Lightweight Binocular Facial Performance Capture under Uncontrolled Lighting

Supplementary Material

Levi Valgaerts[1] Chenglei Wu[1,2] Andrés Bruhn[3] Hans-Peter Seidel[1] Christian Theobalt[1]
[1]MPI for Informatics [2]Intel Visual Computing Institute [3]University of Stuttgart

1 Additional Comparison to Valgaerts et al.

In these experiments, we show that our new scene flow method is an essential improvement over the method of [Valgaerts et al. 2010] for our specific purpose of facial performance capture.

Fig. 1 shows the comparison of our structure-aware smoothness term with the TV regularizer used in [Valgaerts et al. 2010] for the target frame of Fig.6 in the paper. The top results are obtained using TV regularized scene flow and the bottom results using scene flow with our proposed anisotropic regularizer. The two zoom-ins clearly show that anisotropic regularized scene flow produces more realistic tracked features, while TV regularized scene flow induces drift artifacts, such as double folding and degenerate triangles, in particular in the mouth and eye brow region.

Fig. 3 shows a comparison for the target frame depicted in Fig. 2. In the middle and the right column we show tracking results for the same coarse template mesh. The top results are again obtained using TV regularized scene flow and the bottom results using scene flow with our proposed anisotropic regularizer. As for the results in Fig. 1, anisotropic regularized scene flow produces more realistic tracked features, while TV regularized scene flow induces drift artifacts. The better performance of our structure-aware regularization strategy comes from its very selective smoothing behavior, which results in an overall smoother scene structure and scene flow, while still respecting semantically meaningful facial features. It is important to note that our regularization improves both the estimated stereo reconstruction and the estimated scene flow, and thus works on two complementary fronts: An improved stereo reconstruction results in a better assignment of the scene flow vectors in the tracking step, while the improved scene flow estimation leads to a more reliable deformation of the mesh geometry. An example of this is shown in the left column of Fig. 3, where we depict the stereo reconstruction obtained by the scene flow algorithm. We see that our result is smoother than the one produced by TV regularization, while we are still able to preserve sharp facial features. If we try to obtain a comparable level of smoothness with TV regularization, the scene structure and scene flow will suffer from for TV regularization well-known staircasing artifacts, such that important features will be lost. This is clearly visible in Fig. 2.

It also has to be noted that possible artifacts induced by a poorer scene flow estimation could be partly mitigated by choosing a higher weight for the Laplacian regularization of the geometry in the tracking pipeline. We have experienced, however, that increasing the amount of Laplacian regularization will lead to much less expressive facial motion and this is in most cases undesirable.

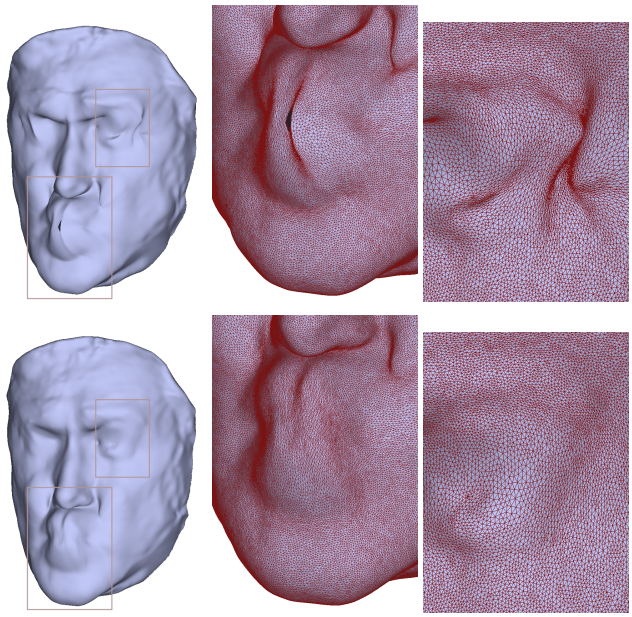

Figure 1: *Novel structure-aware smoothness terms. Top row: results obtained using [Valgaerts et al. 2010]. Bottom row: results obtained using our method. Left column: the tracked coarse mesh geometry. Middle and right column: triangle-overlaid zoom-in into the regions highlighted on the left. Note the better tracking of expressive features such as mouth and eyebrows using our method.*

