Robust Human Body Shape and Pose Tracking

Chun-Hao Huang[1] Edmond Boyer[2] Slobodan Ilic[1]

[1] Technische Universität München

[2] INRIA Grenoble Rhône-Alpes

Marker-based motion capture (mocap.)

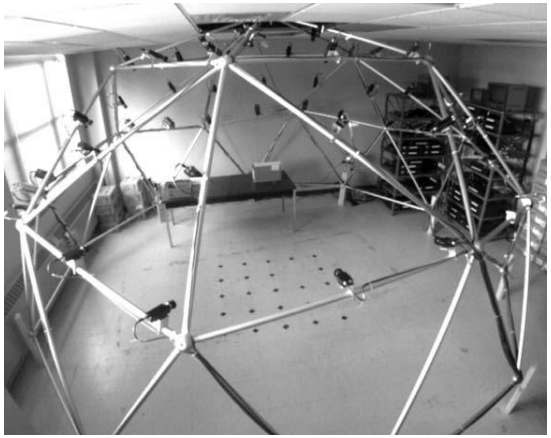
- Adventages:
  - precision, reliability
  - little data (couple of kB/frame for 50 cameras, 5 people)
  - real-time processing, visualization & retargeting.

Giant Studios (L.A. Noire set)

- Disadvantages:
  - Attaching, removing, re-attaching markers is tedious.
  - Markers can interfere with the movement.
  - Markers prevent the simultaneous acquisition of shape and motion.

Marker-less mocap.

- Multiple camera setup is usually required.
- Accquisition of both motion and shape.



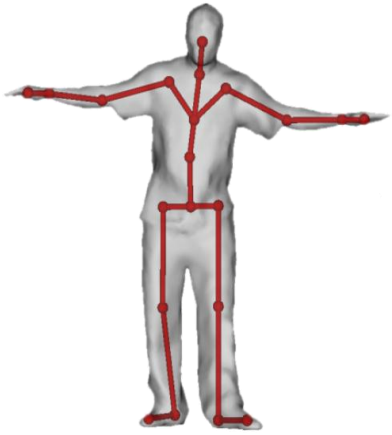
3D dome (CMU)

Grimage (INRIA)

Motivation

- Methods that assume a skeleton usually produce skinning artifacts, require 2nd stage shape refinement.

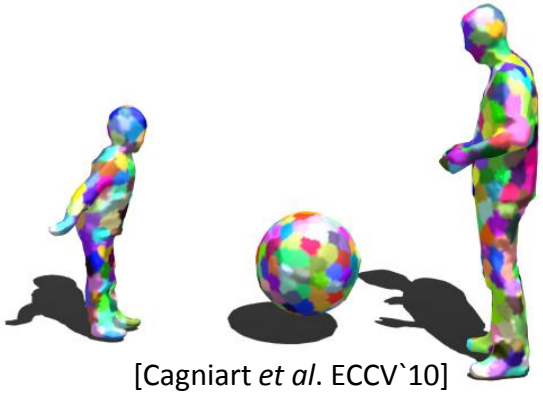
  Degree of freedom (DoF): $N_J \times 6 \ (< 10^2)$



[Vlasic *et al*. ToG`08]
Skeleton-based

- Purely-surface-based methods handle non-rigid surface deformation better, but do not provide the pose.

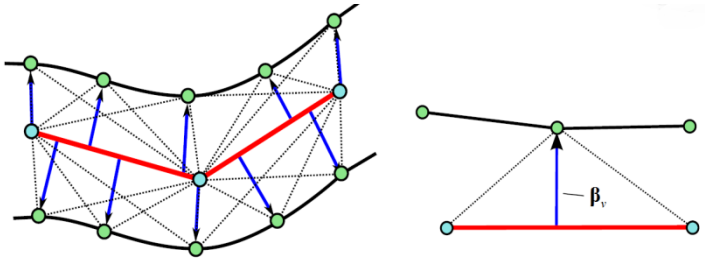
  DoF: $N_P \times 6 \ (\cong 10^3)$



[Cagniart *et al*. ECCV`10]
Surface-based

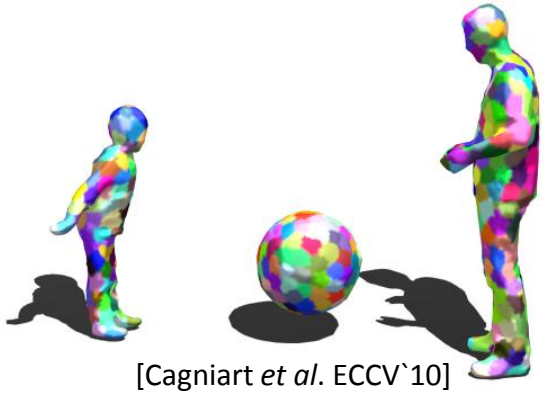
Contribution

- bone differential coordinate



[Straka *et al*. ECCV`12]

