Shape Capture: Performance Capture

Computer Vision for Computer Graphics Seminar
Summer 2013
Max Planck–Institut Informatik

Yeara Kozlov
Supervisor: Christian Theobalt

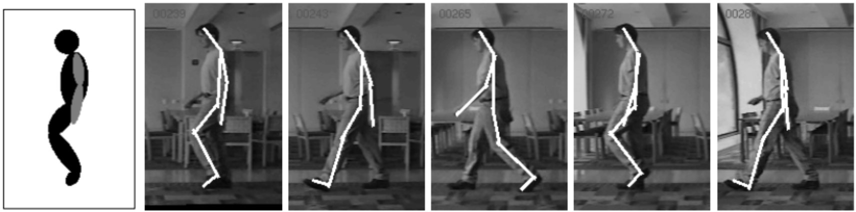
June 18th, 2013

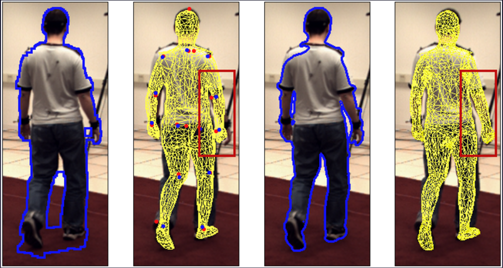
Dense, Marker–less Performance Capture

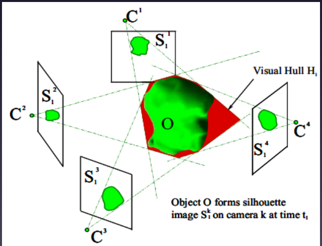
Two approaches for **dense**, **marker-less** performance capture:

- Volumetric deformation of mesh: **Performance Capture from Sparse Multi-View Video**
  - De Aguiar, Edilson, Stoll, Carsten, Theobalt, Christian, Ahmed, Naveed, Seidel, Hans-Peter and Thrun, Sebastian
  - ACM Transactions on Graphics, 2008
- Combined skeleton and surface presentation: **Motion Capture Using Joint Skeleton Tracking and Surface Estimation**
  - Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel
  - CVPR 2009.

Previously on Computer Vision for Computer Graphics



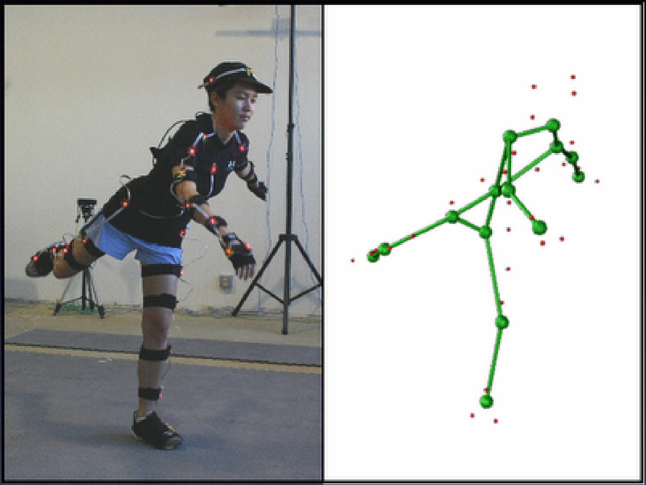




Motion Capture

Your average motion capture suit. Source: New Line Cinema

Marker Based Motion Capture

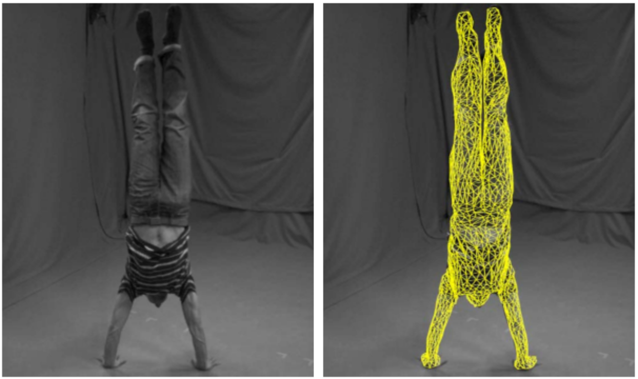


Source: http://kalyankrishna4886.wordpress.com/2010/11/17/motion-capture/

Performance Capture

Problem: **clothing deforms non-rigidly**.

