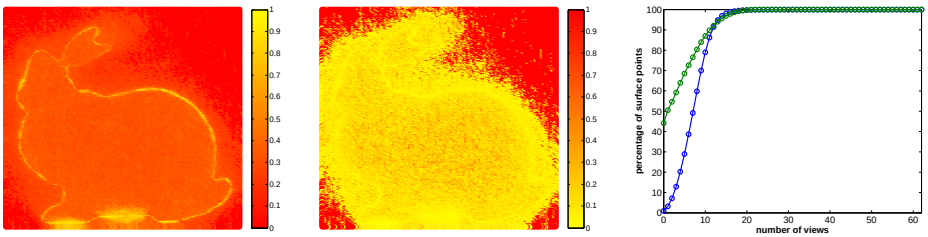
$$\cos^{-1}(\tilde{\mathbf{h}} \cdot \tilde{\mathbf{n}}) \leq t_{\text{ang}}. \tag{4}$$

With the triplet approach we can also easily perform further analysis on views. In particular, we check for 'front-facing' cameras and lights by evaluating the corresponding dot product before assigning them to the inlier set.

Fig. 1 shows our error measure along a ray from the reference camera: For each voxel the size $\#\mathcal{I}$ of the largest inlier set over all hypotheses is computed and then normalized with the maximal number of inliers encountered along the ray. From the ground truth overlay we conclude that this is a good indicator for whether or not a point is part of the surface. The maximum is attained at the correct position for both the narrow- and wide-baseline settings. We observe a second maximum in the wide-baseline case which is due to the back of the object. The inlier set in Fig. 1(d) is comprised of different views for each maximum. Finally, from the observations in the inlier set we can robustly re-estimate the normal at each voxel using least squares. So we not only obtain an indicator for the position of the surface but also its orientation.

3 Results

In this section, we present results for the consistency measure, recovered inlier views, surface orientation, and final 3D models. For these datasets, we determine the largest inlier set for all voxels in a regular 256^3 grid that we manually position in the scene. We found 500 RANSAC iterations at each point and $t_{\text{ang}} = 5^\circ$



(a) (b) (c)

Fig. 2: *Analysis:* (a) 2D slice showing the inlier ratio for the *bunny* data set normalized over the maximum inlier set of the slice (b) L_2-error for the same slice is not discriminant at all. The error is normalized over the maximum L_2-error of the slice. (c) Cumulative percentage of surface points plotted over the number of missed (*blue*) and erroneous (*green*) views.

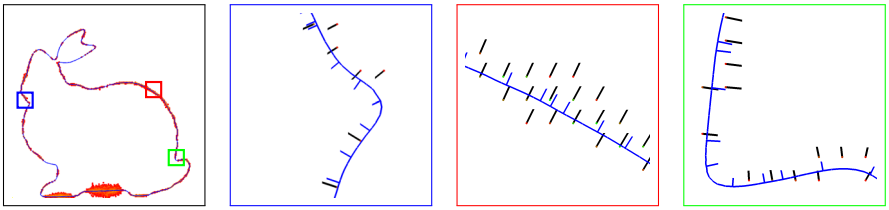


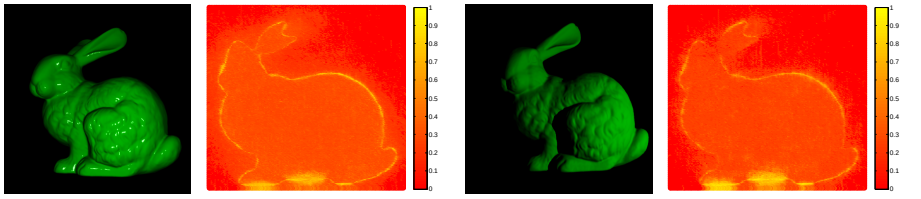
Fig. 3: *Normal Comparison:* Three surface regions with different curvature and visibility. The normals computed by our method (*black*) for the most consistent grid points ($\#\mathcal{I} > 15$) are close to the ground truth (*blue ticks*).

to yield good results — often fewer iterations suffice. Furthermore, we discard image triples with over- or underexposed pixels ($I_i < 0.05$ or $I_i > 0.95$).