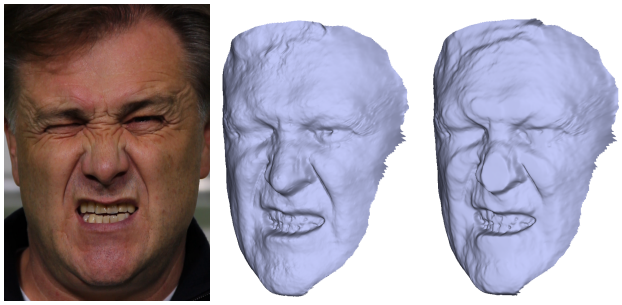

Figure 2: *Novel structure-aware smoothness terms. Left: the right target frame. Middle: the 3D reconstruction estimated by our scene flow method. Right: the 3D reconstruction estimated by [Valgaerts et al. 2010]. For a comparable level of smoothness, TV regularization will produce well-known staircasing artifacts (nose, lips, eyebrows).*

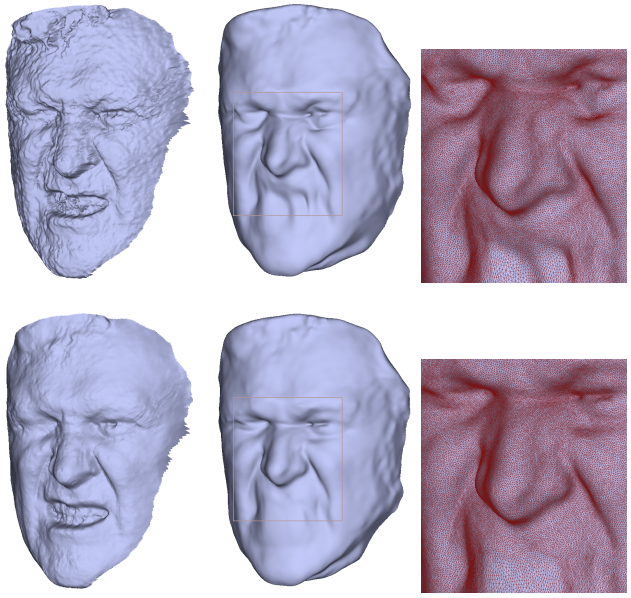

Figure 3: *Novel structure-aware smoothness terms. Top row: results obtained using [Valgaerts et al. 2010]. Bottom row: results obtained using our method. Left: the 3D reconstruction estimated by the scene flow. Middle: the tracked coarse mesh for the same frame. Right: triangle-overlaid zoom-in into the regions highlighted in the middle. Note the artifacts in the mouth and eye region and the slight deformation of the nose for TV regularization.*

2 Additional Results for Motion Refinement

Fig. 4 plots the normalized cross correlation (NCC) between the reprojected image and the corresponding input image for 25 frames of a similar sequence as the one shown in Fig.7 in Sec. 5.3 in the paper. For tracking with motion refinement (blue curve), we see that the NCC is consistently higher than without (red curve). From our experience, motion refinement is important for the realistic capture of expressive facial motion, such as the example of Fig.7 in the paper. For such cases, we found that one refinement step per time instance formed a good compromise between accuracy and computational complexity. This strategy is illustrated in the blue curve by an increase in NNC between each pair of consecutive frames. For less expressive motion such as speech, we found no large improvements in the estimated geometry. For such cases, motion refinement could be applied less frequently or even considered optional.

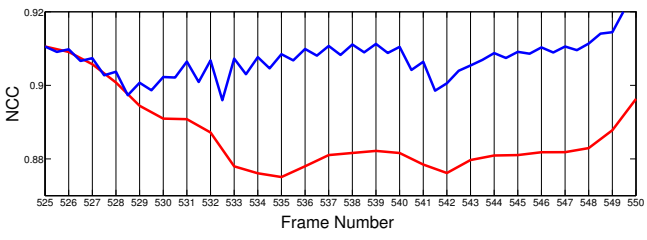

Figure 4: *Motion refinement. Graph of the NCC for a tracking result with (blue) and without (red) motion refinement.*

3 Additional Comparison to Wu et al.

In these experiments, we show that our new shape refinement method is an essential improvement over the method of [Wu et al. 2011] for the purpose of facial performance capture.

Because we can make use of fast sparse linear solvers, e.g. Cholesky decomposition, and because all vertices are optimized simultaneously in each iteration step, we achieve a general speed-up over patch-based non-linear optimization. This is depicted in Fig. 5, where the graph shows that our novel iterative minimization strategy reduces computation time by an order of magnitude compared to the non-linear patch-based optimization of [Wu et al. 2011] for the same sequence with constant parameters. The figure also shows that the proposed shape optimization strategy converges to a lower energy, and thus a better optimal shape, for most frames.