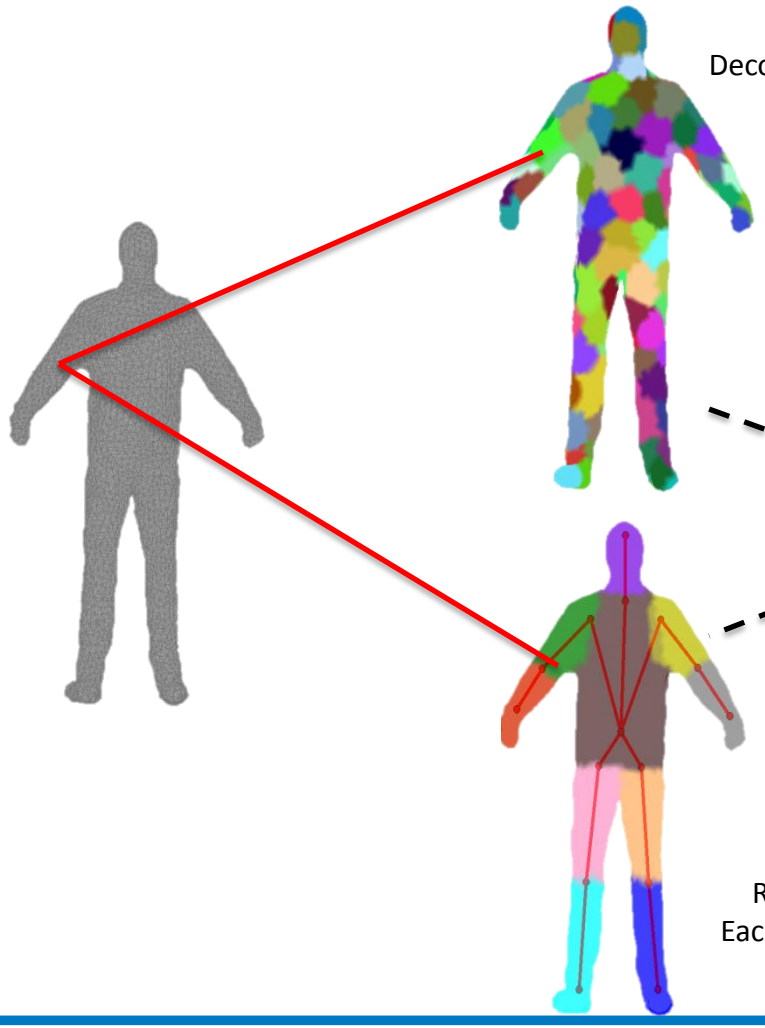
probablistic surface deformation



[Cagniart *et al*. ECCV`10]

- A learning-based outlier rejection scheme.

Preprocessing step: model



Decompose the reference mesh into patches
[Cagniart *et al.* CVPR`10]

Shape parameter

$$\Theta = \left\{ \mathbf{R}_k, \mathbf{c}_k \right\}_{k=1}^{N_P}$$

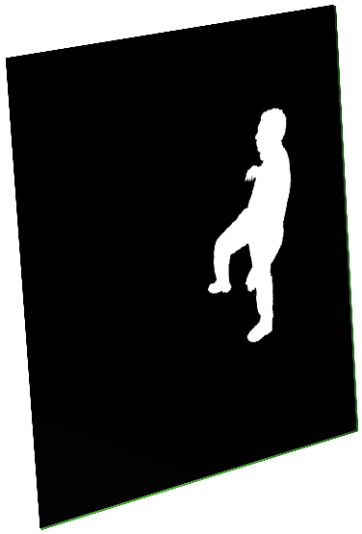
$$\mathcal{M}^0$$

Pose parameter

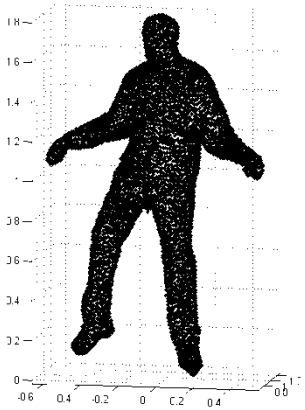
$$\mathbf{J} = \left\{ \mathbf{x}_j \right\}_{j=1}^{N_J}$$

Rig the skeleton inside the mesh.
Each vertex belongs to one body part.
[Baran *et al.* TOG`07]