Goal: Capture **shape**, **movement** and **textural appearance** simultaneously.



Challenges in Performance Capture

- **Representation** - capture enough detail, but avoid over-constraining and limit sensitivity to noise.
- **Algorithm** - robust, recovers from errors, captures enough details and creates plausible results.
- **Parameter Space** - many degrees of freedom.

Performance Capture from Sparse Multi-View Video

- De Aguiar, Edilson, Stoll, Carsten, Theobalt, Christian, Ahmed, Naveed, Seidel, Hans-Peter and Thrun, Sebastian
- ACM Transactions on Graphics, 2008

Input

High resolution 3d scan: triangle mesh $\mathcal{T}_{\triangle}$ with vertices V_i

Multi-view camera input

Output

Location of each vertex at each frame $V_i(t)$

Laplacian Deformation

$$\boldsymbol{L}\boldsymbol{v} = \delta$$

$$\boldsymbol{L} = \boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{G}$$

$$\delta = \boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{D}\boldsymbol{g}$$

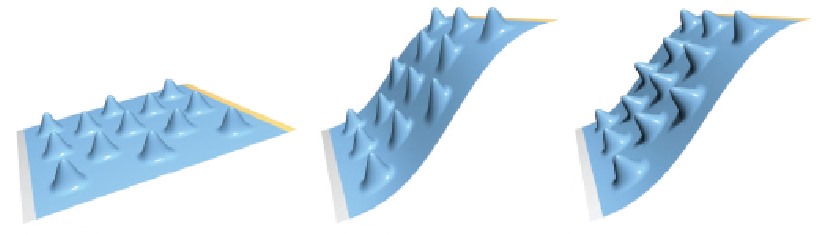
- $\boldsymbol{G}$ Discrete gradient operator.
- $\boldsymbol{g}$ set of tetrahedron gradients
- $\boldsymbol{D}$ diagonal matrix with tetrahedra volumes.
- $\boldsymbol{v}$ vertices

Source: Botsch, 2006

Laplacian Deformation - Local Structures

Problem: Local structures do not transform correctly under linear transformations.

Solution: Extract rotations for each tetrahedra, transform the original tetrahedra to derive a new $\mathbf{g}$.



Source: Botsch, 2006

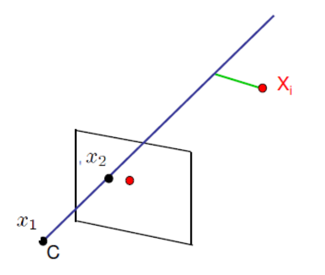
Toolkit - Plücker Line

Plücker line: $L = (n, m)$

Calculated for camera center (x_1) and point on image plane (x_2):

$$n_i = x_1 - x_2$$

$$m_i = x_1 \times n$$



Source: Modelling Reality, MPII, Summer 2012

Algorithm Overview

Algorithm

- Construct initial coarse tetrahedra mesh from high resolution scan: $\mathcal{T}_{\triangle} \rightarrow \mathcal{T}_{tet}$
- Register model to first pose in the sequence.
- For each frame:
  1. Deform global pose using $\mathcal{T}_{tet}$
  2. Transfer deformations to high res mesh: $\mathcal{T}_{tet} \rightarrow \mathcal{T}_{\triangle}$
  3. Infer shape detail on $\mathcal{T}_{\triangle}$.

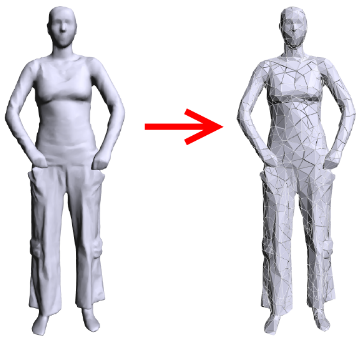
Initialization – Surface to Volume Representation.