Rendered Input Images To separately study the impact of various effects on the consistency measure, we first evaluate the performance of our technique on ray traced images with available ground truth data. We choose 62 light and coinciding camera positions uniformly distributed around the observed object. The 256^3 volume was computed in 11 hours on an Intel quad-core i7 960. Given the totally independent voxels further parallelization is trivial.

Extending Fig. 1, we now show a complete slice through the 3D volume in Fig. 2(a). Again, the size of the largest inlier set is a good surface indicator. The consistency is smeared out in regions with planar surfaces (e.g., the bottom of the bunny). This shading ambiguity is expected in the planar case. Nevertheless, the normals estimated for voxels in that region still yield the correct orientation. Fig. 3 shows the normals corresponding to the most consistent points in the grid together with the ground truth normals. Note that both the 3D points and the estimated surface orientation match the ground truth quite accurately.

We also investigated whether the least squares error for the normal estimated from the inlier set could be used as a discriminant error measure. Fig. 2(b) shows



(a) non-Lambertian (b) shadowed

Fig. 4: *Robustness through outlier handling:* Input image and a 2D slice of the consistency measure for (a) an object deviating from the Lambertian assumption and (b) a data set showing strong shadows.

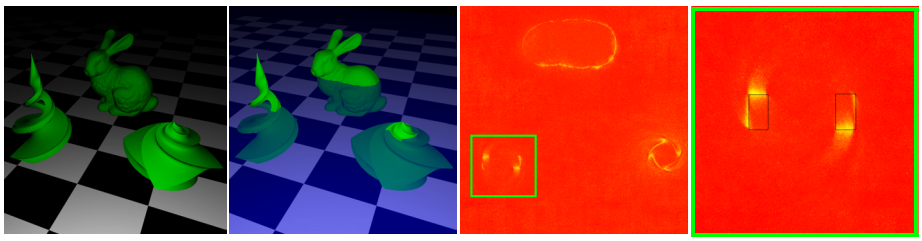


Fig. 5: *Multiple object scene.* Left to right: One of the input images. Visualisation of the selected 2D slice. Consistency in the 2D slice for a complete scene. Closeup of the spiral with ground truth overlaid in *blue*.

that this is not the case: RANSAC works so well that the inlier set leads to a small error for almost all voxels. The ability of RANSAC to select the correct subset of views at each voxel is illustrated in Fig. 2(c). For each point on the ground truth surface, we determine the views $\mathcal{I}_{\mathrm{GT}}$ that actually observe it. We then compute the number of missed views $\#\left(\mathcal{I}_{\mathrm{GT}} \setminus \mathcal{I}\right)$ and the number of erroneously included views $\#\left(\mathcal{I} \setminus \mathcal{I}_{\mathrm{GT}}\right)$. The cumulative histogram (blue) indicates that for about 50% of the surface points, we miss at most 7 views. The green curve shows that we select less than 5 erroneous views in 70% of the cases.

Our approach is also robust against different non-ideal conditions. Fig. 4(a) uses a plastic material with specular highlights in the reflection direction. In Fig. 4(b), we move the light sources away from the camera positions to add cast shadows to the images. Even the shiny material which violates the Lambertian assumption underlying photometric stereo has no apparent effect on the quality of the surface contour. The consensus-based reconstruction automatically discards those views as outliers that do not exhibit sufficiently diffuse behaviour. A similar argument holds for self-shadowing as in Fig. 4(b).

Finally, the flexibility of the consensus-based approach becomes most apparent when considering scenes with multiple surfaces instead of a single object. Fig. 5 shows our technique on such a challenging data set that we rendered with 89 views distributed on a hemisphere. The two arms of the spiral to the left are only partially recovered (see closeup). The missing parts are due to the surface facing downwards and are thus not seen by most of the views. The bunny is well recon-

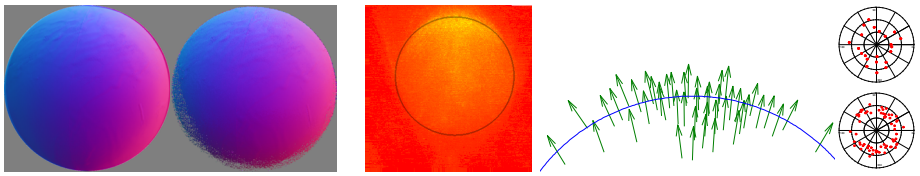
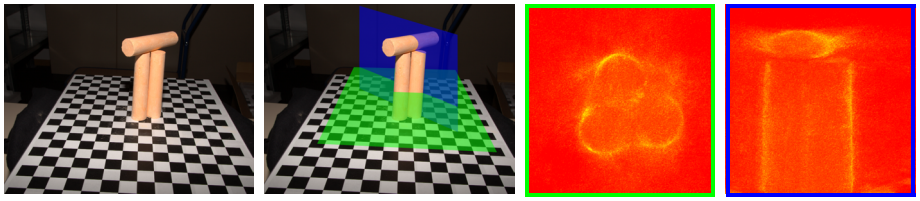