Preprocessing step: input data

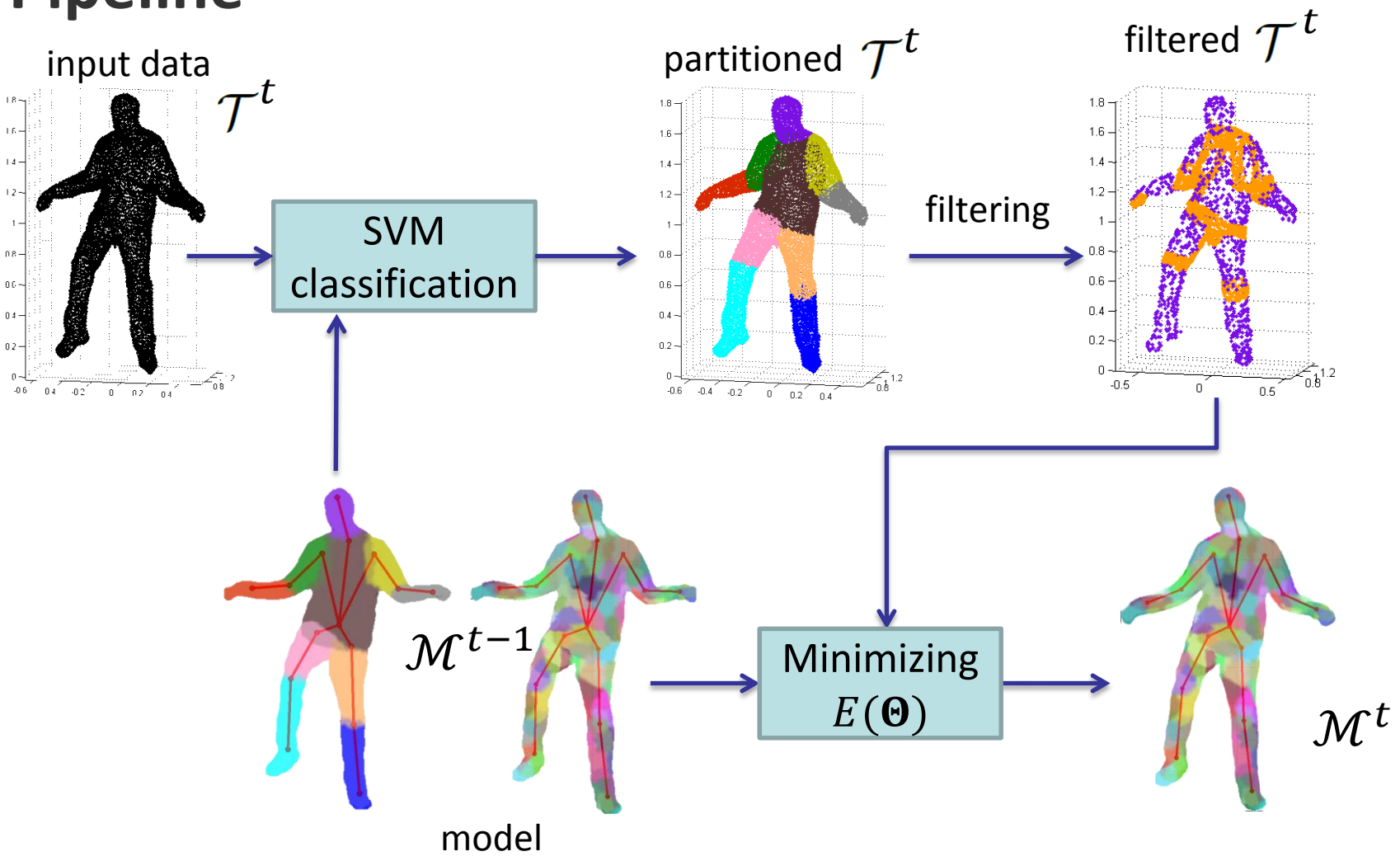


$$\mathcal{T} = \{\mathbf{y}_i\}_{i=1}^{N_T}$$



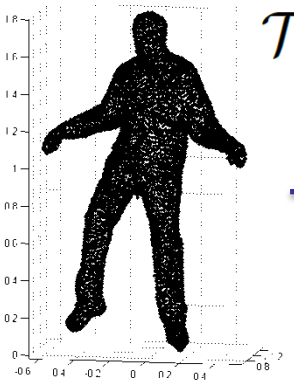
For each time stamp t, visual hull is reconstruced from silhouettes, which serves as our observations

Pipeline

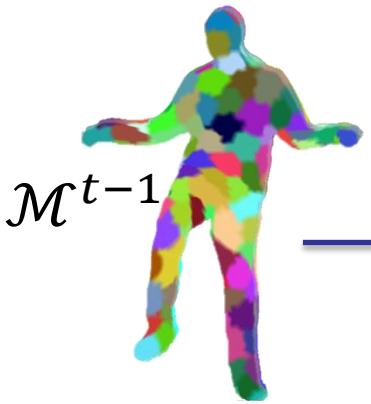

input data
$\mathcal{T}^t$
SVM classification
partitioned $\mathcal{T}^t$
filtering
filtered $\mathcal{T}^t$
$\mathcal{M}^{t-1}$
model
Minimizing $E(\mathbf{\Theta})$
$\mathcal{M}^t$

Pipeline

input data



$\mathcal{T}^t$



$\mathcal{M}^{t-1}$

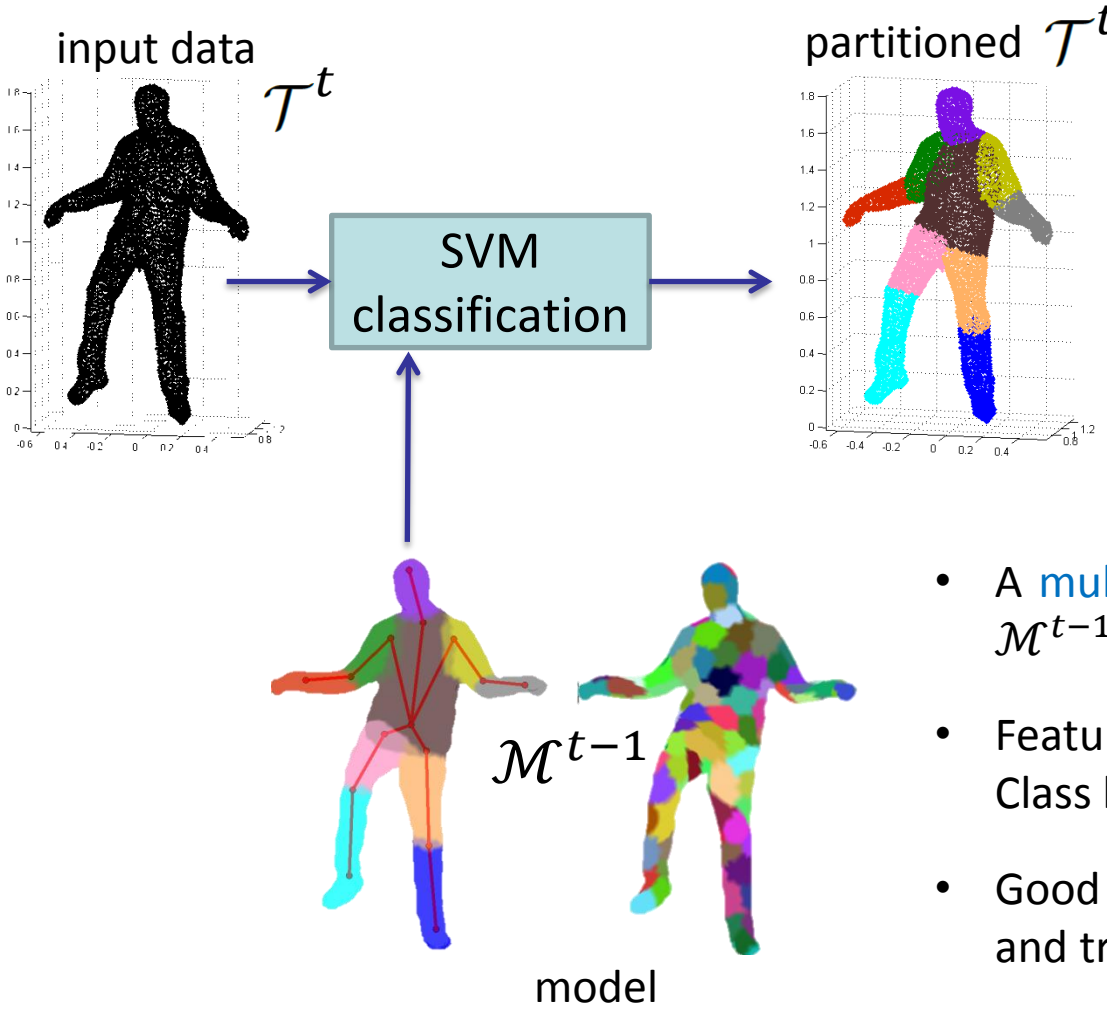
model

Minimizing
$E(\boldsymbol{\Theta})$

[Cagniart *et al*. ECCV`10]