High resolution triangle mesh $\mathcal{T}_\triangle$ with vertices V_i defines a **surface.**

Build a lower resolution tetrahedra representation - $\mathcal{T}(v_i) = \sum c_i T_{\triangle_i}$- $\mathcal{T}_\triangle$ – represent a **volume.**

DoF: $30 - 40k$ triangles $\rightarrow 5 - 6k$ tetrahedra.

Register model to the pose in the first frame by Iterative Closest Point.



Tracking the Global Pose

- For each frame, generate **silhouettes** by background subtraction.
- Find **correspondences** between consecutive frames using SIFT features. Choose good correspondences by reprojecting (Plücker line).
- Solve Laplacian problem iteratively:

$$\boldsymbol{L}\boldsymbol{v} = \delta$$

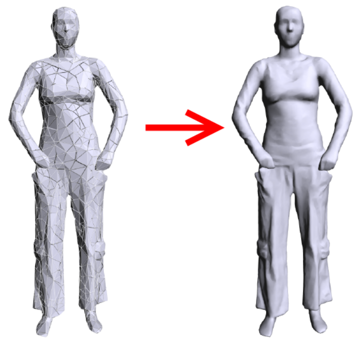
- Constrains are factored into δ.
- Minimizes non-rigid reformations in mesh.

Transfer Pose to High Resolution Surface Scan

Each triangle vertex is linearly interpolated from initial vertex weights:

$$V_{\triangle} = c_i V_{tet_i}$$

Multiple tetrahedra vertices in support region ensure smooth transfer of pose.



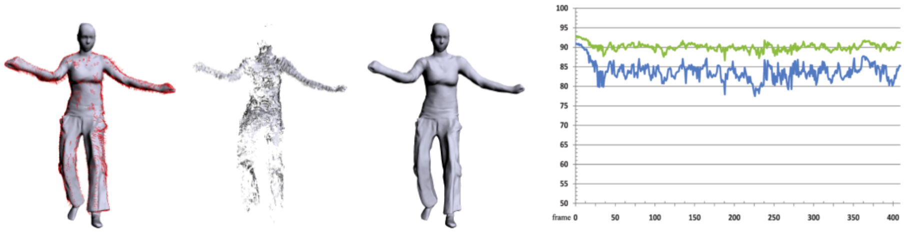
Recovering Surface Details

High frequency details cannot be recovered from the coarse representation.

Build high resolution constrains from rim vertices and silhouette projection.

Solve least squares Laplacian, which minimizes:

$$argmin\{\underbrace{\|Lv-\delta\|^2}_{\text{Laplacian matrix}}+\underbrace{\|CV-q\|^2}_{\text{silhouette constraints}}\}$$



Results (I)


Performance Capture
from Sparse Multi-view Video
Edilson de Aguiar[1] Carsten Stoll[1] Christian Theobalt[2] Naveed Ahmed[1]
Hans-Peter Seidel[1] Sebastian Thrun[2]
[1] MPI Informatik
[2] Stanford University
(with audio)

Results (II)

Complex pose are reconstructed accurately.

Results (III)

Clothes behave a single mesh, hands and feet are not captured

Results (IV)

Folds (texture) does not match new topology.

Results (V)

Laplacian deformation heavily penalizes folds and corners.

Discussion and Future Work

- **Closed surface mesh** model does not fully reflect the real world.
- No **separation** between different surfaces.

Proposals: Model clothing as more than one mesh, not necessarily closed, or add a different energy measure.

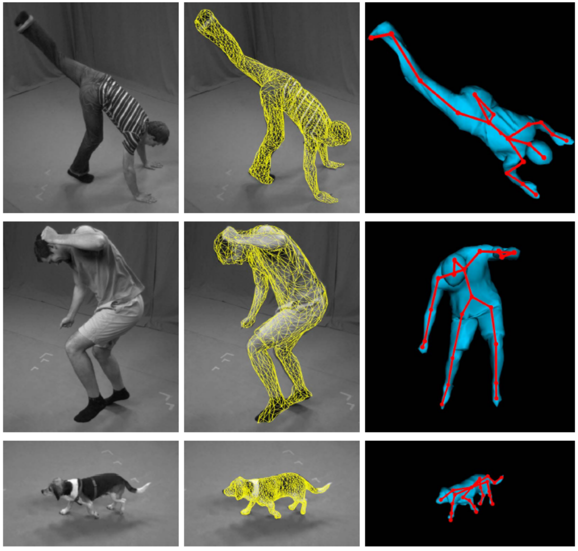
- High resolution texture details are captured **once**.

