Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints

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
Non-Rigid Structure from Motion

input monocular video
Non-Rigid Structure from Motion

input monocular video

corresponding 3D reconstructions
Non-Rigid Structure from Motion

input monocular video

corresponding 3D reconstructions
Non-Rigid Structure from Motion
Non-Rigid Structure from Motion
Non-Rigid Structure from Motion
Non-Rigid Structure from Motion
Non-Rigid Structure from Motion

input frame
Non-Rigid Structure from Motion

input frame

3D reconstructions by different NRSfM methods
Which 3D Reconstruction is Accurate?

input frame

3D reconstructions by different NRSfM methods
Non-Rigid Structure from Motion


line
flyby
zigzag
Non-Rigid Structure from Motion

semi-circle  zigzag  tricky
Non-Rigid Structure from Motion

circle  flyby
Non-Rigid Structure from Motion

line

circle

semi-circle
Related Work

\[
\begin{pmatrix}
W_1 \\
\vdots \\
W_f
\end{pmatrix}
= \begin{pmatrix}
R_1 \\
\vdots \\
R_f
\end{pmatrix}
\begin{pmatrix}
X_1 \\
\vdots \\
X_f
\end{pmatrix}
\]

Measurement matrix is known  Camera matrix is unknown  3D shape matrix is unknown
2fn entries < 6f variables + 3fn variables

Dense NRSfM

\[ \min_{S,R} \frac{\lambda}{2} \| W - RS \|^2_F + \sum_{f,i,p} \| \nabla S^f_i(p) \| + \tau \| P(S) \|_* \]

Dense NRSfM

C3DPO

C3DPO

C3DPO

Deep NRSfM

Related Work

M. Sahasrabudhe and Z. Shu et al. ICCV Workshops, 2019.


Our Neural Dense NRSfM

• N-NRSfM does not require 3D supervision and does not rely on annotated 2D data collections
• **Core contribution**: a new neural deformation model component
• Shapes are recovered during the network training through backpropagation
Our Neural Dense NRSfM

• N-NRSfM does not require 3D supervision and does not rely on annotated 2D data collections
• Core contribution: a new neural deformation model component
• Shapes are recovered during the network training through backpropagation
Method Overview

\[ W \rightarrow R \rightarrow S \]

- refer. frame
- axis-angle representation
- 3D shapes in a canonical frame
Method Overview
Method Overview

\[ W \times R = \text{axis-angle representation} + \text{mean shape} S \]

3D shapes in a canonical frame
Method Overview

\[ W \] = \[ R \] + \[ S \]

\[ \text{mean shape } \bar{S} \]

\[ f_\theta \]

\[ \text{deformation auto-decoder} \]

\[ \text{3D shapes in a canonical frame} \]
Method Overview

\[ \text{W} \rightarrow \text{R} \rightarrow \text{mean shape } \overline{S} \rightarrow \text{structured latent space } z = \{ z_t \} \rightarrow \text{deformation auto-decoder } f_\theta \rightarrow \text{3D shapes in a canonical frame } \text{S} \]
Method Overview

\[ W R S \]

\[ \text{structured latent space } z = \{ z_t \} \]

\[ f_\theta \]

\[ \text{3D shapes in a canonical frame} \]
Our Differentiable Energy Function

\[ E = E_{\text{data}}(\theta, z, R) + \beta E_{\text{temp}}(\theta, z) + \gamma E_{\text{spat}}(\theta, z) + \eta E_{\text{traj}}(\theta, z) + \omega E_{\text{latent}}(z) \]

- \( \theta \) denotes learned network parameters
- \( z \) is the latent space function
- \( f_{\theta}(z) \) are displacements from the mean shape
- \( R \) is a block-diagonal matrix of truncated camera rotations
- \( \beta, \gamma, \eta, \omega \) are the weights balancing the terms
Our Differentiable Energy Function

\[ E = E_{\text{data}}(\theta, z, R) + \beta E_{\text{temp}}(\theta, z) + \gamma E_{\text{spat}}(\theta, z) + \eta E_{\text{traj}}(\theta, z) + \omega E_{\text{latent}}(z) \]