4 Additional Results and Validation

Here, we show additional results to the ones already shown in the main paper. Fig. 6 extends Fig. 11 from the paper by presenting reconstructions for two additional facial poses. Figs. 8 and 9 illustrate the acquisition of stereo sequences with a GoPro. Moreover, they extend Fig. 13 from the paper by showing additional reconstructions. Finally, Fig. 10 depicts a visual comparison between the result of our method and a laser scan for a similar pose.

5 Limitations

The shading-based refinement step can lead to artifacts on the boundaries caused by shadows. This is visible in the outdoor capture results of Fig. 9 around the nose and chin of the actor. In the video we demonstrate how we can reduce these artifacts, but this comes at the expense of less detail.

In Fig. 11 we show a failure case of the tracking pipeline in the presence of strong and fast moving shadow boundaries. The top row shows an actor leaving a building and moving from the shadow into the sunlight. The middle row shows the estimated scene flow for the top frames, where a red color indicates strong motion. Although there is hardly any motion of the face, the moving shadow boundary is erroneously interpreted as physical surface motion. As a result, the tracked mesh deforms in an unrealistic way, as depicted in the bottom row.

References

VALGAERTS, L., BRUHN, A., ZIMMER, H., WEICKERT, J., STOLL, C., AND THEOBALT, C. 2010. Joint estimation of motion, structure and geometry from stereo sequences. In *Proc. ECCV*, Springer LNCS, vol. 6314, 568–581.

WU, C., VARANASI, K., LIU, Y., SEIDEL, H.-P., AND THEOBALT, C. 2011. Shading-based dynamic shape refinement from multi-view video under general illumination. In *Proc. ICCV*.

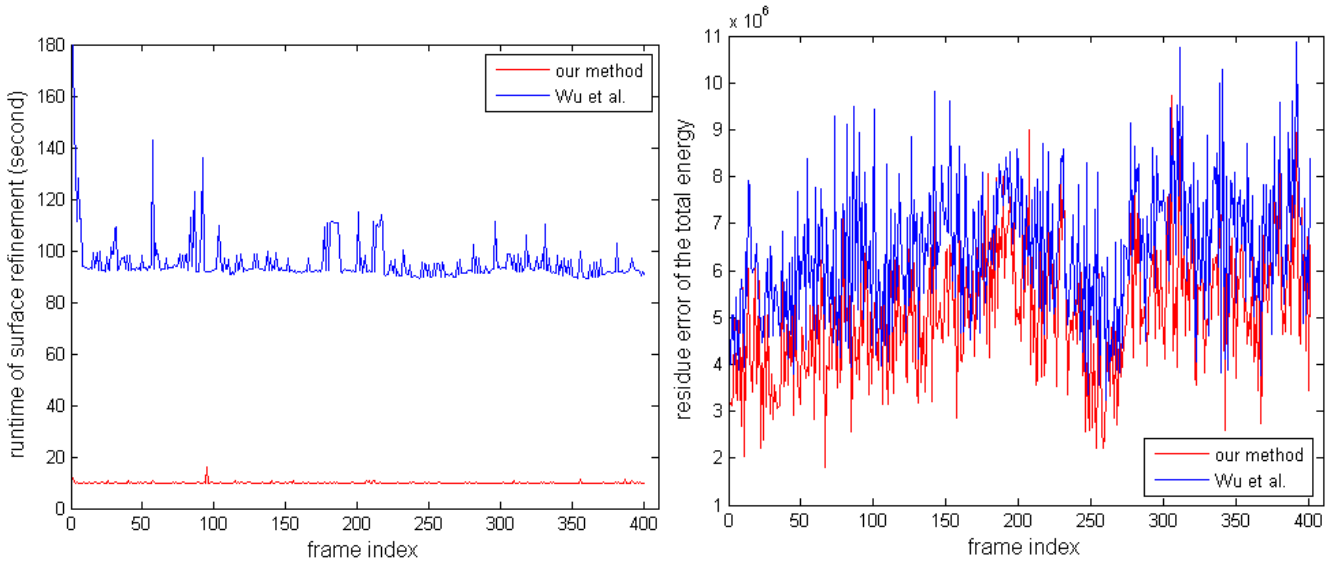

Figure 5: *Fast Iterative Minimization. Left: run time per frame for a sequence refinement with constant parameters. Right: shading energy per frame. Red: our optimization, blue: optimization of [Wu et al. 2011].*