Proposal: Capture surface scan in more than one pose, blend in surface detail from closest poses.

Performance Capture from Sparse Multi–View Video – Summary

- The algorithm **automatically** captures spatio-temporal **coherent shape, motion and texture**.
- Results are **highly plausible**.
- Novel use of skeleton-less **volume deformation** and surface deformations.
- Closed mesh model **penalizes heavily derivations from volume constancy and smoothness assumption**.
- **Cannot be used as a complete solution** for performance capture at industry standard.

Motion Capture Using Joint Skeleton Tracking and Surface Estimation



Problem Statement

Motion Capture Using Joint Skeleton Tracking and Surface Estimation

- Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel
- CVPR 2009.

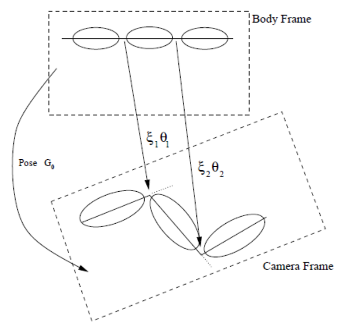
Input

- Video from eight cameras
- 3D surface mesh $\mathcal{M}$
- Skeleton joint locations θ

Output

- Skeleton motion. $\theta(t)$
- Geometry with constant connectivity. $\mathcal{M}(t)$

Toolkit - Kinematic Chain Skeletons

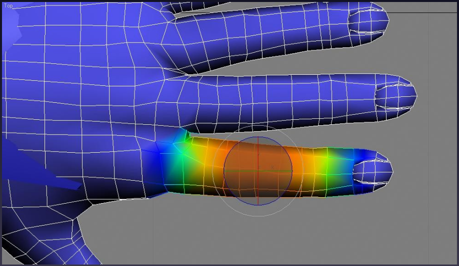


Global Pose: $\chi = (\theta_0\hat{\xi_0}, \Theta)$ 36 DoF

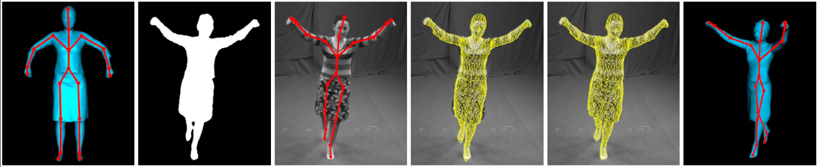
$$T_\chi V_i = \prod_{j=0}^{n_{k_i}} \exp\left(\theta_{\tau_{k_i}(j)}\hat{\xi}_{\tau_{k_i}(j)}\right) V_i$$

Skinning Weights

$$V_i' = \sum_j \underbrace{\phi_{i,j}}_{\substack{\text{bone}\\\text{weight}}} \quad \underbrace{T_j}_{\substack{\text{bone}\\\text{transformation}}} \quad V_i$$



Source: http://img202.imageshack.us/img202/299/31606317.jpg

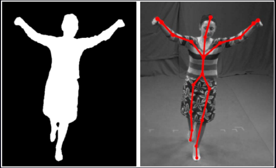


(a) Mesh model (b) Segmented images (c) Estimated skeleton (d) Deformed surface (e) Estimated surface (f) Estimated 3D Model

Algorithm

0. Build initial model.
1. Skeleton based pose estimation - local optimization.
2. Local optimization error handing with global optimization.
3. Surface estimation using Laplacian surface deformation.

