# Consolidating Segmentwise Non-Rigid Structure from Motion

Vladislav Golyanik[1,2] André Jonas[2] Didier Stricker[2,3]
[1]MPI for Informatics [2]University of Kaiserslautern [3]DFKI

## Abstract

*This paper introduces a new segmentwise technique which consolidates multiple principles for non-rigid structure from motion (NRSfM) into a single energy-based framework. The energy functional of our Consolidating Monocular Dynamic Reconstruction (CMDR) approach is optimised by non-linear least squares and includes terms allowing to define the deformation model and additional constraints simultaneously in the metric and trajectory spaces. The proposed method achieves high accuracy on several tested sequences while providing robustness and scalability due to the spatial scene segmentation and the new lifted spatial Laplacian term. CMDR is flexible and easy to implement, thanks to the unified optimisation framework. It allows for scenario-specific extensions and can be used for rapid prototyping of new NRSfM methods.*

## 1 Introduction

Non-rigid structure from motion (NRSfM) addresses 3D reconstruction of dynamic scenes from monocular image sequences relying on motion and deformation cues [1, 2]. Bregler *et al.* [1] showed that 2D point tracks are sufficient for monocular non-rigid reconstruction, and their pioneering work has entailed multiple successor methods for sparse reconstruction with various additional assumptions about the nature of motions and deformations [3, 4, 5, 6, 7]. In the last several years, the advent of accurate dense multi-frame optical trackers [8, 9] paved the way for dense NRSfM [10]. In dense NRSfM, points of interest often constitute connected regions. While many principles tested for sparse NRSfM can be directly generalised for the dense case, the transition from sparse to dense brings additional possibilities to constrain the solution space. Thus, several methods combine reconstruction and filtering of the recovered point clouds [2, 11]. Another example is the application of new in the context of NRSfM mathematical techniques (*e.g.,* tensor calculus as shown for the semi-dense case in [12]).

Along with that, many ideas were predominantly demonstrated in isolation. The reason is a large variety of frameworks and optimisation tools for dense NRSfM. Moreover, formulations are often interleaved with optimisation methods, are not transferable to other optimisation frameworks and even minor changes might result in a redesign of the optimisation techniques. The recent NRSfM challenge [13] has shown how diverse the performances of different methods can be in different scenarios. Having a new data set at hand, it is not always easy to predict how a method will perform on it. We believe that an important direction in NRSfM research is consolidating multiple ideas into a single framework which would enable a convenient integration and testing of different ideas. Though not gained much attention so far, it is a next logical step for boosting new ideas and discoveries in NRSfM.

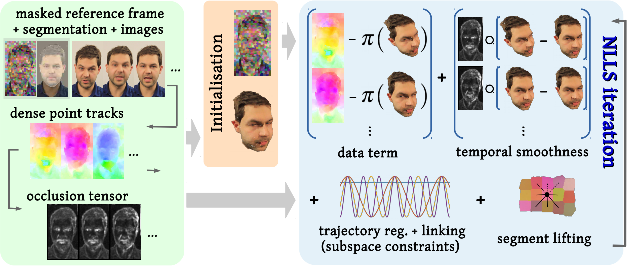

**Figure 1:** An overview of the proposed scalable segmentwise CMDR approach and the entire pipeline for non-rigid 3D reconstruction from monocular image sequences.

### 1.1 Contributions

Our main contribution is a new unifying energy-based *Consolidating Monocular Dynamic Reconstruction (CMDR)* approach. Our motivation for a single energy-based framework is to remove the burden of combinability of different principles for dense NRSfM, and we believe that this goal can be achieved by separating the problem formulation from the optimisation techniques. For optimisation, we employ non-linear least squares (NLLS) which brings an additional advantage of consistently handling linear and non-linear effects (see Sec. 3). We propose a new effective lifted spatial Laplacian term which assigns a weight to every pair of neighbouring segments. In systematic experiments (see Sec. 4), CMDR achieves high reconstruction accuracy with reduced runtime, due to the segmentwise processing, and competes with several recent approaches. An overview of CMDR is shown in Fig. 1.

## 2 Related Work

Several NRSfM approaches are based on energy minimisation [2, 11, 14, 15]. Among sparse settings, [14] employs locally smooth manifold learning. Agudo and

Moreno-Noguer formulate sparse reconstruction for orthographic camera model as an energy minimisation problem and solve it by linear least squares [15]. In CMDR, we optimise up to several millions of parameters simultaneously and solve dense reconstruction with NLLS. Among dense settings, Garg *et al.* [2] uses variational optimisation framework with energy splitting and alternating solving of two subproblems, and Golyanik *et al.* [11] optimise an energy with a frequency domain regulariser. All these approaches are highly diverse in the energy optimisation methods. In contrast, CMDR unifies multiple principles into a single energy-based framework.

