Technische Universität Kaiserslautern
DYNAMICS
augmented VISION

Accurate 3D Reconstruction of Dynamic Scenes from Monocular Image Sequences with Severe Occlusions

Vladislav Golyanik, Torben Fetzer and Didier Stricker
University of Kaiserslautern, German Research Center for Artificial Intelligence

WACV 2017
IEEE 2017 Winter conference on Applications of Computer Vision

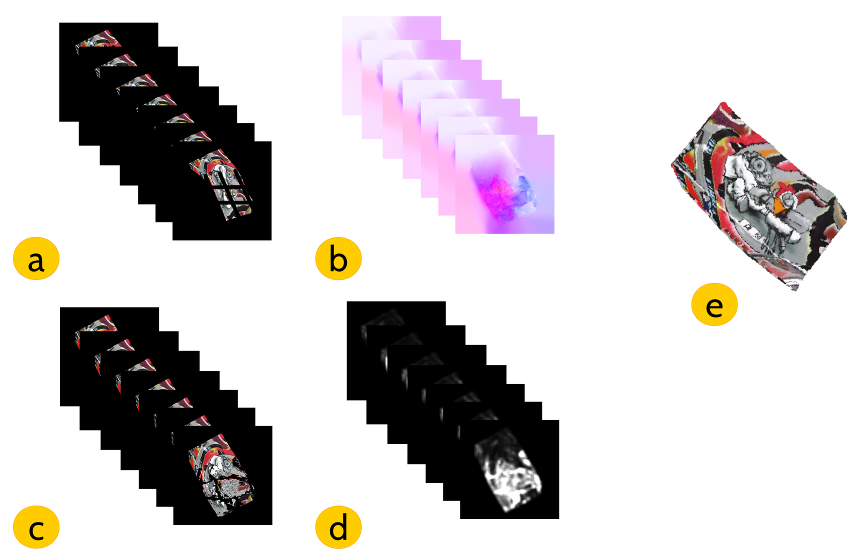
Introduction and motivation

The work tackles dense NRSfM in scenarios with large occlusions or inaccurate point tracks. A new *hybrid NRSfM* framework is proposed. The core method SPVA allows to regularize time-varying structure on the per-pixel level, given an **occlusion indicator** and a **shape prior.** Shape prior is estimated from several first non-occluded frames of the sequence under non-rigid deformations.

	NRSfM	*Template-based reconstruction*	*hybrid NRSfM*
Shape prior	does not require a template	requires a template (obtained in an external procedure, often under rigidity assumption)	generates a shape prior on the fly *under non-rigid deformations*

Application scenarios: minimally invasive surgery, reconstruction and tracking of long sequences under occlusions, specular effects, brightness inconsistency. The framework also proposes a method to obtain a template for template-based methods by relaxing the assumption of a known accurate reconstruction.

Overview of the framework



(a): the input to the pipeline is an image sequence of a non-rigidly deforming scene. (b) first stage of the pipeline is point tracking with multi-frame optical flow [2]. (c) the second stage is occlusion tensor estimation (shown in (d); brighter values indicate higher occlusion probability of the pixel). Next, a shape prior is estimated relying on the total intensity criterion. The correspondences, the occlusion tensor and the estimated shape prior are inputs for the **Shape Prior based Variational Approach (SPVA)**. (e) example of a shape prior.

References

[1] R. Garg, A. Roussos, and L. Agapito. *Dense variational reconstruction of non-rigid surfaces from monocular video*. In CVPR, 2013.
[2] R. Garg, A. Roussos, and L. Agapito. *A variational approach to video registration with subspace constraints*. IJCV, 2013.
[3] B. Taetz, G. Bleser, V. Golyanik, and D. Stricker. *Occlusion-aware video registration for highly non-rigid objects*. In WACV, 2016.
[4] Y. Dai, H. Li, and M. He. *A simple prior-free method for non-rigid structure-from-motion factorization*. IJCV, 2014.
[5] R. Yu, C. Russell, N. Campbell, and L. Agapito. *Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video*. In ICCV, 2015.
*[6] AMP = Accelerated Metric Projections. In WACV 2017.