Local Optimization (I)



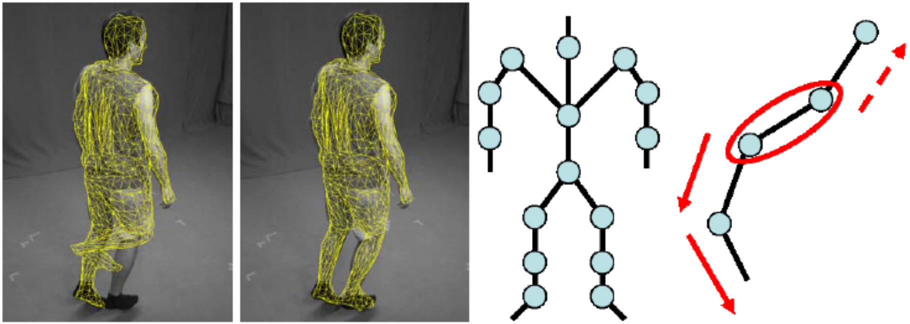
Extract (V_i, x_i) 2D-3D constraints from contour and texture features.
Transform vertices:

$$T_\chi V_i = \prod_{j=0}^{n_{k_i}} \exp\left(\theta_{\tau_{k_i}(j)} \hat{\xi}_{\tau_{k_i}(j)}\right) V_i$$

Error term:

$$\left\| \prod (T_\chi V_i) \times n_i - m_i \right\|_2$$

1. Local Optimization- Error Measurement



Energy per limb:

$$E_k(\chi) = \frac{1}{K} \sum_{\{i, k_i = k\}} \| \prod (T_\chi V_i) \times n_i - m_i \|_2^2$$

Error is propagated through the kinematic chain.

2. Fixing Errors by Global Optimization

Project the pose to a smaller subspace by fixing parameters for non-labelled bones.

$$P(\chi) \to \tilde{\chi} \in \mathbb{R}^m,\ m \leq d$$

Solve using a particle based global optimizer.

$$\arg\min_{\tilde{\chi}} \{ \underbrace{E_S(P^{-1}(\tilde{\chi}))}_{\text{silhouette consistency}} + \gamma\ \underbrace{E_R(\tilde{\chi})}_{\text{prediction deviation}} \}$$

$\gamma = 0.01$

2. Global Optimization - Error Terms

Projection:

$$E_S^c(\chi) = \frac{1}{\mathsf{area}(S_c^p)} \sum (S_c^p(\chi) - S_c) + \frac{1}{\mathsf{area}(S_c)} \sum (S_c - S_c^p(\chi))$$

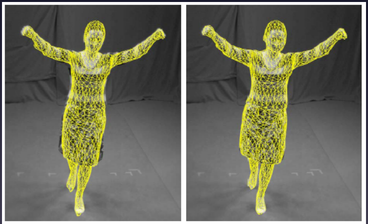
Prediction:

$$E_R(\tilde{\chi}) = \|\tilde{\chi} - P(\tilde{\chi})\|_2^2$$

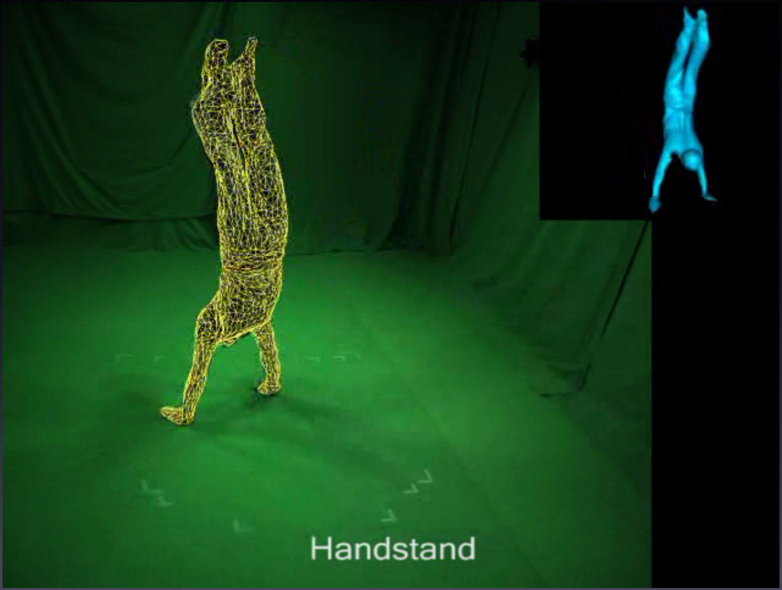
3. Refined Surface Estimation

- Decouple skeleton from mesh.
- Constraints – correspondences between rim vertices and SIFT features.
- 2D-2D constrained surface Laplacian deformation.