The usefulness of patch-based processing was previously shown in NRSfM [16, 17]. In contrast to previous methods, we initialise segments either from the reference image by appliying a SLIC superpixel approach [18], or directly on the 3D solution initialisation with the segmentation method [19]. Several methods simultaneously formulate constraints in metric and trajectory spaces [6, 20] (probabilistic formulation). Similarly, we integrate two terms for expressing the deformation model and additional constraints in metric and trajectory spaces. The trajectory space smoothness term was rarely used in energy-based NRSfM so far. It allows integration of subspace constraints on point trajectories and originates from [5]. In CMDR, we impose smoothness of neighbouring trajectories by a total variation of trajectory coefficients. A similar regulariser was used in NRSfM [21] and multi-frame optical flow estimation [8, 9]. As a spatial segment regulariser, we employ a weighted lifted Laplacian which was previously applied in template-based non-rigid tracking from an RGB-D camera [22] and multi-frame RGB-D scene flow estimation [23].

Lately, *scalability* is increasingly gaining attention in NRSfM, referring to the method's support of a wide range of scenarios and types of data [24, 25]. In CMDR, scalability is enabled by different reconstruction granularity levels, thanks to the segmentation and piecewise affine modelling. Thus, it is possible to process data within time bounds by varying the segment size. Moreover, CMDR reconstructs different types of deformations (*e.g.,* faces, cloths, spinal deformation capture, surfaces of internal organs), with the number of points ranging from few thousands to as much as $2 \cdot 10^5$ (an example is the *barn owl* sequence in Sec. 4).

# 3 Consolidating Monocular Dynamic Reconstruction (CMDR)

In this Section, we first introduce notations and then describe the proposed CMDR framework for segmentwise reconstruction.

## 3.1 Assumptions and Notations

Given a *measurement matrix* $\mathbf{W} = [\mathbf{W}_1\mathbf{W}_2 \ldots \mathbf{W}_F]^\mathsf{T} \in \mathbb{R}^{2F\times P}$, *i.e.,* a set of $P$ 2D points tracks through $F$ views, the objective of NRSfM is the recovery of a time-varying geometry $\mathbf{S} = [\mathbf{S}_1\mathbf{S}_2 \ldots \mathbf{S}_F]^\mathsf{T} \in \mathbb{R}^{3F\times P}$ and camera poses $\mathbf{R} = [\mathbf{R}_1\mathbf{R}_2 \ldots \mathbf{R}_F]^\mathsf{T}$, where $f \in \{1, \ldots, F\}$ is a frame index. In CMDR, we cluster points into $L$ segments $\mathcal{S}_l$, $l \in \{1, \ldots, L\}$. Every segment represents a connected region with correspondences across all $\mathbf{S}_f$. $\mathbf{W}$ is registered to the origin of the coordinate system so that the translation is resolved and we have to estimate the per-frame camera rotation $\mathbf{R}_f$. For the orthographic camera, projection operator $\pi(\cdot)$ is $\mathbf{I}_{2\times3}$ identity matrix, though the derivations hold for both the calibrated perspective and orthographic cameras.

## 3.2 The Energy Functional with Segments

The proposed target segmentwise energy functional and can be interpreted for multiple granularity levels:

$$\begin{aligned}\mathbf{E}(\mathbf{R},\mathbf{T},\mathbf{A},\mathbf{w}) = \alpha\,\mathbf{E}_{\text{data}}(\mathbf{R},\mathbf{T}) + \beta\,\mathbf{E}_{\text{temp}}(\mathbf{T}) +\\ \gamma\,\mathbf{E}_{\text{linking}}(\mathbf{T},\mathbf{A}) + \rho\,\mathbf{E}_{\text{reg}}(\mathbf{A}) + \zeta\,\mathbf{E}_{\text{lifting}}(\mathbf{T},\mathbf{w}).\end{aligned} \tag{1}$$

All notations used in Eq. (1) will become clear by the end of this section, with all terms discussed in detail. See Fig. 1 for an overview of the entire pipeline. During initialisation, we compute an over-segmentation of the reference frame and cluster the trajectories. All points in a segment share common parameters.