Fig. 6: *Real sphere.* Left: normal maps for the single-view case (classical least squares and our approach). Right: consistency, normals (*blue:* ground truth), and positions of camera (*top*) and light source (*bottom*) in the multi-view setting.



(a) (b) (c) (d)

Fig. 7: *Chalk dataset:* (a) Input image. (b) Visualisation of two slices. (c-d) Consistency measure for the horizontal and vertical slice.

structed and for the flower on the right we recover even the sharply twirled edges.

Results on Real Images We spray-painted a sphere with chalk and took several linear 14 bit images with a Canon EOS 5D after calibrating its intrinsic parameters. Computing the point light source positions from the reflections on five shiny spheres while keeping the camera fixed allows us to estimate the surface orientation using classical least squares as in Eq. (1). A multi-view photometric reconstruction technique should ideally fall back to the classical photometric stereo method if the camera positions $\mathbf{C}_i$ all coincide. The comparison in Fig. 6 proves that our algorithm recovers the correct surface normals in the single-view case. We also used the same setup to capture a wide-baseline multi-view dataset. In this case it is hard to determine the correct surface position from the smeared-out consistency in Fig. 6. The recovered normals, however, still follow the curvature even at voxels that are slightly off the surface.

To evaluate another multi-view reconstruction, we placed four pieces of chalk on a turnable checkerboard. 84 images were taken from different heights with a flash attached to the camera and pointing at the scene. We determine the extrinsic camera parameters with the camera calibration toolbox [2]. The light directions are then given by the fixed distance between flash and the optical axis of the camera (which we measured as 175 mm). In Fig. 7 we show the results for the *Chalk* dataset. Note that we do not compute an object mask or any occluding contours. With the proposed measure, the overall shape of the surface is clearly visible in the individual slices. We capture the smoothly curving parts as well as the sharp edge at the end of the sticks in Fig. 7(d). We observe, however, that a lot of voxels inside the object have quite high consistency scores. There is still

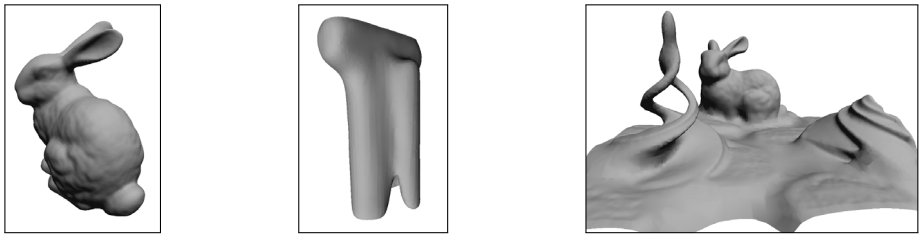


Fig. 8: *Rendering novel views:* Smooth surface extraction can be applied as an additional step if required, e.g. for rendering. (Thresholds used: $t_{cs} = 15/15/25$).

a difference between the consistency scores for on and off the surfaces but it is less pronounced than in the synthetic test cases. Some problems occur in surface regions with low local visibility, e.g., where the sticks touch in Fig. 7(c).

3.1 Surface Extraction

The aim of our work is to analyze to which extend the reconstruction of possibly disconnected, oriented points is possible for uniformly colored objects. But if required, the flexibility of our scene representation fits almost any surface extraction method as an additional step to create a triangle mesh or introduce regularization. We show this in Fig. 8 for a basic scheme of running Poisson surface reconstruction [8] on all points with $\#\mathcal{I} > t_{cs}$.

