# Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints – Supplementary Material –

Vikramjit Sidhu[1,2] Edgar Tretschk[1]
Vladislav Golyanik[1] Antonio Agudo[3] Christian Theobalt[1]

[1]Max Planck Institute for Informatics, SIC [2]Saarland University, SIC
[3]Institut de Robòtica i Informàtica Industrial, CSIC-UPC

In this supplementary document, we provide more details on the evaluation with noisy point tracks (Sec. 1), comparison to FML [8] (Sec. 2), comparisons on the Kinect sequences [9] (Sec. 3), applications of the trained auto-decoder (Sec. 4) as well as an overview of the parameters used in our experiments (Sec. 5).

## 1 Evaluation with Noisy Point Tracks

We also evaluate the accuracy of our method in the presence of noisy point tracks on the *actor mocap* sequence [4]. We follow the methodology described in [6], *i.e.,* we add Gaussian noise with 0 mean and standard deviation $\sigma_n = r \max |\mathbf{W}_s|$ with $r$ varying depending on the level of noise being added. We find that compared to the initial point tracks without added noise, $e_{3D}$ grows by 26% for $r = 0.01$ and 31% for $r = 0.02$. We observe that the shapes degrade gradually. Next, we also find that after adding uniform noise with $\sigma = 3$ pixels, the period of the composite sequence reconstructed in Sec. 5.2 of the main matter is correctly detected, and the result is similar to those shown in Fig. 3 of the main paper. This experiment indicates that our method is robust to noisy point tracks also in the period detection task.

## 2 Alignment of FML Shapes to the 3D Ground Truth

We compare our N-NRSfM to FML [8] on the *actor mocap* sequence [4] which provides ground truth rendered images, ground truth 3D shapes and ground truth dense point tracks. Since FML [8] is an image-based monocular 3D reconstruction technique based on a 3D morphable face model [2], its reconstructions have a different number of points ($\sim$6,000) compared to the ground truth of the *actor mocap* ($\sim$3,000). FML [8] covers a larger head area, including the neck and ears. Hence, to calculate $e_{3D}$ for the FML [8] on the *actor mocap* sequence, we have to align the FML [8] reconstructions with the ground truth meshes.

We follow a multi-stage alignment approach for partially-overlapping shapes. We first re-scale and rotate the FML [8] reconstructions so that they roughly match the scale and orientation of the ground truth. Second, we apply orthogonal

Procrustes alignment in the local coordinate system (*i.e.,* the shapes are registered to the origin of the coordinate system) and register FML [8] meshes and ground truth using several manually selected prior landmarks, these landmarks can be seen in Fig 1.

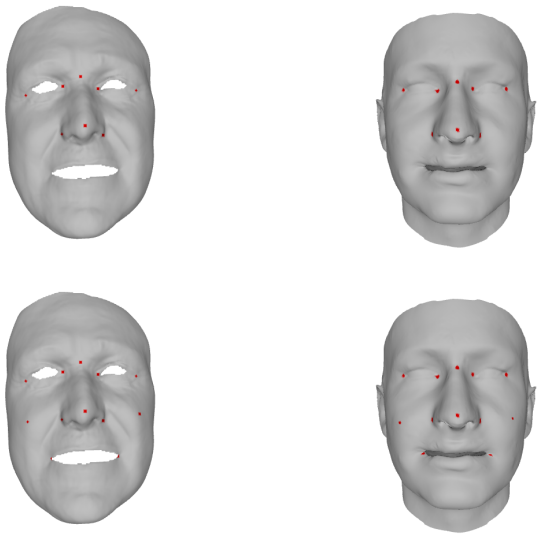

**Fig. 1.** Landmarks used to align the ground truth meshes of the *actor mocap* sequence (left column) and the FML reconstructions (right column). The top row shows the landmarks for the initial Procrustes alignment step, and the bottom row shows the landmarks for the final orthogonal Procrustes alignment and translation resolution.

Finally, we resolve the translation using the same set of landmarks. After the alignment, we establish point correspondences between the FML [8] reconstructions and ground truth shapes using a nearest-neighbour rule with the help of a $k$-d tree and compute $e_{3D}$. Note that this computation favors FML [8] since $e_{3D}$ of N-NRSfM is computed using fixed ground truth correspondences instead of nearest neighbours. Some selected final alignments can be seen in Fig. 2.

## 3 Evaluation with the Kinect Sequences [9]

We follow the evaluation methodology for the Kinect sequences [9] proposed in [6]. We crop the colour images using the rectangular bounding boxes defined by four points with the coordinates $\{[253\ 132]^\mathsf{T}, [253\ 363]^\mathsf{T}, [508\ 363]^\mathsf{T}, [508\ 132]^\mathsf{T}\}$ across 193 frames of the *paper* sequence and four points with the coordinates $\{[203\ 112]^\mathsf{T}, [203\ 403]^\mathsf{T}, [468\ 403]^\mathsf{T}, [468\ 112]^\mathsf{T}\}$ across 313 frames of the *t-shirt* sequence. The lower-left corner of the images is the origin of the coordinate system.

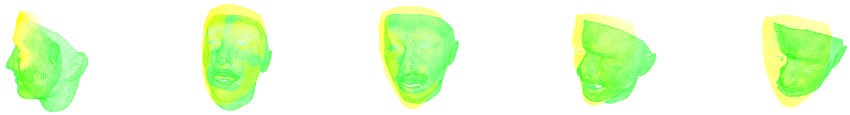

**Fig. 2.** Selected final alignments of the FML reconstructions with the ground truth meshes of the *actor mocap* sequence. The green points represent the FML reconstructions; the points in yellow represent the vertices of the ground truth meshes.