Energy functional of SPVA

$$\underset{\mathbf{R},\mathbf{S}}{\operatorname{argmin}} \frac{\lambda}{2} \underbrace{\|\mathbf{W}-\mathbf{R}\mathbf{S}\|_{\mathcal{F}}^2}_{\text{data term}} + \frac{\gamma}{2} \underbrace{\|\Gamma(\mathbf{S}-\mathbf{S}_{\text{prior}})\|_{\mathcal{F}}^2}_{\text{shape prior term}} + \underbrace{\sum \|\nabla \mathbf{S}_f^i(p)\|}_{\text{total variation}} + \tau \underbrace{\|\mathrm{P}(\mathbf{S})\|_*}_{\text{nuclear norm}}$$

Shape prior granularity

per sequence	$\frac{\gamma}{2}\|\mathbf{S}-\mathbf{S}_{\text{prior}}\|_{\mathcal{F}}^2$	$\Gamma = \mathbf{I}$
per frame	$\frac{\gamma}{2}\|\Gamma(\mathbf{S}-\mathbf{S}_{\text{prior}})\|_{\mathcal{F}}^2$	Γ is diagonal
per pixel per frame	$\frac{\gamma}{2}\|\tilde{\Gamma}(\tilde{\mathbf{S}}-\tilde{\mathbf{S}}_{\text{prior}})\|_{\mathcal{F}}^2$	$\tilde{\Gamma} \in \mathbb{R}^{3FN \times 3FN}$

SPVA: Variational NRSfM with a Shape Prior

Input: measurements $\mathbf{W}$, $\mathbf{S}_{prior}$, parameters $\lambda, \gamma, \tau, \theta, \eta = \theta\tau$
Output: non-rigid shape $\mathbf{S}$, camera poses $\mathbf{R}$
1: **Initialisation:** $\mathbf{S}$ and $\mathbf{R}$ under rigidity assumption [46]
2: **STEP 1. Fix S, find an optimal R** ***framewise*:**
3: $\text{svd}(\mathbf{WS}(\mathbf{SS}^\mathsf{T})^{-1}) = \mathbf{U\Sigma V}^\mathsf{T}$
4: $\mathbf{R} = \mathbf{UCV}^\mathsf{T}$, where
$\mathbf{C} = \text{diag}(1, 1, \ldots, 1, \text{sign}(\det(\mathbf{UV}^\mathsf{T})))$
5: **STEP 2. Fix R; find an optimal S:**
6: **while** not converge **do**
7: **Primal-Dual:** fix $\bar{\mathbf{S}}$; *find an intermediate* $\mathbf{S}$ *(Eq. (9))*
8: **Initialisation:** $q_f^i(p) = 0$
9: **while** not converge **do**
10: $\mathbf{D}_q = \begin{pmatrix} \nabla^* q_1^1(1) & \cdots & \nabla^* q_1^1(N) \\ \vdots & \ddots & \vdots \\ \nabla^* q_F^3(1) & \cdots & \nabla^* q_F^3(N) \end{pmatrix}$
11: $\mathbf{S} = (\lambda \mathbf{R}^\mathsf{T}\mathbf{R} + \gamma + \frac{1}{\theta}\mathbf{I})^{-1}$
12: $(\lambda \mathbf{R}^\mathsf{T}\mathbf{W} + \frac{1}{\theta}\bar{\mathbf{S}} + \gamma \mathbf{S}_{\text{prior}} - \mathbf{D}_q)$
13: **for** $f = 1, \ldots, F$; $i = 1, \ldots, 3$; $p = 1, \ldots, N$ **do**
14: $q_f^i(p) = \frac{q_f^i(p) + \sigma\nabla \mathbf{S}_f^i(p)}{\max(1, \|q_f^i(p) + \sigma\nabla \mathbf{S}_f^i(p)\|)}$
15: **end while**
16: **Soft-Impute:** fix $\mathbf{S}$; *find an intermediate* $\bar{\mathbf{S}}$ *(Eq. (10))*
17: $\text{svd}(\mathrm{P}(\mathbf{S})) = \mathbf{UDV}^\mathsf{T}$, where $\mathbf{D} = \text{diag}(\sigma_1, \ldots, \sigma_r)$
18: $\bar{\mathbf{S}} = \mathbf{UD}_\eta\mathbf{V}^\mathsf{T}$, where
$\mathbf{D}_\eta = \text{diag}(\max(\sigma_1 - \eta, 0), \ldots, \max(\sigma_r - \eta, 0))$
19: **end while**