The reconstruction of the *bunny* appears slightly smoothed compared to the ground-truth. This is due to the regular grid and its limited resolution which prevent a faithful reconstruction of the high-frequency details. In future work, this could be addressed by an adaptive sampling since the measure itself does not rely on the uniform grid structure. For the *chalk* dataset, the regularization removes some of the sharp edges that were recognizable in the voxel-based reconstruction, see Fig. 7(d). Again the single pieces of chalk are easily discernible.

4 Discussion and Future Work

Abandoning almost all possible assumptions, e.g., baseline sizes, surface smoothness, visibility data, or object segmentation leads to a very challenging reconstruction problem. We have shown that reconstructions for narrow- and wide-baseline settings are possible even for scenes with multiple, discontinuous surfaces, but the lack of a per-voxel normalization limits the current approach. If we knew $\mathcal{I}_{\mathrm{GT}}(\mathbf{p})$, it would make sense to rather consider $\#(\mathcal{I} \cap \mathcal{I}_{\mathrm{GT}})$. Currently, if $\#\mathcal{I}_{\mathrm{GT}}(\mathbf{p}_1) \ll \#\mathcal{I}_{\mathrm{GT}}(\mathbf{p}_2)$ for two surface points, then usually $\#\mathcal{I}(\mathbf{p}_1) \ll \#\mathcal{I}(\mathbf{p}_2)$. That means that $\mathbf{p}_1$ will get a much lower score than $\mathbf{p}_2$ even if all non-occluded views are consistent. For example, the back of the bunny in Fig. 2 shows large consensus sizes while the paws are only seen in few views and have lower consistency values. Thus, we observe a certain dependency on the view distribution that we would like to investigate further. Another issue is, that the quality of the

real world results is not as good as we expected after experiencing the robustness on synthetic data. While the normals can be recovered quite accurately, the depth reconstruction merits further inspection. Finally, we would like to remark that the approach we presented will not outperform more specialized solutions if their respective prerequisites are met. But it is more flexible and we believe that this aspect will become even more important in the future.

Acknowledgments This work was supported in part by the DFG Emmy Noether fellowship GO 1752/3-1.

References

1. Bonfort, T., Sturm, P.F.: Voxel carving for specular surfaces. In: ICCV (2003)
2. Bouguet, J.Y.: Camera calibration toolbox for matlab. Available at http://www.vision.caltech.edu/bouguetj/calib_doc/. (2012)
3. Coleman, E.N., Jain, R.: Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. Computer Graphics and Image Processing 18, 309–328 (1982)
4. Esteban, C.H., Vogiatzis, G., Cipolla, R.: Multiview photometric stereo. PAMI 30(3), 548–554 (2008)
5. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: ICCV (2007)
6. Higo, T., Matsushita, Y., Ikeuchi, K.: Consensus photometric stereo. In: CVPR (2010)
7. Joshi, N., Kriegman, D.: Shape from varying illumination and viewpoint. In: ICCV (2007)
8. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction and its applications. In: Eurographics Symposium on Geometry Processing (2006)
9. Lim, J., Ho, J., Yang, M.H., Kriegman, D.J.: Passive photometric stereo from motion. In: ICCV (2005)
10. Maki, A., Cipolla, R.: Obtaining the shape of a moving object with a specular surface. In: BMVC (2009)
11. Maki, A., Watanabe, M., Wiles, C.: Geotensity: Combining motion and lighting for 3d surface reconstruction. IJCV 48(2), 75–90 (2002)
12. Moses, Y., Shimshoni, I.: 3d shape recovery of smooth surfaces: Dropping the fixed-viewpoint assumption. PAMI 31(7), 1310–1324 (2009)
13. Nehab, D., Weyrich, T., Rusinkiewicz, S.: Dense 3D reconstruction from specularity consistency. In: CVPR (2008)
14. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
15. Simakov, D., Frolova, D., Basri, R.: Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In: ICCV (2003)
16. Slabaugh, G.G., Culbertson, W.B., Malzbender, T., Stevens, M.R., Schafer, R.W.: Methods for volumetric reconstruction of visual scenes. IJCV 57(3), 179–199 (2004)
17. Weber, M., Blake, A., Cipolla, R.: Towards a complete dense geometric and photometric reconstruction under varying pose and illumination. In: BMVC (2002)
18. Yoshiyasu, Y., Yamazaki, N.: Topology-adaptive multi-view photometric stereo. In: CVPR (2011)
19. Zhang, L., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. In: ICCV (2003)