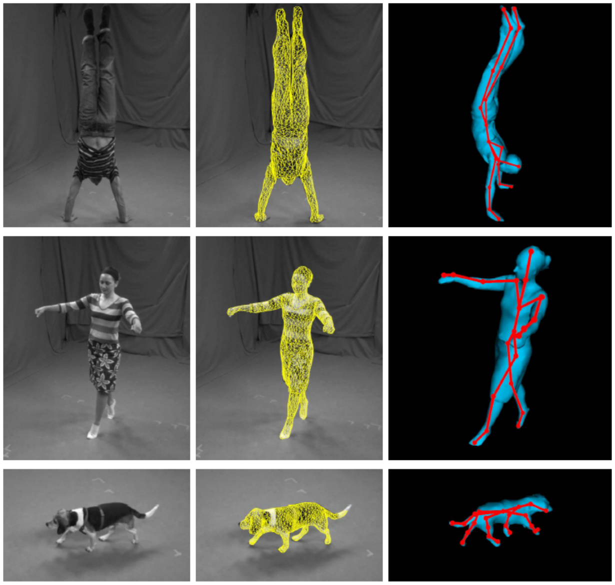
$$argmin\{\underbrace{\|Lv-\delta\|_2^2}_{\text{Laplacian matrix}}+\alpha\underbrace{\|C_{sil}V-q_{sil}\|_2^2}_{\text{silhouette constraints}}\}$$



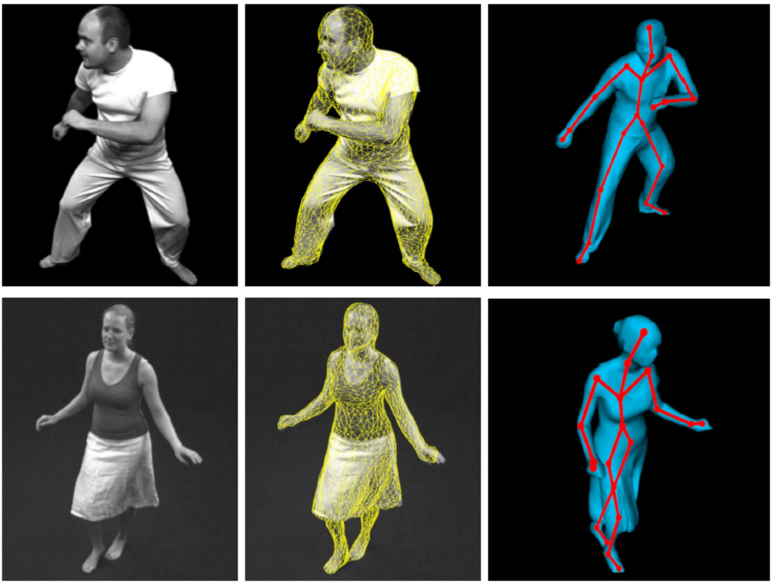
Results (Movie)



Results (II)



Results (III)