The data term constrains segmentwise projections of the recovered shapes to agree with the 2D measurements:

$$\begin{aligned}&\mathbf{E}_{\text{data}}(\mathbf{R},\mathbf{T}) =\\ &\sum_{f=1}^{F} \left\| \mathbf{W}_f - \pi\Big( \mathbf{R}_f \left[ g(\mathbf{T}_1^f, \mathcal{S}_1) \;\ldots\; g(\mathbf{T}_L^f, \mathcal{S}_L) \right] \Big) \right\|_\epsilon^2,\end{aligned} \tag{2}$$

where $g(\mathbf{T}_l^f, \mathcal{S}_l)$ is an affine 7 DoF transformation $\mathbf{T}_l^f = \{\mathbf{R}_l^f, \mathbf{t}_l^f, s_l^f\}$ of segment $\mathcal{S}_l$, $l \in \{1, \ldots, L\}$ in frame $f$, with the segment pose $\mathbf{R}_l^f$, translation $\mathbf{t}_l^f$ and scale $s_l^f$. $\|\cdot\|_\epsilon$ denotes Huber loss defined as

$$\|\alpha\|_\epsilon = \begin{cases} \alpha, & \text{for } |\alpha| \leq \epsilon \\ 2\sqrt{\alpha} - 1, & \text{for } |\alpha| > \epsilon. \end{cases} \tag{3}$$

Robust norms such as $\ell^1$ or a Huber norm often lead to more accurate results in the presence of outliers.

The temporal smoothness term imposes similarity on reconstructions of adjacent frames. It is expressed in terms of differences of per-segment frame-to-frame transformations:

$$\mathbf{E}_{\text{temp}}(\mathbf{T}) = \sum_{f=2}^{F} \sum_{l=1}^{L} \left\| \mathbf{\Phi}_f^l \circ (\mathbf{T}_l^f - \mathbf{T}_l^{f-1}) \right\|_\epsilon^2, \tag{4}$$

with the per-frame and per-segment weights $\mathbf{\Phi}_f = \{\mathbf{\Phi}^1, \mathbf{\Phi}^2, \dots, \mathbf{\Phi}^L\}$ and $\circ$ denoting Hadamard product. The $\mathbf{\Phi}_f^l$ weights are set from the prior knowledge about the scene and segment transformations. They can be also influenced by an indicator of external and self-occlusions in the scene.

The linking term expresses assumptions about the deformation complexity of the scene. Here, we rely on known basis trajectories $\Theta$ sampled from discrete cosine transform (DCT) at regular intervals:

$$\mathbf{E}_{\text{linking}}(\mathbf{S}, \mathbf{A}) = \|\mathbf{\Psi} - (\Theta \otimes \mathbf{I}_3)_{3F \times 3K}\, \mathbf{A}_{3K \times L}\|_\epsilon^2, \quad (5)$$

$$\text{where } \mathbf{\Psi} = \begin{bmatrix} g(\mathbf{T}_1^1, \mathcal{S}_1) & g(\mathbf{T}_2^1, \mathcal{S}_2) & \dots & g(\mathbf{T}_L^1, \mathcal{S}_L) \\ g(\mathbf{T}_1^2, \mathcal{S}_1) & g(\mathbf{T}_2^2, \mathcal{S}_2) & \dots & g(\mathbf{T}_L^2, \mathcal{S}_L) \\ & & \vdots & \\ g(\mathbf{T}_1^F, \mathcal{S}_1) & g(\mathbf{T}_2^F, \mathcal{S}_2) & \dots & g(\mathbf{T}_L^F, \mathcal{S}_L) \end{bmatrix}, \quad (6)$$

$$\text{and } \Theta = \begin{pmatrix} \theta_{11} & \dots & \theta_{1K} \\ \vdots & \ddots & \vdots \\ \theta_{F1} & \dots & \theta_{FK} \end{pmatrix}, \text{ with}$$

$$\theta_{fk} = \frac{\sigma_k}{\sqrt{2}} \cos\left(\frac{\pi}{2F}(2f-1)(k-1)\right) \text{ and} \quad (7)$$

$$\sigma_k = \begin{cases} 1 & \text{if } k = 1, \\ \sqrt{2} & \text{otherwise.} \end{cases}$$

In Eq. (5), $\otimes$ denotes Kronecker product and $\mathbf{A}$ holds coefficients of linear combinations which approximate trajectories of recovered 3D points. The linking term connects or "links" the recovered trajectories to unknown though valid combinations of basis trajectories.