Figure 6: *Results for a pair of Canon cameras. From top to bottom: the left input image, the corresponding reconstructed mesh, the mesh overlaid with a checkerboard pattern to demonstrate geometric coherence, the mesh colored using projective texturing. All results are taken from a single capture over more than 300 frames and are in full-vertex correspondence. They are shown in chronological order, starting from the template mesh at the far left.*

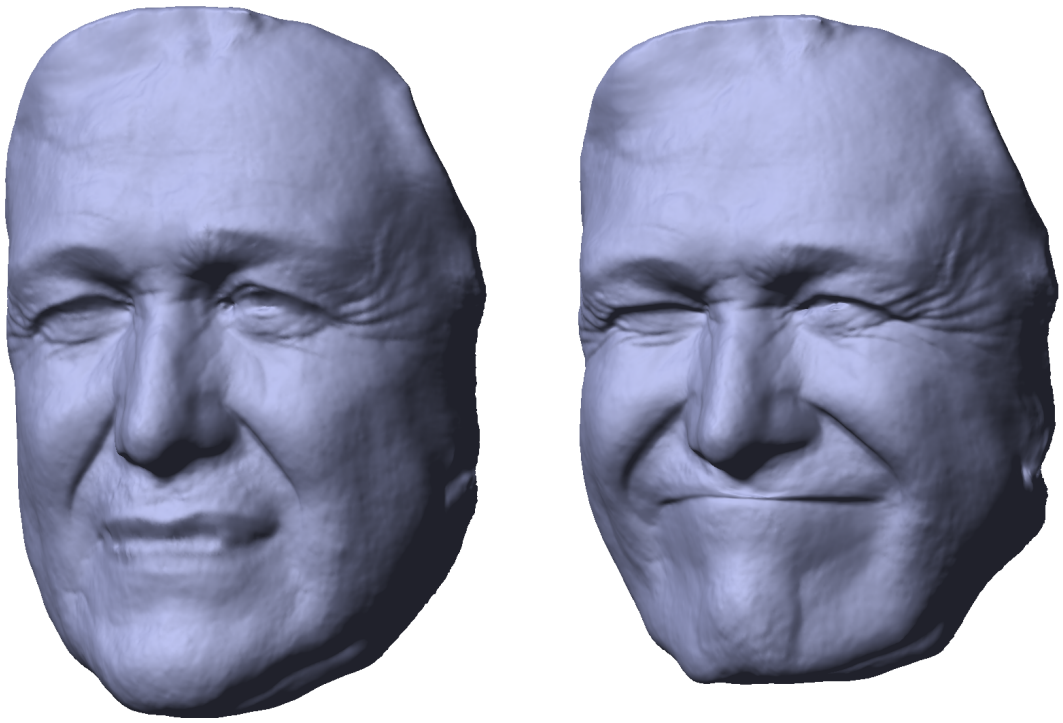

Figure 7: *Two enlarged results for the pair of Canon cameras that show the reconstructed surface detail.*