\( \theta \) denotes learned network parameters
\( z \) is the latent space function
\( f_\theta(z) \) are displacements from the mean shape
\( R \) is a block-diagonal matrix of truncated camera rotations
\( \beta, \gamma, \eta, \omega \) are the weights balancing the terms

\[ S_t = \bar{S} + f_\theta(z_t) \]
Our Differentiable Energy Function

\[ E = E_{\text{data}}(\theta, z, R) + \beta E_{\text{temp}}(\theta, z) + \gamma E_{\text{sphat}}(\theta, z) + \eta E_{\text{traj}}(\theta, z) + \omega E_{\text{latent}}(z) \]

\[ \theta \] denotes learned network parameters
\[ z \] is the latent space function
\[ f_\theta(z) \] are displacements from the mean shape
\[ R \] is a block-diagonal matrix of truncated camera rotations
\[ \beta, \gamma, \eta, \omega \] are the weights balancing the terms

\[ S_t = \bar{S} + f_\theta(z_t) \]

\[ \frac{\partial E}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial E}{\partial S_t} \frac{\partial S_t}{\partial \theta} \]
Our Differentiable Energy Function

\[ E = E_{\text{data}}(\theta, z, R) + \beta E_{\text{temp}}(\theta, z) + \gamma E_{\text{spat}}(\theta, z) + \eta E_{\text{traj}}(\theta, z) + \omega E_{\text{latent}}(z) \]
Our Differentiable Energy Function

\[ E = E_{\text{data}}(\theta, z, R) + \beta E_{\text{temp}}(\theta, z) + \gamma E_{\text{spat}}(\theta, z) + \eta E_{\text{traj}}(\theta, z) + \omega E_{\text{latent}}(z) \]
Temporal Regularisation

\[ E_{\text{temp}}(\theta, z) = \sum_{t=1}^{T-1} \| f_\theta(z_{t+1}) - f_\theta(z_t) \|_\epsilon \]

\[ E_{\text{tra}j}(\theta, z) = \| (1_T \otimes \tilde{S}) + f_\theta(z) - (\Phi \otimes I_3)A \|_\epsilon, \quad \Phi = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,K} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \cdots & \phi_{T,K} \end{pmatrix} \]
Temporal Regularisation

\[ E_{\text{temp}}(\theta, z) = \sum_{t=1}^{T-1} \| f_\theta(z_{t+1}) - f_\theta(z_t) \|_\epsilon \]

\[ E_{\text{tra}}(\theta, z) = \| (I_T \otimes \bar{S}) + f_\theta(z) - (\Phi \otimes I_3) A \|_\epsilon, \quad \Phi = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,K} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \cdots & \phi_{T,K} \end{pmatrix} \]

\[ \begin{array}{ccc}
S & \Theta & A \\
3F \times P & 3F \times 3K & 3K \times P \\
\end{array} \]

\[ X\text{-coordinate of trajectory of hand} = a_1 \times + a_2 \times + \ldots + a_k \times \]

Spatial Regulariser

\[ E_{\text{s}}(\theta, z) = \sum_{t=0}^{T-1} \sum_{p \in S_t} \left\| p - \frac{1}{|\mathcal{N}(p)|} \sum_{q \in \mathcal{N}(p)} q \right\|_1 - \lambda \sum_{t=1}^{T} \left\| \mathcal{P}_z(G_tS_t) \right\|_2 \]

- Laplacian smoothing
- Depth control

\[ E_{\text{s}} \]

no  \[ E_{\text{s}} \]

\[ E_{\text{s}} \]
Latent Space Constraints

\[ E_{\text{latent}}(z) = \| \mathcal{F}(z) \|_1 \quad \mathcal{F}(\cdot) \text{ denotes the Fourier transform operator} \]
Structure from Recurrent Motion

X. Li et al., CVPR, 2018.
Dynamic Shape Prior in NRSfM

V. Golyanik et al., arXiv.org, 2019.
Dynamic Shape Prior in NRSfM

V. Golyanik et al., arXiv.org, 2019.
Dynamic Shape Prior in NRSfM

V. Golyanik et al., arXiv.org, 2019.
$E_{\text{latent}}(z) = \| \mathcal{F}(z) \|_1$

$\mathcal{F}(\cdot)$ denotes the Fourier transform operator
Latent Space Constraints

\[ E_{\text{latent}}(\mathbf{z}) = \| \mathcal{F}(\mathbf{z}) \|_1 \]

\( \mathcal{F}(\cdot) \) denotes the Fourier transform operator.
Experimental Results