The regularisation term imposes a temporal coherence constraint in terms of discrepancies of 3D trajectories of neighbouring clusters. Since the recovered 3D trajectories are parameterised by $\mathbf{A}_k$, the regularisation term can be expressed as

$$\mathbf{E}_{\text{reg.}}(\mathbf{A}) = \sum_{l=1}^{L} \sum_{k=1}^{K} \mathbf{w}(l)\ \|\nabla \mathbf{A}_{k,l}\|_\epsilon^2, \quad (8)$$

where $\mathbf{w}(l)$ is a connectivity vector for every $\mathcal{S}_l$. To compute gradients of trajectory coefficients, Eq. (8) requires segment adjacencies. We compute an adjacency lookup table from the spatial arrangement of the segments.

Finally, we propose the lifted weighted Laplacian term accounting for the spatial segment coherency:

$$\mathbf{E}_{\text{lifting}}(\mathbf{T}, \mathbf{w}) = \sum_{f=1}^{F} \sum_{\forall w_{j,h}} \left(\zeta_1\ \left\|w_{j,h}^2(\mathbf{T}_j - \mathbf{T}_k)\right\|_2^2 + \zeta_2\ \left\|(1 - w_{j,h}^2)\right\|_2^2\right), \quad (9)$$

with $w_{j,h} = \pm 1$ if two segments move coherently and $w_{j,h} = 0$ if two segments belong to unrelated parts of a scene. If $|w_{j,h}|$ is small, the mutual influence between $\mathcal{S}_j$ and $\mathcal{S}_h$ is attenuated. Whether two segments move coherently or independently, is determined automatically by finding an optimal $\mathbf{w}$. The weights $\zeta_1$ and $\zeta_2$ — controlling the subterms of the lifting term — are fixed in all experiments. Solely the $\zeta$ weight is varied depending on assumptions about the scene. Our $\mathbf{E}_{\text{lifting}}$ is a robust spatial segment regulariser as $\mathbf{w}$ allows for an approximate detection of topological boundaries.

### 3.3 Initialisation and Energy Optimisation

We initialise $\mathbf{R}$ and $\mathbf{S}$ assuming predominant rigid component in the scene with [26]. The segments are obtained either with the superpixel approach [18] — if a reference frame is available — or with a graph-cut method run on the 3D initialisation [19]. For the reference frame — which can be any frame of the sequence — all $\mathbf{T}_l^f$ are initialised as $\mathbf{R} = \mathbf{I}$, $\mathbf{t} = \mathbf{0}$ und $s = 1$. Every segment obtained after rigid initialisation can be planarised. In the Huber loss, we use $\epsilon = 0.1$ and set $\alpha = \beta = \gamma = \rho = \zeta = 1.0$, $\zeta_1 = 0.2$, $\zeta_2 = 0.8$ in all experiments. The target energy functional (1) contains $M = FP + L(10F + 7)$ residual blocks, *i.e.,* $FP$ for the data term, $(F-1)L$ for temporal smoothness, $FL$ for linking, $8L$ for regularisation (assuming eight neighbours per $\mathcal{S}_l$) and $8FL$ for lifting. In total, there are $Q = 3F + 7LF + 3KL + 8L$ parameters, *i.e.,* $3F$ for global poses, $7LF$ for segment orientations for every frame, $3KL$ for the DCT coefficients and $8L$ for $\mathbf{w}$.

Let $\mathrm{x}$ be a set of unknowns in the target energy functional. Suppose the residuals are compactly denoted by $f_r(\mathrm{x})$, $r \in \{1, \dots, M\}$ and stacked into a multivariate vector-valued function $\mathbf{F}(\mathrm{x}) : \mathbb{R}^Q \to \mathbb{R}^M$: $\mathbf{F}(\mathrm{x}) = [f_1(\mathrm{x}), f_2(\mathrm{x}), \dots, f_r(\mathrm{x})]^\mathsf{T}$. The target non-linear energy optimisation problem can be compactly written in terms of $\mathrm{x}$ as $\mathrm{x}' = \arg\min_{\mathrm{x}} \|\mathbf{F}(\mathrm{x})\|_2^2$. In every optimisation step, the objective is linearised in the vicinity of the current solution $\mathrm{x}_t$ by the first-order Taylor expansion: $\mathbf{F}(\mathrm{x} + \Delta\mathrm{x}) \approx \mathbf{F}(\mathrm{x}) + \mathbf{J}(\mathrm{x})\Delta\mathrm{x}$, with $\mathbf{J}(\mathrm{x})_{M \times Q}$, *i.e.,* the Jacobian of $\mathbf{F}(\mathrm{x})$. The per-iteration convex and overconstrained objective for $\Delta\mathrm{x}$ reads $\min_{\Delta\mathrm{x}} \|\mathbf{J}(\mathrm{x})\Delta\mathrm{x} + \mathbf{F}(\mathrm{x})\|^2$. An optimum is achieved when the condition $\mathbf{J}(\mathrm{x})\Delta\mathrm{x} = -\mathbf{F}(\mathrm{x})$ holds, and, in practice, is computed in the least-squares sense with Levenberg-Marquardt (LM) algorithm.

