Reconstructing Detailed Dynamic Face Geometry from Monocular Video

Supplementary Material

Pablo Garrido[1] Levi Valgaerts[1] Chenglei Wu[1,2] Christian Theobalt[1]
[1]Max Planck Institute for Informatics [2]Intel Visual Computing Institute

Personalized Blend Shapes and more results Fig. 4 shows a selection of expressions for the generic Emily model and for the four corresponding models derived from it which were used to generate the results shown in Figs. 5, 6, 7 and 8.

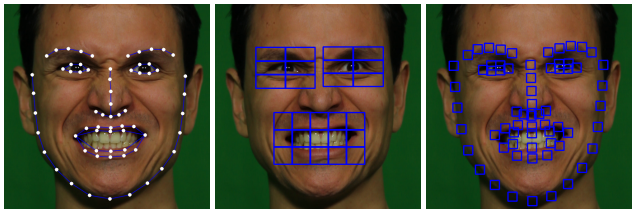

Figure 1: *Key frame selection.*

Key Frame Selection The middle pane in Fig. 1 shows the three rectangular regions of fixed size around the eyes and mouth, selected on an example frame after aligning it with the reference frame f^{t_0} (a neutral rest pose). These regions are used to build the LBP descriptors which are used to automatically find key frames depicting a similar expression as the reference frame. The right pane shows the smaller regions around each of the 66 tracked feature points used to find in-between key frames that share a local appearance with the reference frame around the facial features.

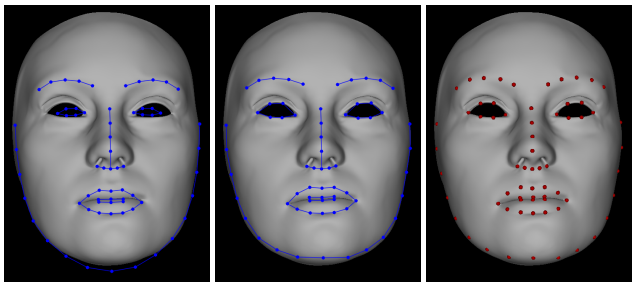

Figure 2: *Coupling the 2D and 3D model.* Left: *Features estimated by the feature tracker in a frontal view of the neutral generic blend shape pose.* Middle: *The manually corrected features.* Right: *The 3D feature vertices on the generic blend shape model.*

Coupling the 2D and 3D Model To couple the 66 sparse features that are tracked in the video to their corresponding 3D positions on the generic blend shape model, we render a frontal snap shot of the neutral pose and use the feature tracker to estimate the facial features. This works for a shaded OpenGL rendering of the model with constant material in front of a black background, but the detected features still need minor manual correction for better alignment. For the Emily blend shape model, the eyes are unnaturally large and these detected features need correction. As the 2D features are the projections of the corresponding 3D points on the blend shape model, correspondences can be easily established by back projection on the mesh. Since all personalized blend shape models used in our results are derived from the same generic Emily model, the indices of the found set of 3D feature vertices has to be the same for all actors. Thus, this step only needs to be completed once and only has to be repeated if a different generic face model is used.

Figure 3: *Comparison of our method with the binocular method of [Valgaerts et al. 2012].* Left to right: *target frame with fast head rotation, binocular result, our result.*

Table 1: *Comparison with binocular method. The average Euclidean distance in mm between the nearest vertices on the meshes reconstructed by the binocular method of [Valgaerts et al. 2012] and our monocular method for the results of Fig. 5 and Fig. 6. The distance was computed over all visible vertices of the 200k vertices in our reconstructions. This Euclidean distance is also visualized in the figures as a heatmap overlay (see error scale in the paper).*

Sequence	Average distance (mm)	Average maximum distance (mm)
Fig. 5 (over 565 frames)	1.71	7.45
Fig. 6 (over 402 frames)	2.91	9.82

Comparison with Binocular Reconstruction The 3D reconstruction quality of the binocular method of [Valgaerts et al. 2012] is quite high, but our monocular method is also able to capture high frequency detail and produces very accurate overlays (see also the comparison in the main paper and the video). In Tab. 1, we provide quantitative results for the comparison in the main paper. It lists the average Euclidean distance between the nearest visible vertices on the binocular and monocular meshes for the sequences of Fig. 5 and Fig. 6 for a mesh size of 200k. The deviation of our monocular result from the binocular results lies in the millimeter range despite the lack of direct depth information. A color coded overlay of this distance for the first sequence is shown in Fig. 5. For this comparison, the nearest vertices between the binocular and monocular result were recomputed for each frame, thus highlighting the shape reconstruction accuracy. However, if we determine the nearest vertices in the reference frame and keep them fixed over all other frames, the average Euclidean distance for the sequence of Fig. 5 becomes 3.27mm. This is because any tangential drift between the monocular and binocular result is additionally measured.