The same bounding boxes are used to crop the reference depth measurements from the depth maps for the evaluation. Furthermore, while computing $e_{3D}$, we filter outliers in the Kinect data by removing points which exceed a predefined distance from their positions in the reference frame. These distances are set to 15 and 30 Kinect depth units for the *paper* and *t-shirt* sequences, respectively.

## 4 Additional Applications of Trained Auto-Decoders

We present here additional applications of a trained auto-decoder. We first use an auto-decoder pre-trained on the synthetic faces sequence [3] with *traj. A* to reconstruct shapes from the synthetic faces observed under *traj. B*. We keep the weights of the auto-decoder fixed and optimise only for the camera poses and latent codes. See Fig. 3 for the selected reconstructions. Some of the facial expressions are recovered correctly, whereas the remaining ones are more dissimilar to ground truth. This preliminary experiment opens up a new direction of learning a category-specific shape auto-decoder instead of training for on each new sequence. This policy also saves training time. In the considered scenario, updating only the latent codes and camera poses is ∼four times faster compared to the training from scratch.

Second, we recover 3D shapes for point tracks of single frames. We optimise the latent code and the camera pose using $\mathbf{E}_{\text{data}}$ only and an auto-decoder pre-trained on a shorter version of several sequences. The 3D shape recovery takes 6, 9 and 59 seconds for the *back* [7] (20,000 points), *actor mocap* [4] (35,000 points) and *barn owl* [5] (203,000 points) sequences, respectively.

## 5 Reproducibility of the Quantitative Results

The hyperparameters used to obtain the best results on the quantitative sequences are summarised in Table 1. The weight of the data term $\mathbf{E}_{\text{data}}$ is kept fixed at $10^2$, the spatial smoothness term $\mathbf{E}_{\text{spat}}$ consists of the Laplacian term with the weight $\gamma$ and the depth control term with the combined weight $\gamma\lambda$. Since none of the quantitative sequences is periodic, we set $\omega = 0$ for these experiments. Since the camera in the Kinect sequences is not moving, we use this as a prior and keep the camera poses fixed during the optimisation. On repeating

**Fig. 3.** Experiment with the category-specific shape decoder. We use a shape auto-decoder trained on the synthetic face with *traj. A* to reconstruct shapes observed under *traj. B*. The top row contains reconstructions on *traj. B* using the auto-decoder trained on *traj. A*, the middle row contains the reconstructions using an auto-decoder trained for *traj. B*, and the bottom row shows the ground truth shapes for *traj. B*. Note the difference in the reconstruction accuracy of facial expressions.

the experiments, the obtained $e_{3D}$ is within $3 \cdot 10^{-3}$ of the values reported in Table 1. In the experiments involving the periodicity prior, we set $\omega = 1$ and disable the trajectory term.

**Table 1.** Hyperparameters leading to the lowest $e_{3D}$ on the quantitative sequences.

| dataset | $B$ | $\gamma$ | $\gamma\lambda$ | $\eta(K)$ | $\beta$ |
|---|---|---|---|---|---|
| *actor mocap* [4] | 32 | $10^{-6}$ | 0 | 0 (NA) | 1 |
| *traj. A* [3] | 10 | $10^{-4}$ | 0 | 10 (30) | 1 |
| *traj. B* [3] | 32 | $10^{-4}$ | $10^{-7}$ | 1 (40) | 1 |
| *expressions* [1] | 32 | $10^{-5}$ | 0 | 1 (7) | 1 |
| Kinect *paper* [9] | 32 | $10^{-5}$ | $10^{-7}$ | 1 (7) | 1 |
| Kinect *t-shirt* [9] | 32 | $10^{-5}$ | $10^{-7}$ | 1 (7) | 1 |

## References

1. Agudo, A., Moreno-Noguer, F.: Global model with local interpretation for dynamic shape reconstruction. In: Winter Conference on Applications of Computer Vision (WACV) (2017)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH. pp. 187–194 (1999)

3. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: Computer Vision and Pattern Recognition (CVPR) (2013)
4. Golyanik, V., Jonas, A., Stricker, D., Theobalt, C.: Intrinsic Dynamic Shape Prior for Fast, Sequential and Dense Non-Rigid Structure from Motion with Detection of Temporally-Disjoint Rigidity. arXiv e-prints (2019)
5. Golyanik, V., Mathur, A.S., Stricker, D.: Nrsfm-flow: Recovering non-rigid scene flow from monocular image sequences. In: British Machine Vision Conference (BMVC) (2016)
6. Kumar, S., Cherian, A., Dai, Y., Li, H.: Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In: Computer Vision and Pattern Recognition (CVPR). pp. 254–263 (2018)
7. Russell, C., Fayad, J., Agapito, L.: Energy based multiple model fitting for non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR). pp. 3009–3016 (2011)
8. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H., Pérez, P., Zollhöfer, M., Theobalt, C.: Fml: Face model learning from videos. In: Computer Vision and Pattern Recognition (CVPR) (2019)
9. Varol, A., Salzmann, M., Fua, P., Urtasun, R.: A constrained latent variable model. In: Computer Vision and Pattern Recognition (CVPR) (2012)