## 4 Experiments

In this section, we describe the results. CMDR is implemented in C++ for a CPU. As NLLS solver, we use *ceres* [27]. We run the experiments on a system with 24 Gb RAM and a quadcore Intel Xeon v3520 processor achieving 2.67 GHz. If the size of a segment equals to one point, we refer to the *dense per-pixel case.*

### 4.1 Quantitative Evaluation

For the quantitative evaluation, we select two *synthetic face* sequences with a wide variation of expressions (99 frames, 28887 points per frame) [2], several sequences from White *et al.* [28], *i.e., coin* (a cloth laid upon a bowl with coins being dropped into the center of the cloth; 45 frames, 2146 points per frame; this sequence was used for evaluation in [20]), *toss* (a rag being tossed onto a cup, with a large deformation occurring when the bump is formed on the surface; 26 frames, 1370 points per frame), *flag mocap* (450 frames, 540 points per frame), *synthetic flag* (60 frames, 9622 points per frame) [8]. We project the 3D sequences of White *et al.* by a smoothly moving virtual camera to generate the ground truth measurements (an angular step of 5° per axis, with the maximum deviation of 20°). Next, we adapt the mocap sequence of Valgaerts *et al.* [29] so that the ground truth geometry, measurements, corresponding images and masks are available for every frame. For the modification, we rotate the ground truth surfaces and project them into an image plane by ray tracing. The projection of the reference frame defines the mask. The ground truth optical flow is obtained as the distances between the projections of the corresponding points in the image plane. The obtained sequences (*actor mocap* and *actor mocap #2*) contain 100 frames each, with $3.5 \cdot 10^4$ points per frame.

We compare several approaches qualitatively in the dense per-pixel case: Trajectory Basis (TB) [5], Metric Projections (MP) [7], Variational Approach (VA) [2], Coherent Depth Fields (CDF) [11], Dense Spatio-Temporal Approach (DSTA) [30], Scalable Monocular Surface Recovery (SMSR) [24], Grassmannian Manifold (GM) [25] and the proposed CMDR. We report mean per-sequence 3D error $e_{3D}$ defined as $e_{3D} = \frac{1}{F} \sum_{f=1}^{F} \frac{\|\mathbf{S}_f^{GT} - \mathbf{S}_f\|_{\mathcal{F}}}{\|\mathbf{S}_f^{GT}\|_{\mathcal{F}}}$, where $\mathbf{S}_f^{GT}$ denotes the $f$-th ground truth shape. $e_{3D}$ for TB, VA, MP and DSTA are replicated from [30], and $e_{3D}$ for CDF and GM are taken from the respective papers.

Tables 1 and 2 summarise $e_{3D}$ for the *synthetic face* data set and the sequences from White *et al.* [28]. On the *synthetic face*, our method achieves the third best accuracy after SMSR [24] and GM [25]. The difference in $e_{3D}$ is volatile and smaller compared to the difference with VA [2] ranked fourth. At the same time, CMDR outperforms SMSR [24] on the *coin*, *toss*, *flag mocap* and *synthetic flag* by a considerable margin. The Probabilistic Point Trajectory Approach [20] achieves a slightly lower $e_{3D} = 0.057$ on *coin*, but our tracks and camera poses differ from those used in [20].

Fig. 2 summarises the accuracy, runtime and number of LM solver iterations as the functions of the segment size for the *actor mocap* sequence. The accuracy variation of CMDR is low — the difference for the segments with 50 pixels compared to 200 pixels is below 12% while the runtime drops by the factor of six. While segment size increases, the decreasing number of segments results in a smaller number of variables and the total number of solver iterations until the convergence decreases up to the factor of two.