Another comparison of our monocular method with the binocular result is shown in Fig. 3 for a frame of the sequence of Fig. 7, which depicts fast rotating head motion. As reported by Valgaerts et al. [2012], purely mesh-based binocular methods are sensitive to occlusions and drift in the presence of strong apparent out of plane head rotation, leading to unnatural deformations in some frames. Our monocular method, on the other hand, robustly tracks a parametric face model and only leaves the blend shape space in the expression correction step by computing a small deformation field.

References

VALGAERTS, L., WU, C., BRUHN, A., SEIDEL, H.-P., AND THEOBALT, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM TOG (Proc. SIGGRAPH Asia) 31*, 6, 187:1–187:11.

Figure 4: *Personalized blend shape models.* Top row: *9 out of 79 expressions of a blend shape model created by an artist, including the neutral pose (courtesy of Faceware Technologies). The neutral expression is shown on the left.* Next four rows: *the same 9 expressions for the derived blend shape models of the actors in our four test sequences (see experimental section in the main paper Fig. 5, 6, 7 and 8). All meshes share the same number of vertices and triangulation. All models span the same semantic dimensions.*

Figure 5: *More overlay results for the first sequence of the main paper (560 frames). The two first rows show the input sequence and our monocular result. The third row shows a comparison with the binocular result of [Valgaerts et al. 2012] by means of a heatmap overlay of the Euclidean distance in mm between the nearest vertices on the meshes reconstructed by the monocular method and the binocular method. The color code ranges from 0mm (green) to 10mm (red), with yellow denoting the mid range 5mm (see also the error scale in the paper).*

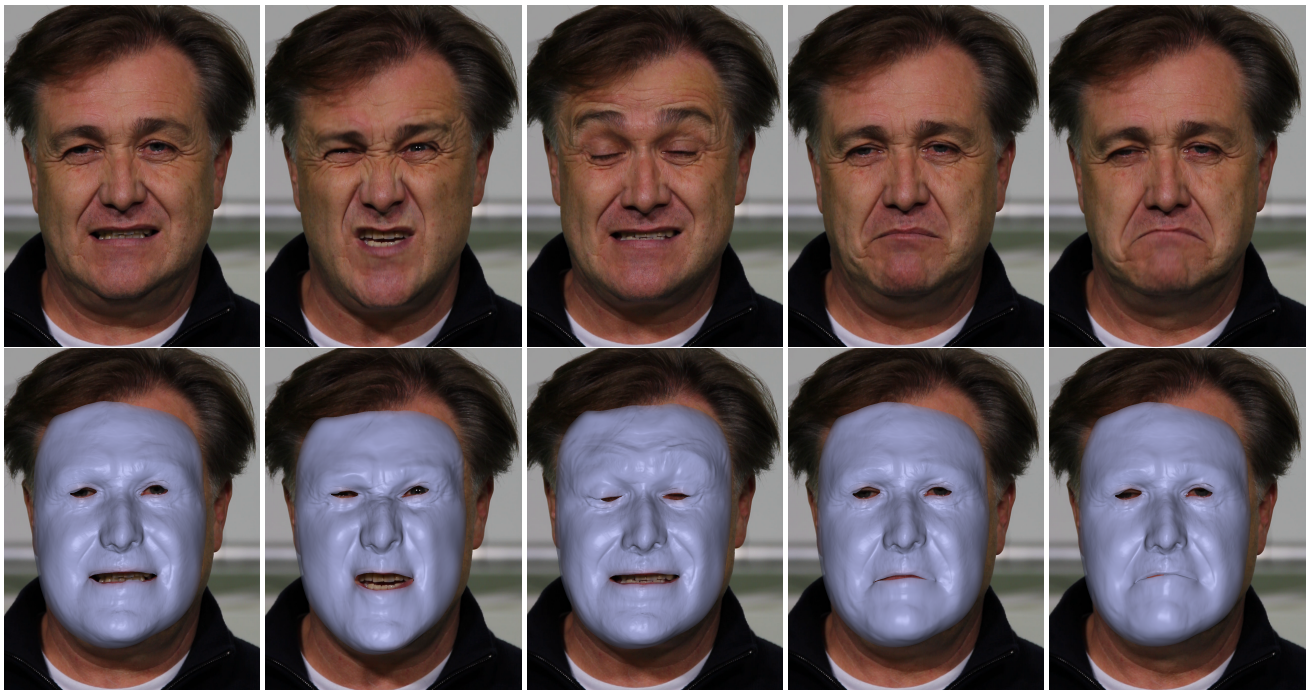

Figure 6: *More overlay results for the second sequence of the main paper (620 frames).*

Figure 7: *More overlay results for the third sequence of the main paper (1000 frames).*

Figure 8: *More overlay results for the fourth sequence of the main paper (650 frames). This is an outdoor sequence recorded under unknown lighting with a lightweight camera system featuring challenging head motion. The right column shows a failure case where our method does not estimate the pose and expression correctly. The supplementary video shows that our method fully recovers afterwards.*