Tomasi&Kanade

projection of an affine update onto SO(3)

$q_f^i(p)$ are dual variables
$\nabla^* = -\text{div}(\cdot)$

per sequence shape prior

Soft-Impute

S is fixed
update S
update $\mathbf{D}_q$
$\bar{\mathbf{S}}$ is fixed
$\bar{\mathbf{S}}$ is fixed
R is fixed

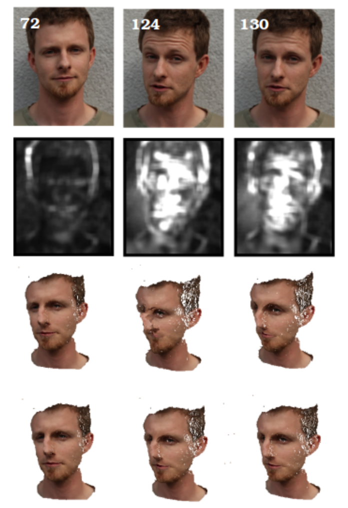
l. 11-12 change according to:

per frame

$$\tilde{\mathbf{S}} = (\lambda \mathbf{R}^\mathsf{T}\mathbf{R} + \gamma\Gamma^\mathsf{T}\Gamma + \frac{1}{\theta}\mathbf{I})^{-1}(\lambda \mathbf{R}^\mathsf{T}\mathbf{W} + \frac{1}{\theta}\bar{\mathbf{S}} + \gamma\Gamma^\mathsf{T}\Gamma\mathbf{S}_{\text{prior}} - \mathbf{D}_q)$$

per pixel per frame

$$\tilde{\mathbf{S}} = (\underbrace{\lambda \hat{\mathbf{R}}^\mathsf{T}\hat{\mathbf{R}}}_{\text{block-diagonal}} + \underbrace{\frac{1}{\theta}\mathbf{I}_{3FN}}_{\text{diagonal}} + \underbrace{\gamma\tilde{\Gamma}^\mathsf{T}\tilde{\Gamma}}_{\text{diagonal}})^{-1}(\lambda \hat{\mathbf{R}}^\mathsf{T}\hat{\mathbf{W}} + \frac{1}{\theta}\tilde{\bar{\mathbf{S}}} + \gamma\tilde{\Gamma}^\mathsf{T}\tilde{\Gamma}\tilde{\mathbf{S}}_{\text{prior}} - \tilde{\mathbf{D}}_q)$$



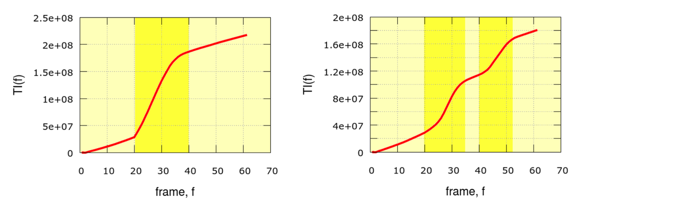
Obtaining a shape prior

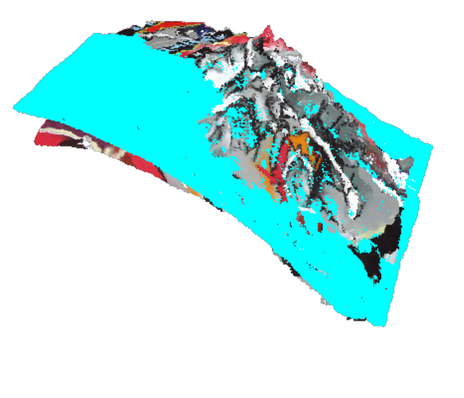
Total intensity (TI) criterion:

$$\sum_{f=1}^{F_{sp}} \left\| \int_\Omega du\, dv \right\|_2 \leq \epsilon$$

Alternatively:

$$\frac{TI(F_{sp}+1) - TI(F_{sp}-1)}{2} \leq \epsilon'$$





initialisation obtained under rigidity assumption overlayed with a shape prior (cyan)

Estimation of an occlusion tensor $\mathbf{E}(x)$

Input: dense flow fields $\mathbf{u}(x;\mathbf{n})$, a reference frame $\mathbf{I}(x, \mathbf{r})$, Gaussian kernel $G_{k\times k}$
Output: occlusion maps $\mathbf{E}(x, \mathbf{n})$
for every frame n $\in \{2, \ldots, F\}$ **do**
$\mathbf{w}(\mathbf{n}, \mathbf{r}) = \mathbf{I}(x, \mathbf{n}) - \mathbf{u}(x; \mathbf{n})$ (backprojection to $\mathbf{I}(x, \mathbf{r})$)
image difference $\mathbf{B}(x) = \mathbf{w}(\mathbf{n}, \mathbf{r}) - \mathbf{I}(x, \mathbf{r}) \triangleq$
for every pixel x **do**
$\mathbf{B}(x) = \|(x_r^{\mathbf{w}} - x_r^{\mathbf{I}})^2 + (x_g^{\mathbf{w}} - x_g^{\mathbf{I}})^2 + (x_b^{\mathbf{w}} - x_b^{\mathbf{I}})^2\|_2$
end for
$\mathbf{E}(x, \mathbf{n}) = \mathbf{B}(x) * G$
postprocess $\mathbf{E}(x)$
end for

$$e_{3D} = \frac{1}{F}\sum_{f=1}^{F} \frac{\|\mathbf{S}_f^{ref} - \mathbf{S}_f\|_{\mathcal{F}}}{\|\mathbf{S}_f^{ref}\|_{\mathcal{F}}}$$

Parallel energy optimisation

$\mathbf{D}_q$ and $\mathbf{S}$ updates as well as multiplications of large matrices are implemented on GPU

```
__global__ void kernel_compute_D_q();
__global__ void kernel_AB_T();
__global__ void kernel_AA_T();
...
```

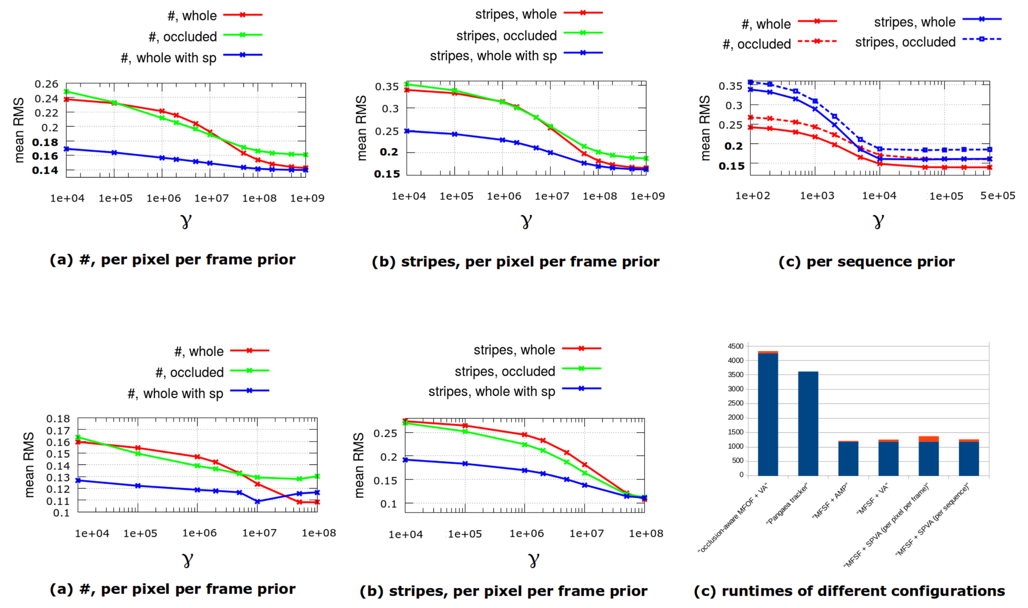
configuration	heart surgery 360×288, 50 fr.	face (new) 241×285, 136 fr.	ASL F5_10_A_H17 720×480, 114 fr.
MFSF [2] + VA [1]	$481.0 + 119.3$	$728.9 + 35.7$	$3114.0 + 400.0$
MFSF [2] + AMP [6]	$481.0 + 20.4$	$728.9 + 26.4$	$3114.0 + 98.0$
occlusion-aware MFOF [3] + VA [1]	$1592.8 + 119.2$	$2693.6 + 35.7$	$11995.3 + 300.5$
MFSF [2] + SPVA	$481.0 + 846.2$	$728.9 + 122.9$	$3114.0 + 1011.0$