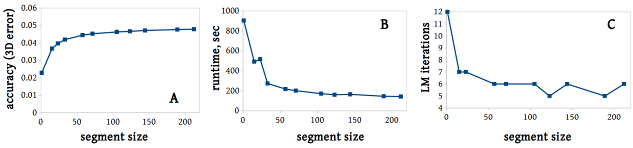

**Figure 2:** $e_{3D}$, runtime and number of LM solver iterations as the functions of the segment size for the *actor mocap*.

**Table 1:** Comparison of multiple methods on 99 frames long *synthetic faces* from [2], observed by two different camera settings. Our method achieves the third best $e_{3D}$ with a volatile difference to SMSR [24] and GM [25].

| *seq.* | TB<br>[5] | MP<br>[7] | VA<br>[2] | DSTA<br>[30] | CDF<br>[11] | SMSR<br>[24] | GM<br>[25] | CMDR<br>(ours) |
|---|---|---|---|---|---|---|---|---|
| Seq. 3 | 0.1252 | 0.0611 | 0.0346 | 0.0374 | 0.0886 | **0.0304** | **0.0294** | **0.0324** |
| Seq. 4 | 0.1348 | 0.0762 | 0.0379 | 0.0428 | 0.0905 | **0.0319** | **0.0309** | **0.0369** |

**Table 2:** Comparison of SMSR [24] and CMDR on the sequences from [28] and [8].

| *approach* | *coin* | *toss* | *flag mocap* | *synth. flag* | *actor mocap* | *actor mocap #2* |
|---|---|---|---|---|---|---|
| SMSR [24] | 0.2424 | 0.4003 | 0.196 | 0.1467 | 0.054 | **0.0145** |
| CMDR | **0.0696** | **0.3064** | **0.0792** | **0.084** | **0.0257** | 0.0228 |

**Table 3:** The summary of the ablation study.

| *seq.* | all terms | no $\mathbf{E}_{\text{reg.}}$ | no $\mathbf{E}_{\text{lift.}}$ | $\mathbf{E}_{\text{data}}$, $\mathbf{E}_{\text{temp.}}$ and $\mathbf{E}_{\text{lift.}}$ | $\mathbf{E}_{\text{data}}$, $\mathbf{E}_{\text{temp.}}$ |
|---|---|---|---|---|---|
| Seq. 3 | 0.0627 | 0.0616 | 0.0906 | 0.0622 | 0.0681 |
| Seq. 4 | 0.0678 | 0.0678 | 0.0967 | 0.0682 | 0.0736 |

**Ablation Study.** We perform ablation study by systematically switching off different combinations of energy terms in the experiment with the *synthetic face*. We obtain the segmentation with the Felzenszwalb's method [19] on 3D coordinates, with 40 points per segment (pps) on average. The results are summarised in Table 3. The ablation study demonstrates that all terms are useful and contribute to the accuracy of CMDR. For seq. 3, all terms except $\mathbf{E}_{\text{reg.}}$ result in the most accurate reconstruction, and the combination of the data, smoothness and lifting terms also works well. For seq. 4, all terms on lead to the most accurate recovery, with a slight decay if $\mathbf{E}_{\text{reg.}}$ is disabled. Otherwise, the $e_{3D}$ patterns of both sequences are similar. The mean CMDR's runtime amounts to eight minutes.