$\mathcal{M}^t$

SVM-based body part classification



- A multi-class linear SVM is trained on $\mathcal{M}^{t-1}$ and tested on $\mathcal{T}^t$
- Feature: 3D coordinate of vertices. Class label: rigid body part label.
- Good compromise between accuracy and training time.

Filtering point cloud

- Bone $\mathcal{T}_b$:

  patches on the bone often move rigidly together.

  → sub-sample the observations.

- Joint $\mathcal{T}_g$:

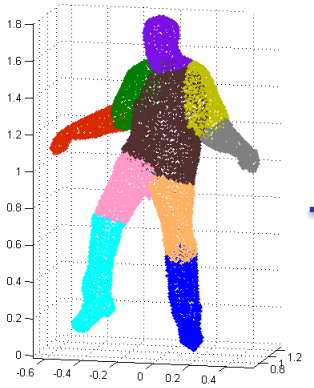
  patches on the joint have non-rigid deformation.

  → keep all the observations.

- Outlier $\mathcal{T}_o$:

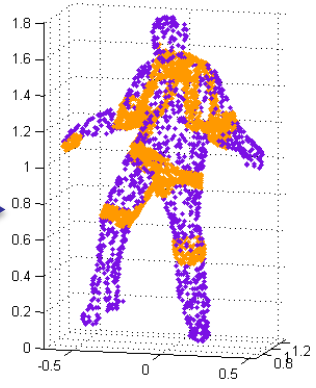
  abandon all the observations.

partitioned $\mathcal{T}^t$



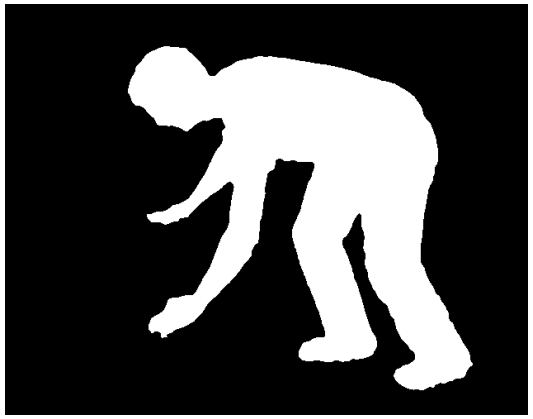
filtering

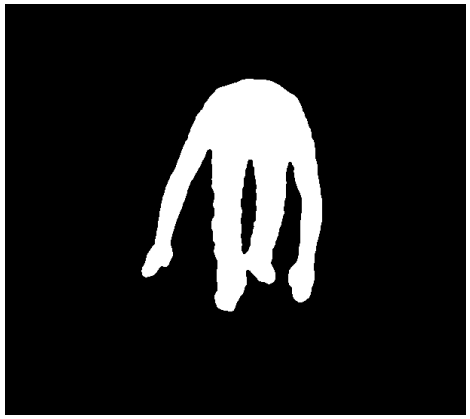
filtered $\mathcal{T}^t$

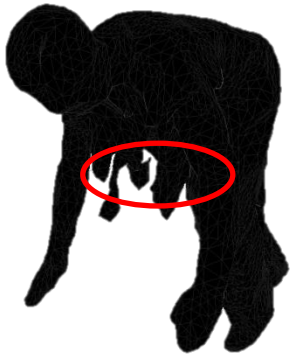


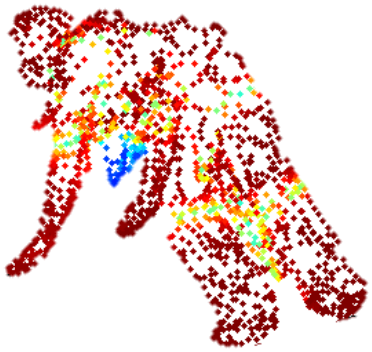
$$\mathbf{y}_i \in \begin{cases} \mathcal{T}_b & \text{if } 0.9 < p_i \\ \mathcal{T}_g & \text{if } 0.5 < p_i \leq 0.9 \\ \mathcal{T}_o & \text{if } p_i \leq 0.5, \end{cases}$$

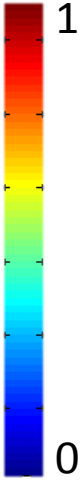
Benefit of SVM classification – outlier removal