Test platform:
- Xeon E5-1650
- GK110 GPU
- 32 GB RAM

Joint evaluation methodology:
- based on a dataset with a ground truth surface geometry and rendered images with occlusions (we choose the mocap flag sequence [2] and introduce large occlusions)

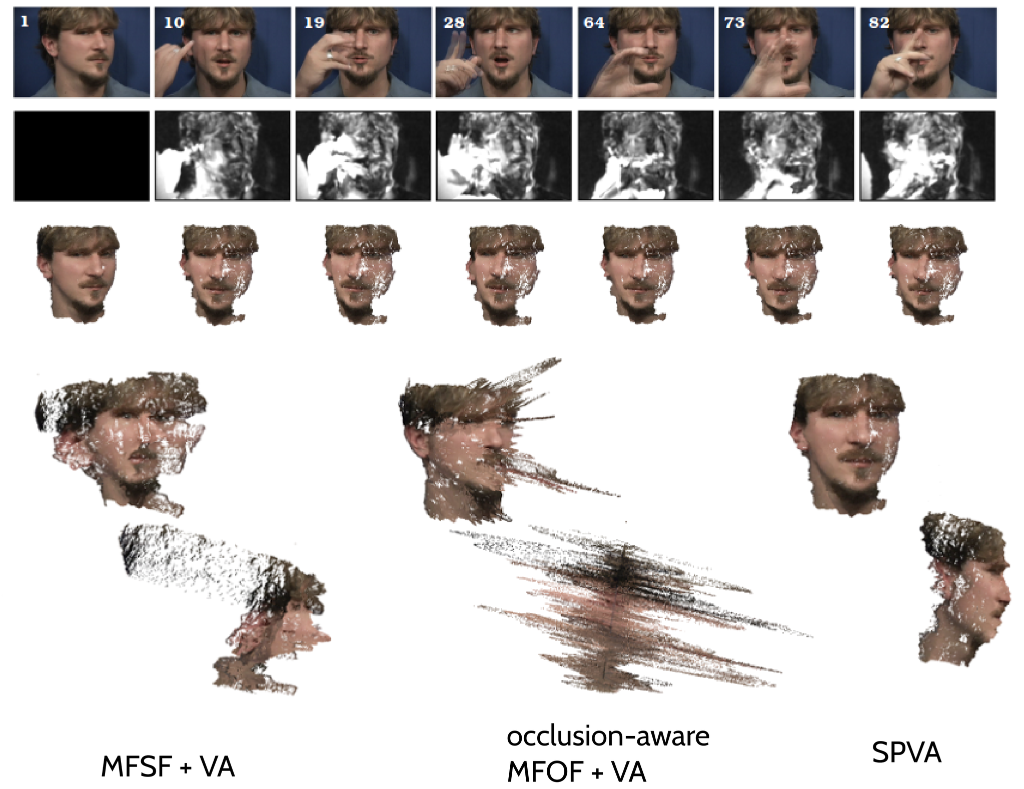
- two patterns are used: # and stripes
- correspondences are computed either with multi-frame subspace flow [2] or occlusion-aware video registration (MFOF) [3]
- accuracy and runtime of different algorithmic pipelines are evaluated