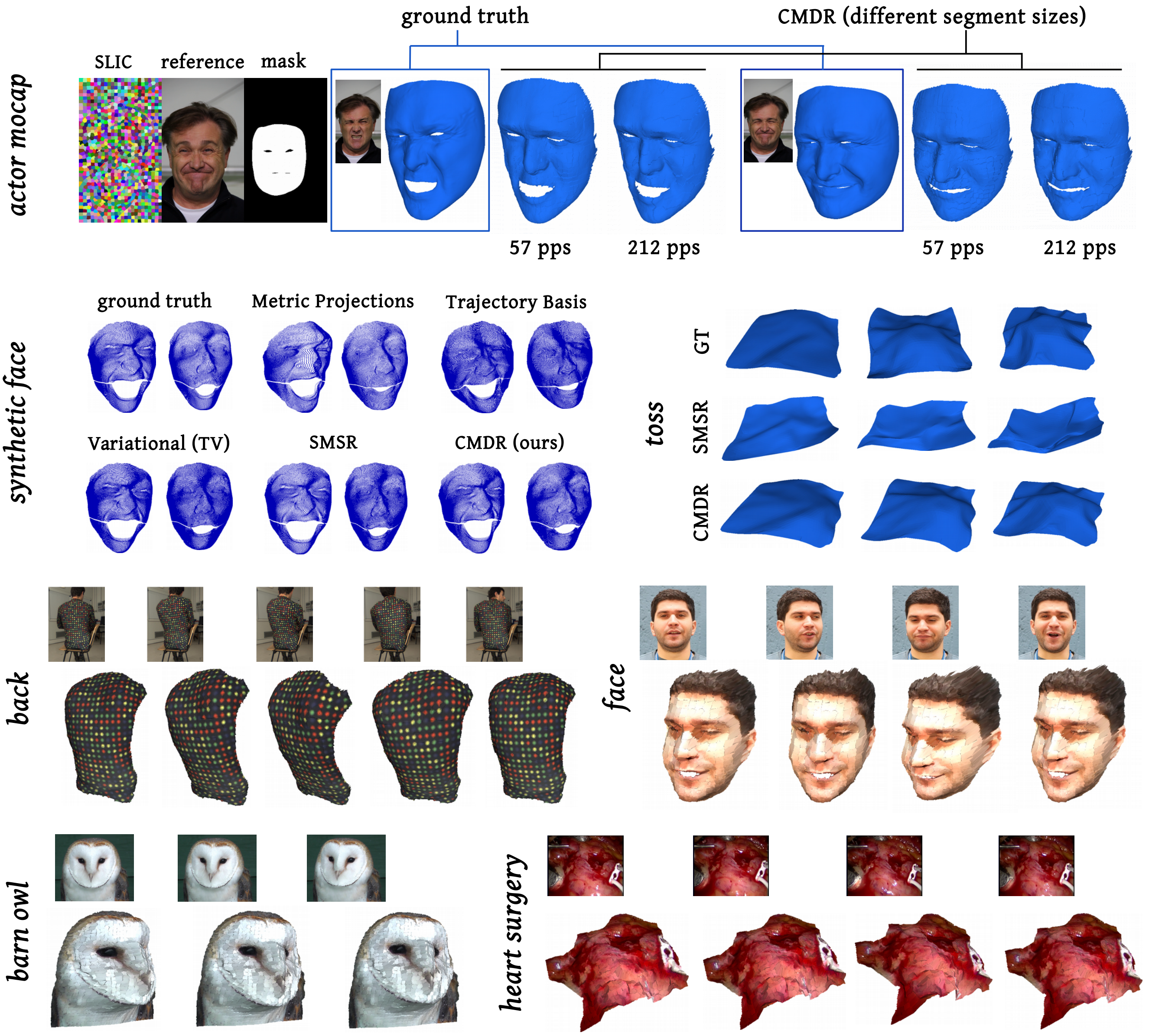

**Figure 3:** Visualisations of selected reconstruction results. For *actor mocap* [29], results of CMDR for two segment sizes are shown (57 points per segment (pps) and 212 pps). On *synthetic face*, no noticeable qualitative difference between VA [2], SMSR [24] and CMDR is observed. On *toss*, SMSR shows a lower accuracy than CMDR due to a less accurate initialisation, as one of the reasons. For *back* [31], *face* [2], *barn owl* [32] and *heart surgery* [33], CMDR outputs highly realistic results while reducing the runtime by the factor of six compared to the dense per point case. Note smooth spatial transitions between the segments. Orientations of the segments cause slightly different hues of neighbouring segments.

## 4.2 Experiments with Real Data

Fig. 3 visualises selected experimental results. On the *actor mocap*, we see qualitatively that CMDR is well-posed w.r.t. the segment size, *i.e.,* the reconstructions are similar and accurately resemble the ground truth irrespective of the segment size. The minor differences in $e_{3D}$ arise due to the varying segment granularity. For the *synthetic faces*, there is no noticeable qualitative difference for VA [2], SMSR [24] and CMDR. Compared to SMSR, CMDR captures the rag deformations in the *toss* sequence more accurately. Additionally, we show the efficiency of CMDR in different real-world scenarios with *face* [2], *back* [31], *barn owl* [32] and *heart surgery* [33] sequences. As only images are initially available, we compute dense point tracks by multiframe optical flow with subspace constraints and explicit handling of small occlusions [9]. CMDR captures realistic geometry in all scenes, see Fig. 3. We set the average segment size to 57 pixels. In *back*, the segments are visible only under a close look. In the remaining sequences, the surfaces are moderately inhomogeneous. Since the segments are planarised and due to differing surface orientations, variations in segment hues are observed. Those, on the one hand, do not substantially affect the perception, and, on the other hand, can be alleviated by surface smoothing (currently not applied), if necessary. The cardiac cycle is distinctly perceptible in the reconstructions of the *heart surgery*.