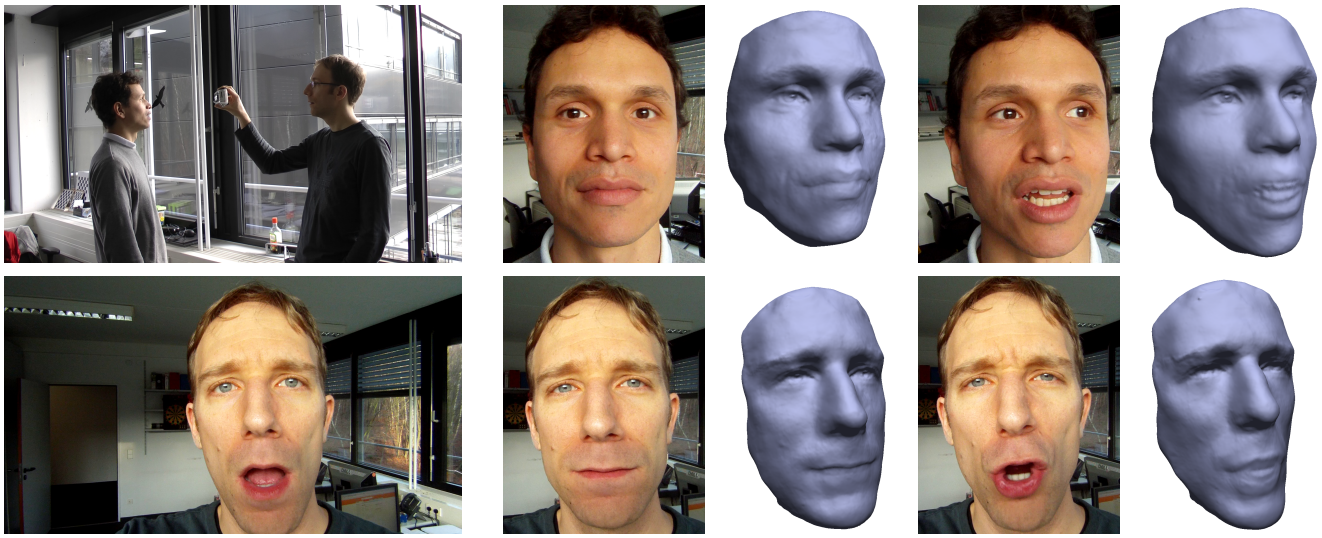

Figure 8: *Indoor results for a pair of GoPro HD helmet cameras. Top row from left to right: set-up of a person being recorded in an uncontrolled indoor environment, the starting frame and the template mesh, a captured result after approximately 100 frames. Bottom row from left to right: the set-up of a person recording himself (note the relatively small image region occupied by the face), the template mesh and a captured result after approximately 100 frames.*

Figure 9: *Outdoor results for a pair of GoPro HD helmet cameras. Top row: a person recording himself in an uncontrolled environment with bright sunlight. Captured results after approximately 100 and 200 frames. Bottom row: a person recording himself under changing illumination. Captured results after approximately 50 and 150 frames.*