## Synthetic Face Sequences

<table>
<thead>
<tr>
<th></th>
<th>TB 7</th>
<th>MP 43</th>
<th>VA 19</th>
<th>DSTA 15</th>
<th>CDF 23</th>
<th>CMDR 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>traj. A</td>
<td>0.1252</td>
<td>0.0611</td>
<td>0.0346</td>
<td>0.0374</td>
<td>0.0886</td>
<td>0.0324</td>
</tr>
<tr>
<td>traj. B</td>
<td>0.1348</td>
<td>0.0762</td>
<td>0.0379</td>
<td>0.0428</td>
<td>0.0905</td>
<td>0.0369</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GM* 37</th>
<th>JM* 36</th>
<th>SMSR 8</th>
<th>PPTA 6</th>
<th>EM-FEM 1</th>
<th>N-NRSfM (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>traj. A</td>
<td>0.0294</td>
<td>0.0280</td>
<td>0.0304</td>
<td>0.0309</td>
<td>0.0389</td>
<td>0.045 / 0.032b</td>
</tr>
<tr>
<td>traj. B</td>
<td>0.0309</td>
<td>0.0327</td>
<td>0.0319</td>
<td>0.0572</td>
<td>0.0304</td>
<td>0.049 / 0.0389a</td>
</tr>
</tbody>
</table>

---

Synthetic Face Sequences

synthetic face (trajectory A)

ground truth  our reconstructions
Synthetic Face Sequences

*synthetic face (trajectory A)*

ground truth  
our reconstructions
Synthetic Face Sequences

synthetic face (trajectory B)

ground truth  our reconstructions
Expressions Sequence

Kinect Paper and T-Shirt

<table>
<thead>
<tr>
<th></th>
<th>TB 7</th>
<th>MP 43</th>
<th>DSTA 15</th>
<th>GM 37</th>
<th>JM 36</th>
<th>N-NRSfM (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper</td>
<td>0.0918</td>
<td>0.0827</td>
<td>0.0612</td>
<td>0.0394</td>
<td>0.0338</td>
<td>0.0332</td>
</tr>
<tr>
<td>t-shirt</td>
<td>0.0712</td>
<td>0.0741</td>
<td>0.0636</td>
<td>0.0362</td>
<td>0.0386</td>
<td>0.0309</td>
</tr>
</tbody>
</table>

Kinect Paper and T-Shirt

images  flow fields  3D reconstructions  latent space
Kinect Paper and T-Shirt

images | flow fields | 3D reconstructions | latent space
Kinect Paper and T-Shirt

images  flow fields  3D reconstructions  latent space
Kinect Paper and T-Shirt

images  flow fields  3D reconstructions  latent space
Kinect Paper and T-Shirt

images  flow fields  3D reconstructions  latent space
Kinect Paper and T-Shirt

images  flow fields  3D reconstructions  latent space
Comparison to FML

input images  GT flow  flow computed on images
Comparison to FML

ground truth meshes
FMLO. Tewari et al., CVPR’19
our reconstructions

A. Tewari et al., CVPR, 2019.
V. Golyanik et al., arXiv.org, 2019.
Comparison to FML

ground truth meshes

FML
Tewari et al., CVPR’19

our reconstructions

A. Tewari et al., CVPR, 2019.
V. Golyanik et al., arXiv.org, 2019.
Comparison to FML

ground truth meshes

FML
Tewari et al., CVPR’19

our reconstructions

A. Tewari et al., CVPR, 2019.
V. Golyanik et al., arXiv.org, 2019.
Comparison to FML

ground truth meshes
FML
Tewari et al., CVPR'19
our reconstructions

A. Tewari et al., CVPR, 2019.
V. Golyanik et al., arXiv.org, 2019.
Experimental Results

Experimental Results

Experimental Results

Sequence source: V. Golyanik et al., BMVC, 2016.
Experimental Results

images  flow fields  3D reconstructions  latent space

Sequence source: V. Golyanik et al., BMVC, 2016.
Experimental Results

input images  no $E_{\text{spat}}$  with $E_{\text{spat}}$

Experimental Results

input images  no $E_{spat}$  with $E_{spat}$

Experimental Results

images  flow fields  3D reconstructions  latent space

Experimental Results

images  flow fields  3D reconstructions  latent space

Experimental Results

Experimental Results

Experimental Results

Experimental Results


*shown with the increased frame rate*
Applications

ground truth

input
(noisy point cloud)

completed shape
Applications

- ground truth
- input (partial point cloud)
- completed shape
Applications

ground truth
(input (partial point cloud))
completed shape
Applications
Applications

direct monocular non-rigid 3D reconstruction
Applications

direct monocular non-rigid 3D reconstruction
Project Page

http://gvv.mpi-inf.mpg.de/projects/Neural_NRSfM/
Project Page

http://gvv.mpi-inf.mpg.de/projects/Neural_NRSfM/

source code available for research purposes
Thank You

Vikram
Edgar
Antonio
Christian
Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints

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