## 5 Conclusion

The proposed CMDR consolidates various principles for dense NRSfM into a unified energy-based framework. Along with the data term and trajectory regularisation as a deformation model, we propose weighted temporal smoothness, trajectory regularisation and the new segment lifting terms. The ablation study has shown that all terms contribute to the accuracy of the method. CMDR achieves high reconstruction accuracy on multiple synthetic and real data sets while bringing the advantage of scalability thanks to segments. One of the promising directions for future work is automatic handling of topological boundaries and we are currently working on adapting CMDR for endoscopic scenarios on parallel hardware.

### Acknowledgement

This work is supported by the BMBF project VIDETE (01IW18002).

## References

[1] Bregler et al.: Recovering non-rigid 3d shape from image streams. In: CVPR. (2000)

[2] Garg et al.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: CVPR. (2013)

[3] Brand, M.: A direct method for 3d factorization of nonrigid motion observed in 2d. In: CVPR. (2005)

[4] Torresani et al.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. TPAMI **30** (2008) 878–892

[5] Akhter et al.: Trajectory space: A dual representation for nonrigid structure from motion. TPAMI **33** (2011)

[6] Gotardo, P.F.U., Martinez, A.M.: Non-rigid structure from motion with complementary rank-3 spaces. In: CVPR. (2011)

[7] Paladini et al.: Optimal metric projections for deformable and articulated structure-from-motion. IJCV **96** (2012) 252–276

[8] Garg, R., Roussos, A., Agapito, L.: A variational approach to video registration with subspace constraints. IJCV **104** (2013)

[9] Taetz et al.: Occlusion-aware video registration for highly non-rigid objects. In: WACV. (2016)

[10] Russell et al.: Dense non-rigid structure from motion. In: 3DIMPVT. (2012)

[11] Golyanik et al.: Introduction to coherent depth fields for dense monocular surface recovery. In: BMVC. (2017)

[12] Parashar et al.: Isometric non-rigid shape-from-motion with riemannian geometry solved in linear time. TPAMI **PP** (2018)

[13] Jensen et al.: A benchmark and evaluation of non-rigid structure from motion. In: arXiv.org. (2018)

[14] Rabaud, V., Belongie, S.: Re-thinking non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR). (2008)

[15] Agudo, A., Moreno-Noguer, F.: Simultaneous pose and non-rigid shape with particle dynamics. In: CVPR. (2015)

[16] Taylor et al.: Non-rigid structure from locally-rigid motion. In: CVPR. (2010)

[17] Lee et al.: Consensus of non-rigid reconstructions. In: CVPR. (2016)

[18] Achanta et al.: Slic superpixels compared to state-of-the-art superpixel methods. TPAMI **34** (2012)

[19] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. IJCV **59** (2004) 167–181

[20] Agudo, A., Moreno-Noguer, F.: A scalable, efficient, and accurate solution to non-rigid structure from motion. CVIU (**167**) 121–133

[21] Olsen, S.I., Bartoli, A.: Implicit non-rigid structure-from-motion with priors. JMIV **31** (2008)

[22] Zollhöfer et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Trans. Graph. **33** (2014)

[23] Golyanik et al.: Multiframe scene flow with piecewise rigid motion. In: 3DV. (2017)

[24] Ansari et al.: Scalable dense monocular surface reconstruction. In: 3DV. (2017)

[25] Kumar et al.: Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In: CVPR. (2018)

[26] Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. IJCV **9** (1992) 137–154

[27] Agarwal et al.: Ceres solver. (`http://ceres-solver.org`)

[28] White et al.: Capturing and animating occluded cloth. ACM TOG **26** (2007)

[29] Valgaerts et al.: Lightweight binocular facial performance capture under uncontrolled lighting. ACM Trans. Graph. (TOG) (2012)

[30] Dai et al.: Dense non-rigid structure-from-motion made easy – a spatial-temporal smoothness based solution. In: ICIP. (2017)

[31] Russell et al.: Energy based multiple model fitting for non-rigid structure from motion. In: CVPR. (2011)

[32] Golyanik et al.: Nrsfm-flow: Recovering non-rigid scene flow from monocular image sequences. In: BMVC. (2016)

[33] Stoyanov, D.: Stereoscopic scene flow for robotic assisted minimally invasive surgery. In: MICCAI. (2012)