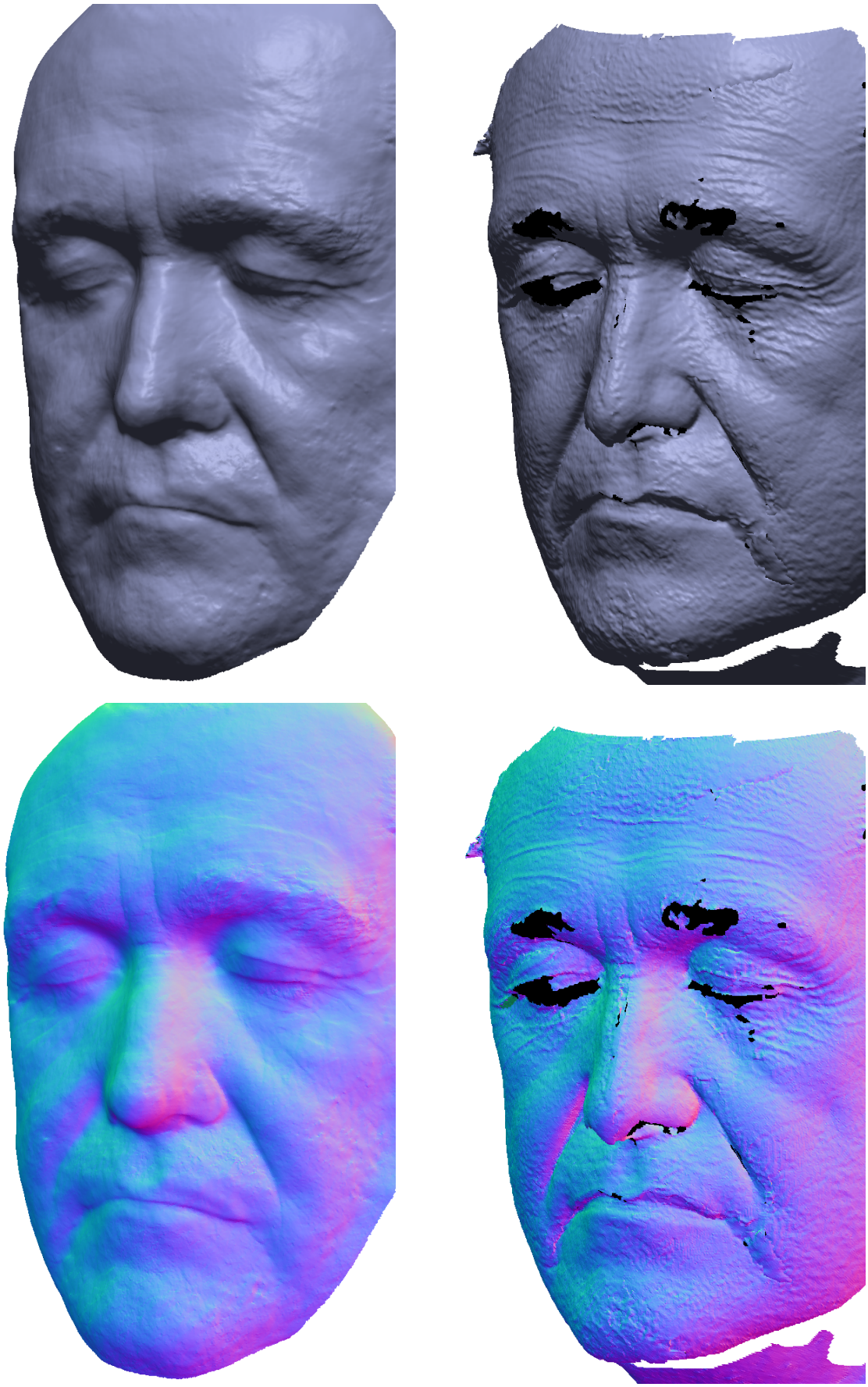

Figure 10: *Our reconstruction on a static face (top left) comes close to a laser scan of a similar face pose (top right, no hole filling done). The bottom row shows the meshes colored by normal orientation.*

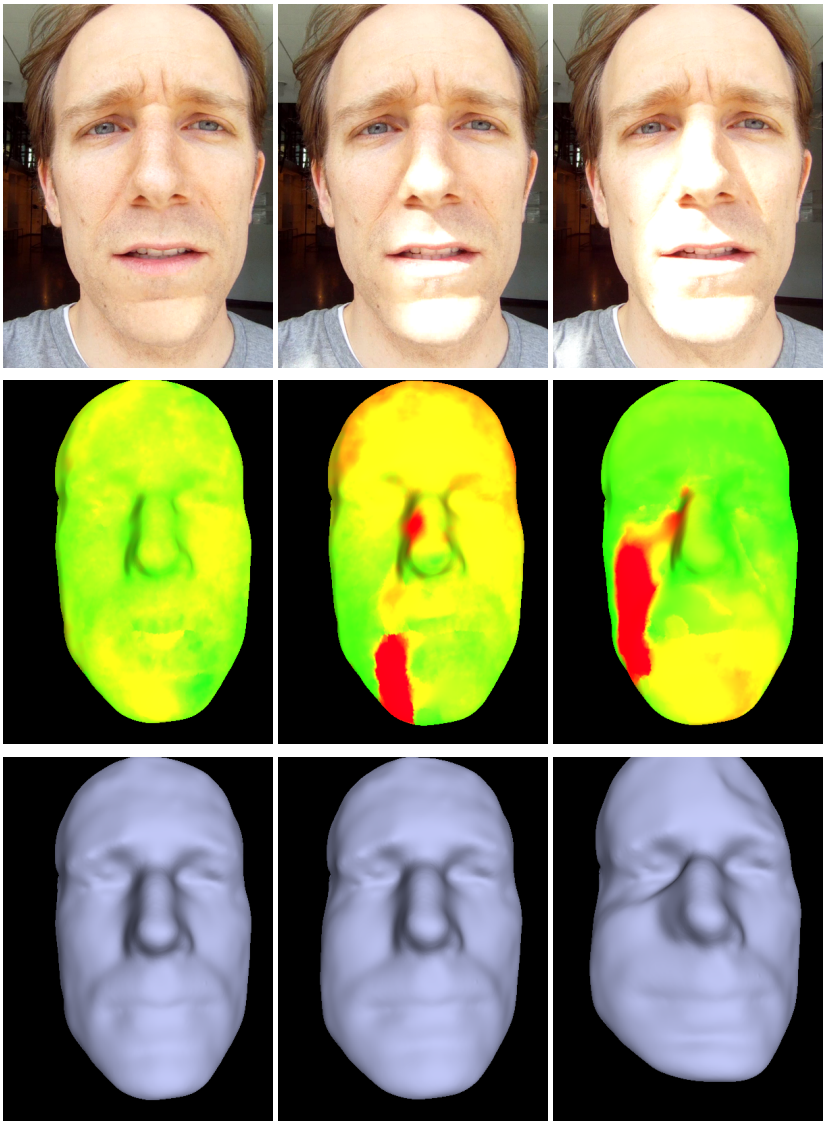

Figure 11: *Limitations of the tracking pipeline in the presence of strong moving shadows.*