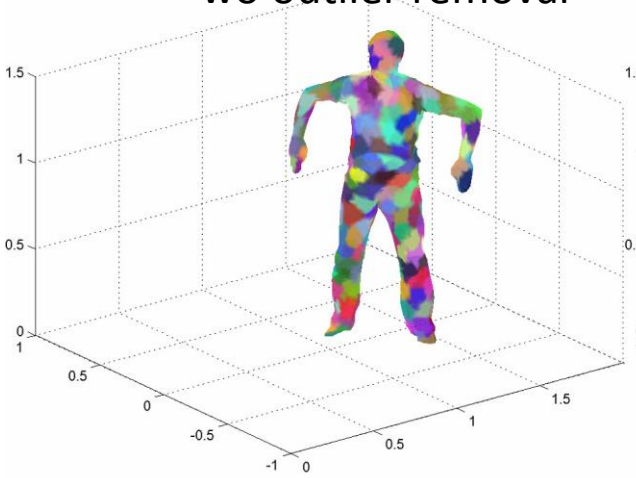




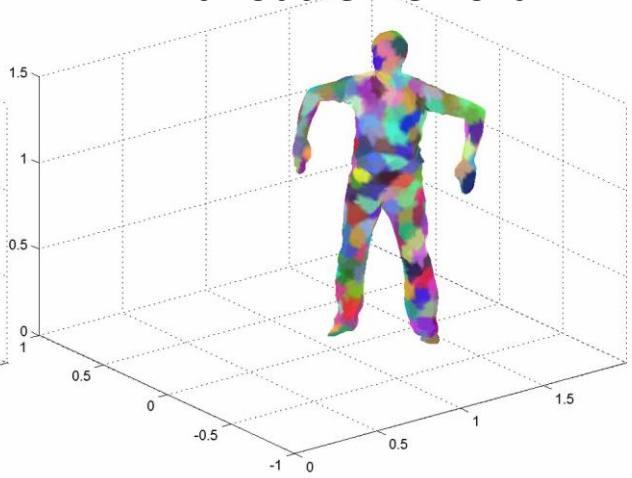




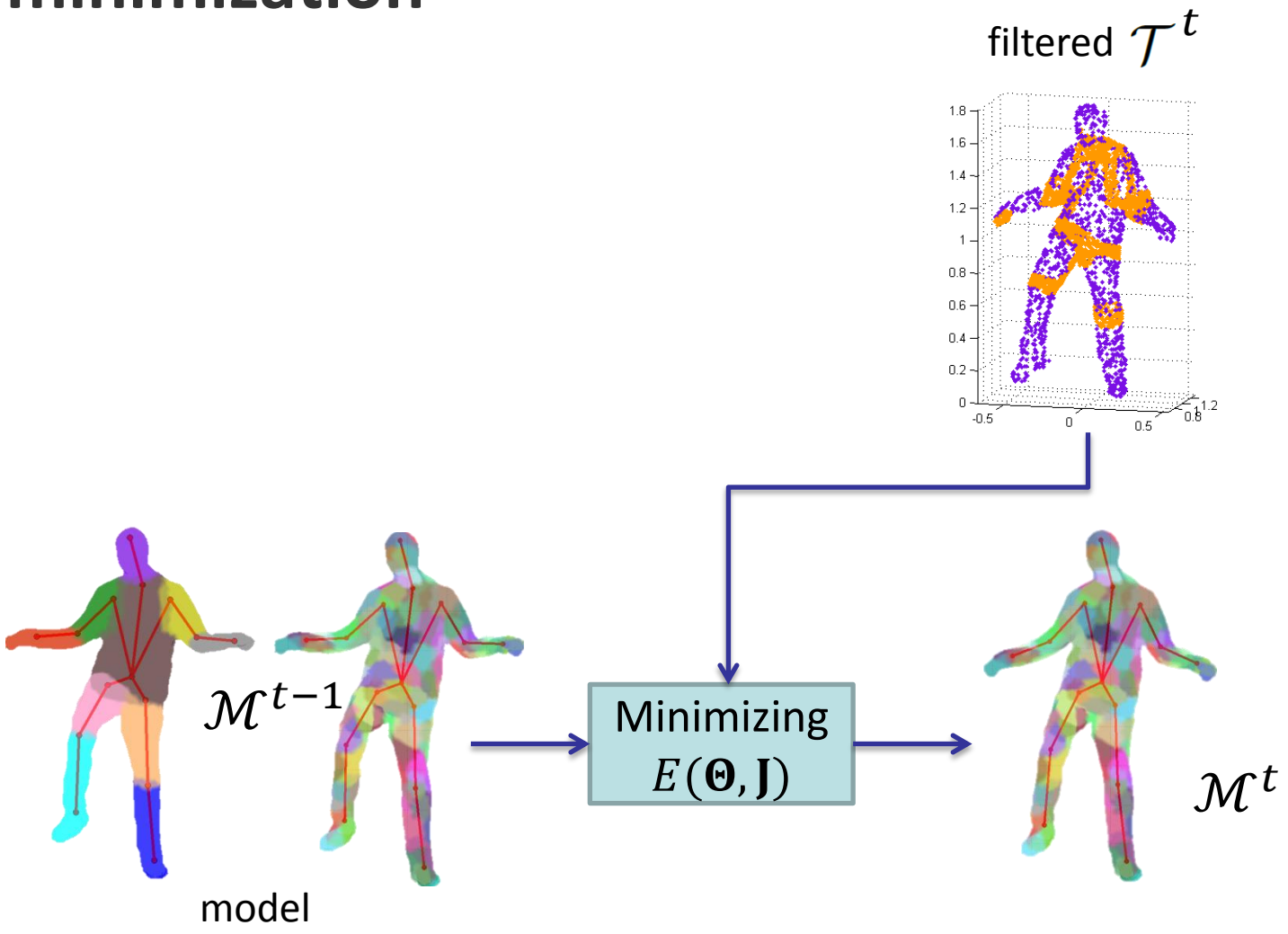
wo outlier removal



with outlier removal

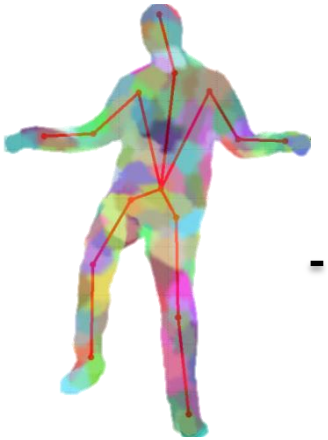


Energy minimization

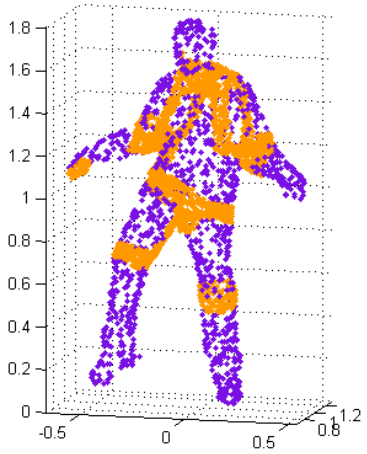


Energy function

$\mathcal{M}^{t-1}$



$E(\Theta, \mathbf{J})$



filtered $\mathcal{T}^t$

$$E(\Theta, \mathbf{J}) = E_{\text{data}}(\Theta) + E_{\text{rigid}}(\Theta) + E_{\text{bone}}(\Theta, \mathbf{J})$$