Model evaluation

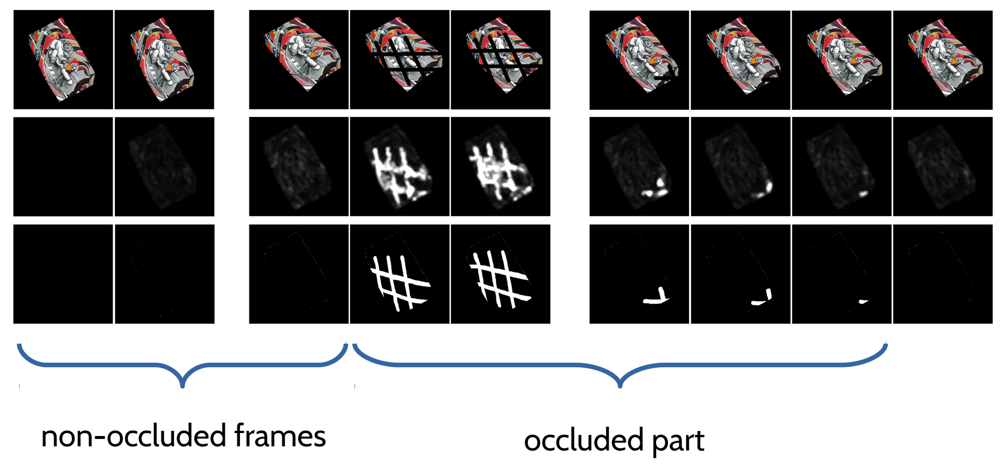


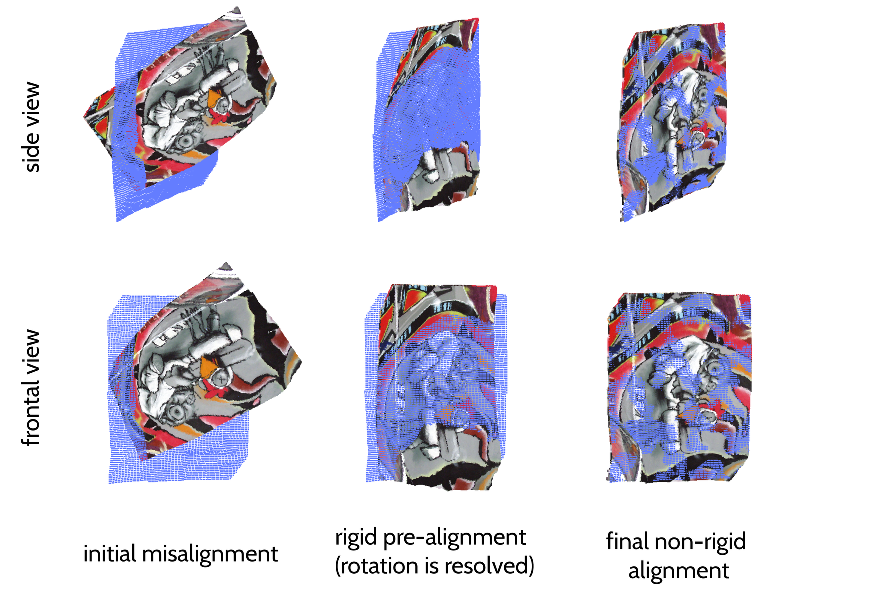
algorithm combination	mean RMS, #	mean RMS, *stripes*
o.a. MFOF [3]+VA [1]	0.181 (0.219)	0.195 (0.209)
Pangaea tracker [5]	**0.172 (0.191)**	**0.172 (0.191)**
MFSF [2]+AMP [6]	0.297 (0.381)	0.460 (0.523)
MFSF [2]+VA [1]	0.239 (0.252)	0.341 (0.355)
MFSF [2]+SPVA, p. pix.	**0.143 (0.161)**	**0.167 (0.189)**
MFSF [2]+SPVA, p. seq.	**0.140 (0.160)**	**0.160 (0.184)**