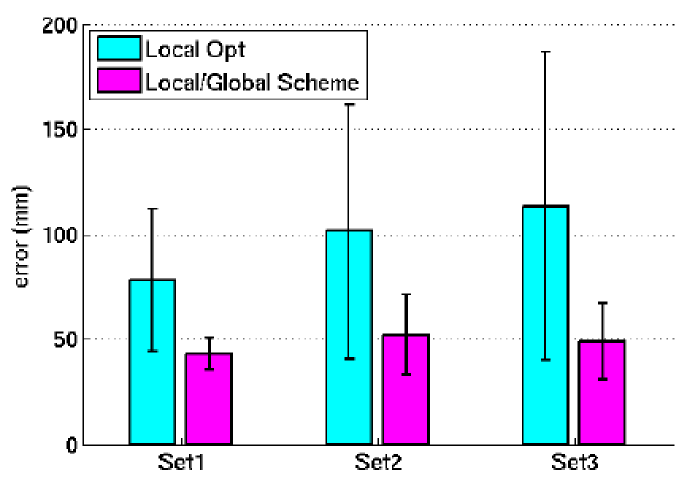


Results - Reduction in Global Optimization DoF

Sequence	Frames	Views	Model	%DoF
Handstand	401	8	Scan	3.3
Wheel	281	8	Scan	0.2
Dance	574	8	Scan	4.0
Skirt	721	8	Scan	0.2
Dog	60	8	Scan	98.3
Lock [25]	250	8	S-f-S	33.9
Capoeira1 [10]	499	8	Scan	3.4
Capoeira2 [10]	269	8	Scan	11.8
Jazz Dance [10]	359	8	Scan	43.8
Skirt1 [10]	437	8	Scan	7.2
Skirt2 [10]	430	8	Scan	6.5
HuEvalI S4 [23]	1258	4	SCAPE	79.3

Average dimensionality of global search space in percentage of full search space.

Results - Local Only vs. Local-Global Optimization Error Measurements



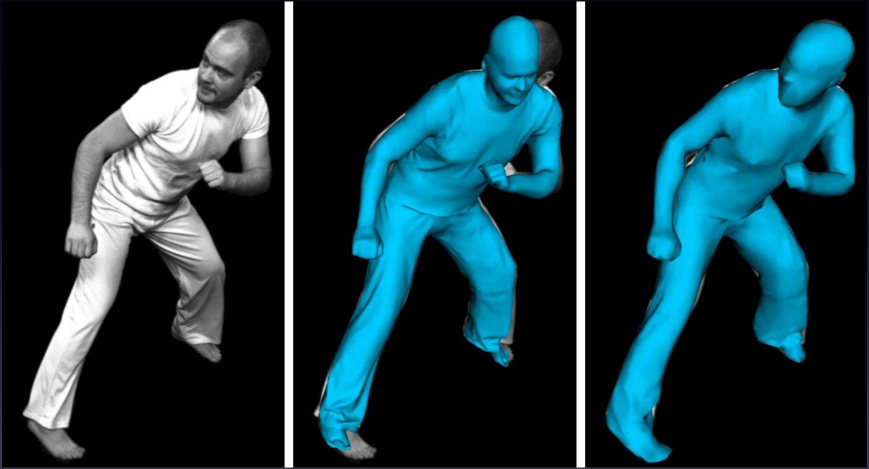
Discussion

- Choosing **threshold** for labelling bones as error.
- No way to **validate** against a ground truth.
- **Recovers feet, hands, and head position accurately**.
- **High freq** details are not well recovered.

Motion Capture Using Joint Skeleton Tracking and Surface Estimation

- Novel local-global optimization approach is an **effective and robust.**
- **Accurately** captures global pose including extremities.
- Allows **automatic** tracking.
- **High frequency details are not well recovered**.

Two Approaches to Marker–Less Performance Capture



Left: Input image. Center: Output of first paper. Right: Output of second paper

Two Approaches to Marker–Less Performance Capture

Similarities

- **Accurate, automatic, robust.**
- Require **very little user input** and **no supervision.**
- **Two-levels optimization** strategy.
- **Laplacian deformation**, high freq results are penalized in similar ways.
- Similar issues **measuring accuracy**.
- Can only handle limited **topology**.

Two Approaches to Marker–Less Performance Capture

Differences

- **Optimization** strategies.
- **Coarse representation** models.
- **Pose accuracy** vs. **high freq detail**.
- **Novelty** of model vs. of optimization strategy.

Future Work

- Allow clothing to **deform independently**.
- Add different **energy measures**.
- **Combine** the best of both papers into one performance capture pipeline.

Sources I

- Christoph Bregler and Jitendra Malik.
  Tracking people with twists and exponential maps.
  In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 8–15. IEEE, 1998.
- Mario Botsch, Robert Sumner, Mark Pauly, and Markus Gross.
  Deformation transfer for detail-preserving surface editing.
  In *Vision, Modeling & Visualization*, pages 357–364. Citeseer, 2006.
- Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun.
  Performance capture from sparse multi-view video.
  In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM, 2008.

Sources II

- Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and H-P Seidel.
  Motion capture using joint skeleton tracking and surface estimation.
  In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1746–1753. IEEE, 2009.
- Carsten Stoll.
  A volumetric approach to interactive shape editing.
  Max-Planck-Institut für Informatik, 2007.