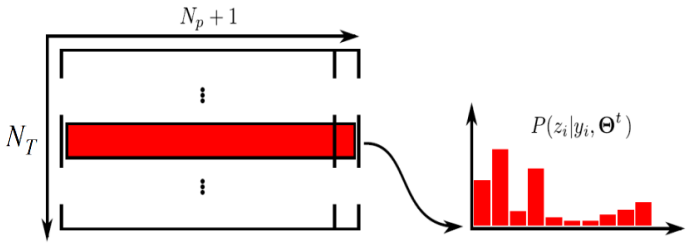
- $E_{\text{data}}(\Theta)$: how well the surface explains the observations.
- $E_{\text{rigid}}(\Theta)$: smooth the motion of neighboring patches.
- $E_{\text{bone}}(\Theta, \mathbf{J})$: keep the relationship between the mesh and the skeleton.

Data term

- $E_{\text{data}}(\boldsymbol{\Theta})$: how well the surface explains the observations.
- A probablisitic Iterative Closest Point (ICP) approach.
- Each observation has a soft assignment to every patch, updated in each iteration.
- Let observation i correspond to vertex v_i^k in P_k with a soft assigment w_i^k.

$$E_{\text{data}}\left(\boldsymbol{\Theta}\right) = \sum_{i=1}^{N_T} \sum_{k=1}^{N_P+1} w_i^k \left\| \mathbf{y}_i - \mathbf{x}_{v_i^k} \right\|^2$$

A probablistic point of view [Cagniart *et al*. ECCV`10]



- Can be interpreted as EM algorithm.
- The likelihood: Gaussian mixture model

$$P(\mathbf{y}_i \mid \boldsymbol{\Theta}) = \sum_{k=1}^{N_P+1} \Pi_k P(\mathbf{y}_i \mid z_i = k, \boldsymbol{\Theta})$$

- E-step: update the soft assignment.

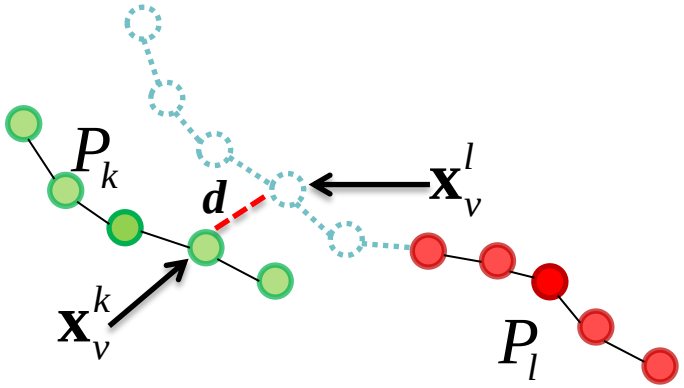
$$P(z_i = k \mid \mathbf{y}_i, \boldsymbol{\Theta}) = \frac{\Pi_k P(\mathbf{y}_i \mid z_i = k, \boldsymbol{\Theta})}{\sum_{l=1}^{N_P+1} P(\mathbf{y}_i \mid z_i = l, \boldsymbol{\Theta})} \qquad -\ln(\cdot)$$

- M-step: minimize sum of negative log likelihood (energy).

$$E_{\text{data}}(\boldsymbol{\Theta}) = \sum_{i=1}^{N_T} \sum_{k=1}^{N_P+1} w_i^k \left\| \mathbf{y}_i - \mathbf{x}_{v_i^k} \right\|^2$$

Rigidity energy [Cagniart *et al*. CVPR`10]

- $E_{\text{rigid}}(\boldsymbol{\Theta})$: smooth the motion of neighboring patches



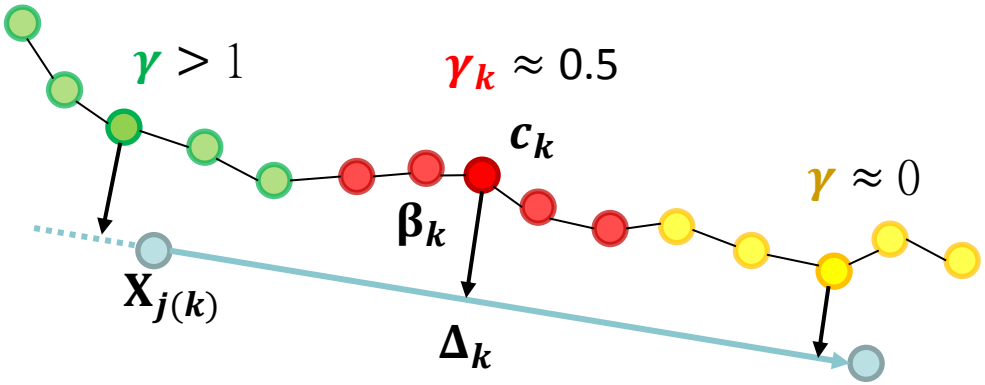
$$E_{\text{rigid}}(\boldsymbol{\Theta}) = \sum_{k=1}^{N_P} \sum_{P_l \in N_k} \sum_{v \in P_k \cup P_l} \frac{\left\| \mathbf{x}_v^k - \mathbf{x}_v^l \right\|^2}{\mathbf{d}}$$

For each patch, the real location and the predicted location should be consistent.