frame 22, regulariser strength increases from the left to the right

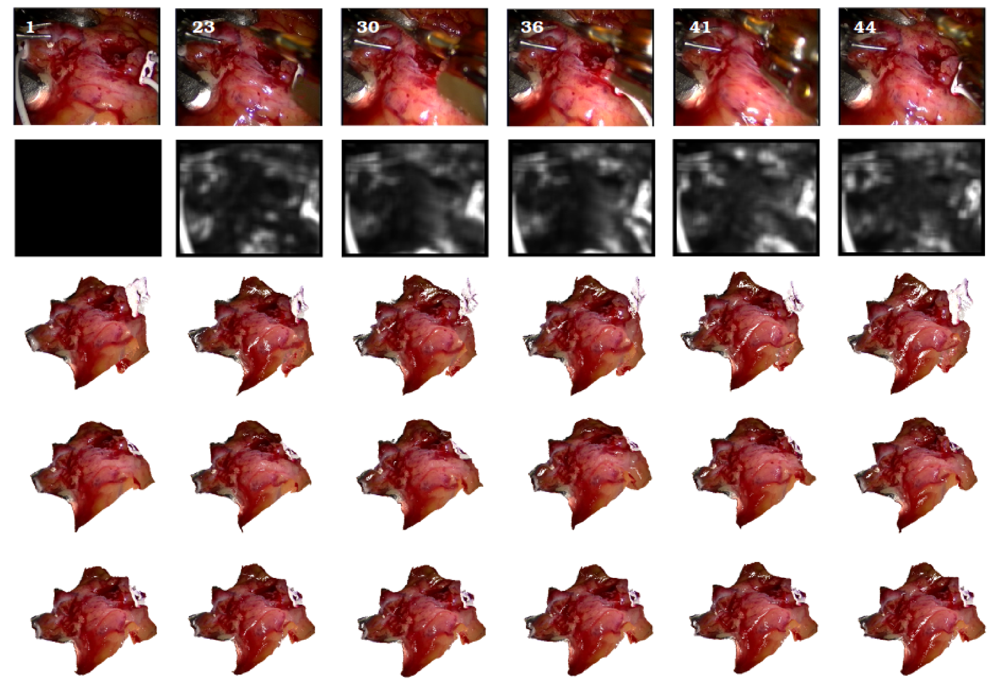


input #-sequence

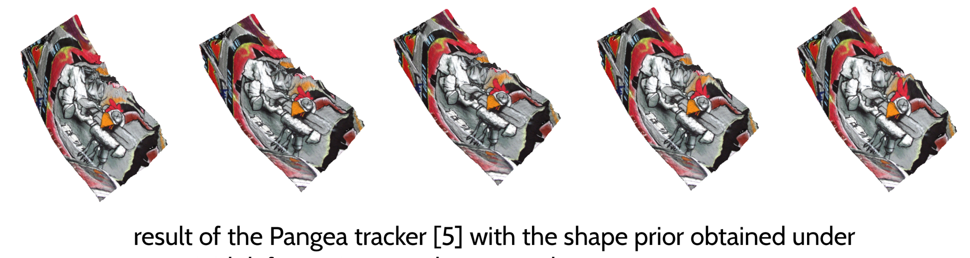




non-rigid alignment of the ground truth geometry with an exemplary reconstruction for correspondence establishment. Here, we use Extended CPD. The point correspondences are eventually used in the quantitative evaluation (3D error metric).



results on the *heart surgery* sequence (different algorithmic pipelines)



result of the Pangea tracker [5] with the shape prior obtained under non-rigid deformations used as a template