$\boldsymbol{\Theta}$ is implicitly encoded in $\mathbf{x}_v^k$ and $\mathbf{x}_v^l$

Bone-binding energy

- $\boldsymbol{\beta}$ coordinate : A relative displacement from patch to bone



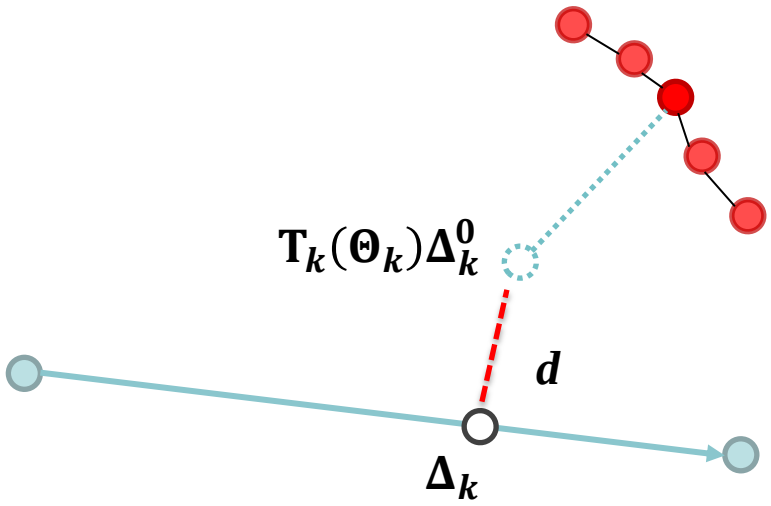
$$\boldsymbol{\beta_k = \Delta_k - c_k}$$

$$\boldsymbol{\Delta_k = \gamma_k X_{j(k)} + (1 - \gamma_k) X_{j\_child(k)}}$$

$\boldsymbol{\gamma_k}$ is computed such that $\boldsymbol{\beta_k}$ is perpendicular to the bone.

Bone-binding energy

- $E_{\text{bone}}(\boldsymbol{\Theta}, \mathbf{J})$: keep $\boldsymbol{\beta}$ consistent after transformation.



$$E_{\text{bone}}(\boldsymbol{\Theta}, \mathbf{J}) = \sum_{k=1}^{N_P} w_k \left\| \boldsymbol{\beta}_k - T_k(\boldsymbol{\Theta}_k)\boldsymbol{\beta}_k^0 \right\|^2$$

$$= \sum_{k=1}^{N_P} w_k \underbrace{\left\| \boldsymbol{\Delta}_k - T_k(\boldsymbol{\Theta}_k)\boldsymbol{\Delta}_k^0 \right\|^2}_{\boldsymbol{d}}$$

For each patch, $\boldsymbol{\Delta_k}$ predicted from the patch and $\boldsymbol{\Delta_k}$ from the bone should be consistent.

w_k is weighed according to γ_k

Energy function

- $E_{\text{data}}(\Theta)$: how well the surface explains the observations.

$$E_{\text{data}}(\Theta) = \sum_{i=1}^{N_T} \sum_{k=1}^{N_P+1} w_i^k \left\| \mathbf{y}_i - \mathbf{x}_v^k \right\|^2$$

- $E_{\text{rigid}}(\Theta)$: smooth the motion of neighboring patches.

$$E_{\text{rigid}}(\Theta) = \sum_{k=1}^{N_P} \sum_{P_l \in N_k} \sum_{v \in P_k \cup P_l} \left\| \mathbf{x}_v^k - \mathbf{x}_v^l \right\|^2$$

- $E_{\text{bone}}(\Theta, \mathbf{J})$: keep the relationship between mesh

$$E_{\text{bone}}(\Theta, \mathbf{J}) = \sum_{k=1}^{N_P} w_k \left\| \Delta_k - T_k(\Theta_k) \Delta_k^0 \right\|^2$$

regularization terms
or
deformation prior

Minimizing the energy

$$E(\Theta, \mathbf{J}) = \lambda_{\mathrm{d}} E_{\mathrm{data}}(\Theta) + \lambda_{\mathrm{r}} E_{\mathrm{rigid}}(\Theta) + \lambda_{\mathrm{b}} E_{\mathrm{bone}}(\Theta, \mathbf{J})$$

- $\lambda_{\mathrm{d}} = 10, \lambda_{\mathrm{r}} = 1, \text{and } \lambda_{\mathrm{b}} = 1$
- Each term is quadratic in terms of variables.
- Standard Gauss-Newton optimization is thus feasible.
- 3 - 4s per frame (including SVM training time).

Quantitative results

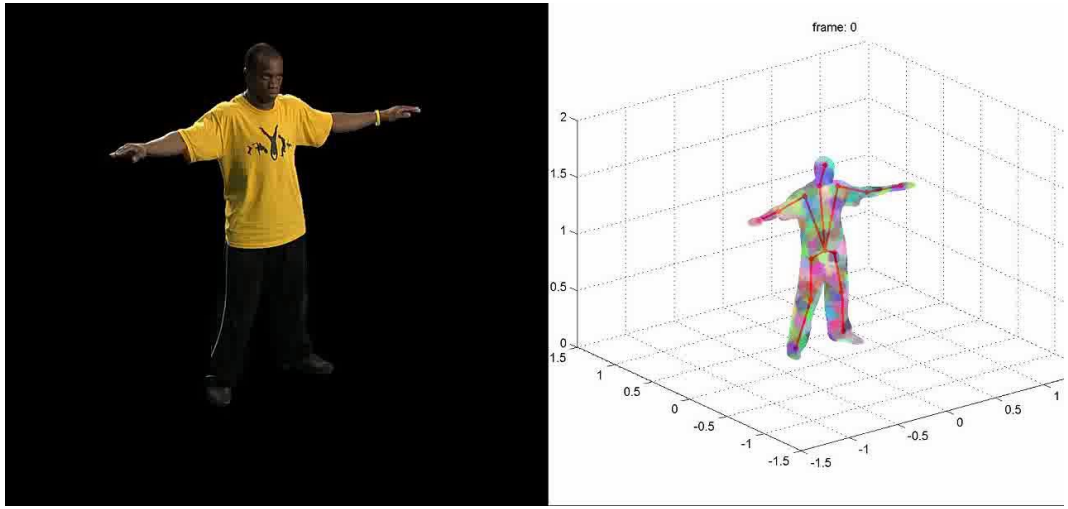
- **Pose**: 70.86*mm* error in *Walking* sequence from `HumanEvaII` benchmark. (error < 80*mm* typically corresponds to a correct pose [Sigal *et al*. IJCV`12]).
- **Shape**: reprojection error (%)

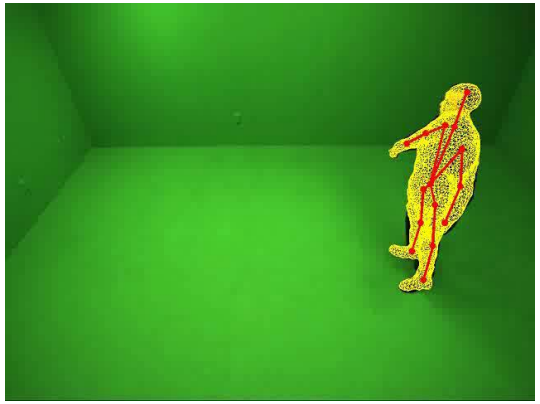
Sequence	our	surface-based [4]	inverse kinematic
Handstand 1 [1]	**15.53**	20.13	23.04
Wheel [1]	**10.28**	10.30	14.35
Skirt [1]	**11.94**	12.55	21.43
Dance [1]	9.95	**9.90**	15.01
Crane [2]	**10.79**	11.20	16.33
Handstand 2 [2]	**12.84**	13.97	15.16
Bouncing [2]	**9.87**	9.95	14.64
Free [3]	**14.12**	14.69	-

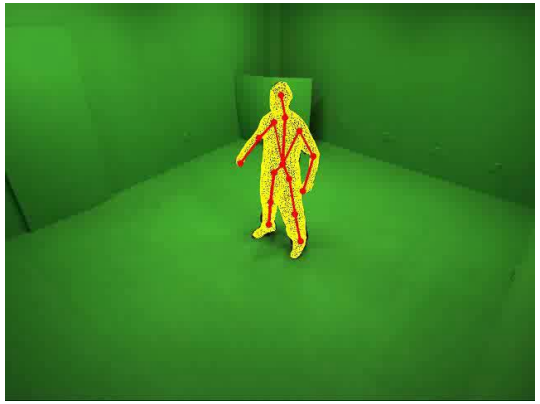
1. Gall *et al*. CVPR`09
2. Vlasic *et al*. ToG`08
3. Starck *et al*. CGA`07
4. Cagniart et al. ECCV`10

Qualitative results

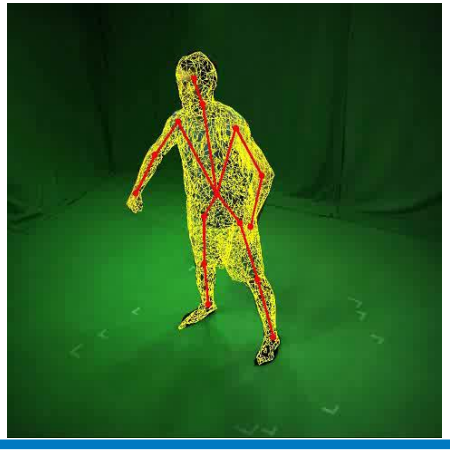
[Starck *et al*. 2007]

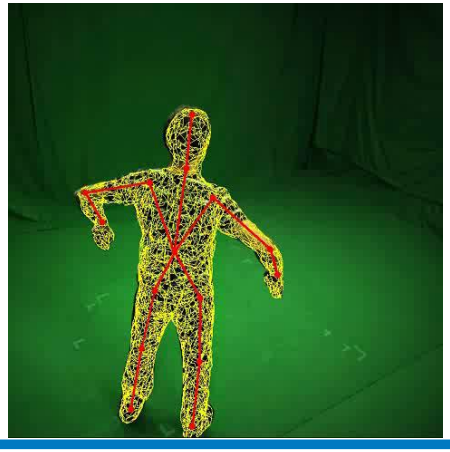


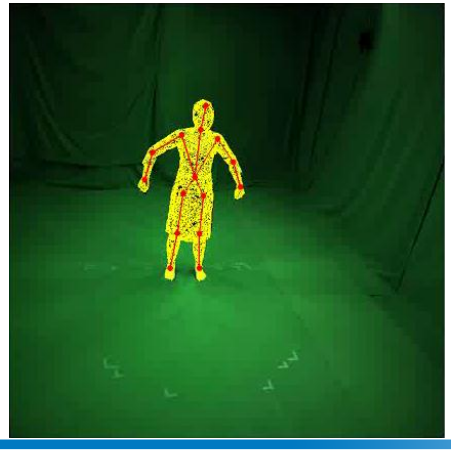




[Vlasic *et al*. 2008]







[Gall *et al*. 2009]

Conclusion and future work

- A method that jointly recovers the pose and the shape of human body has been proposed.
- We introduce a novel SVM-based classification scheme that filters target point clouds and thus helps better correspondence search.
- Future directions include alleviating the requirement of background substraction, and exploiting more photometric information.
- More experiment results in the poster session.

Thank you!

Questions?