Motion Capture of Hands in Action using Discriminative Salient Points

Luca Ballan   Aparna Taneja   Jürgen Gall   Luc Van Gool   Marc Pollefeys

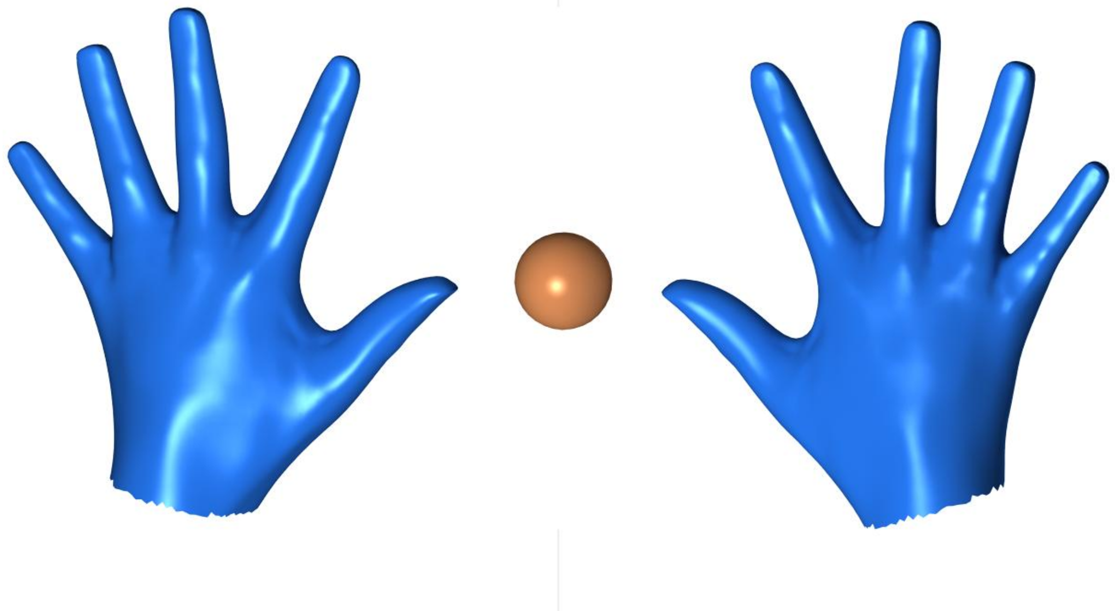
ETH Zürich

MAX-PLANCK GESELLSCHAFT

Goal: Markerless Motion Capture

Scene recorded from multiple viewing angles

+

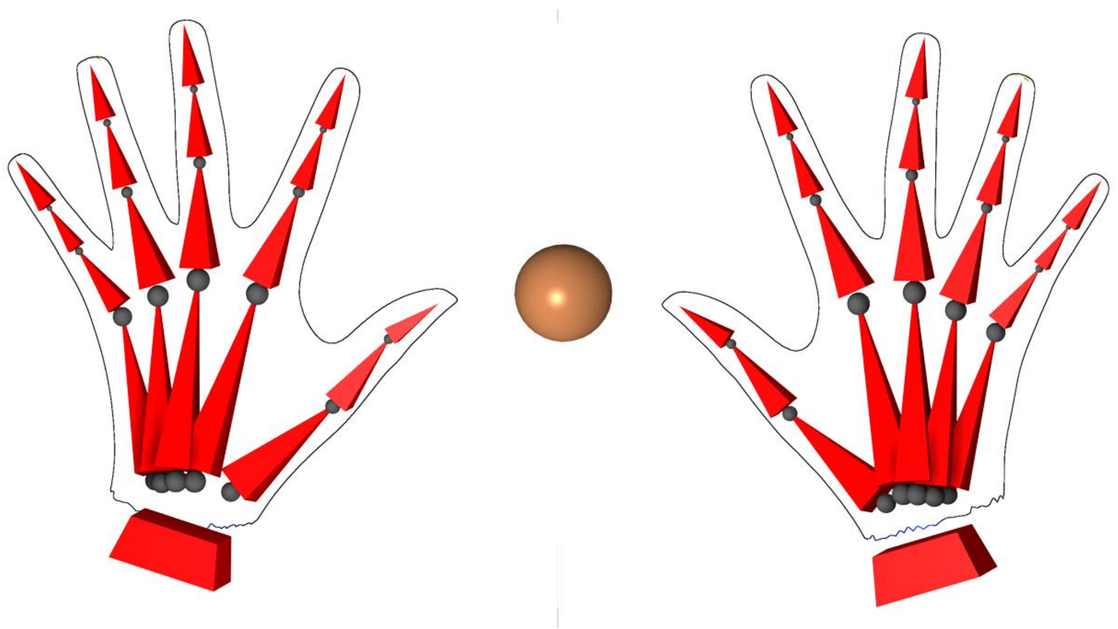


Template Model

Goal: Markerless Motion Capture

Scene recorded from multiple viewing angles

+

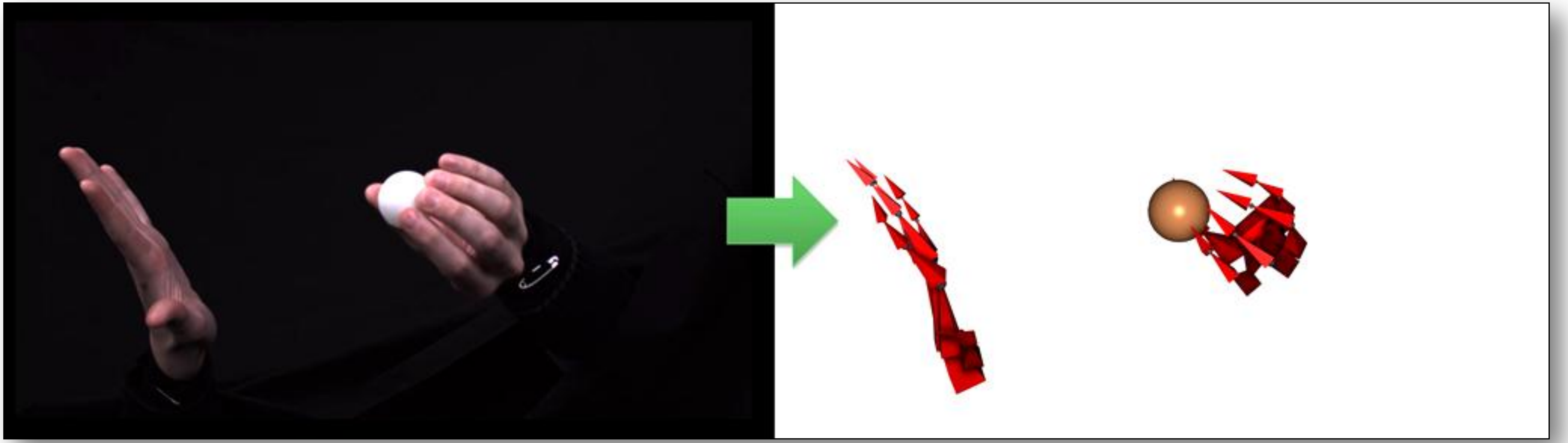


Template Model
+
Kinematic Structure

Goal: Markerless Motion Capture

Input:

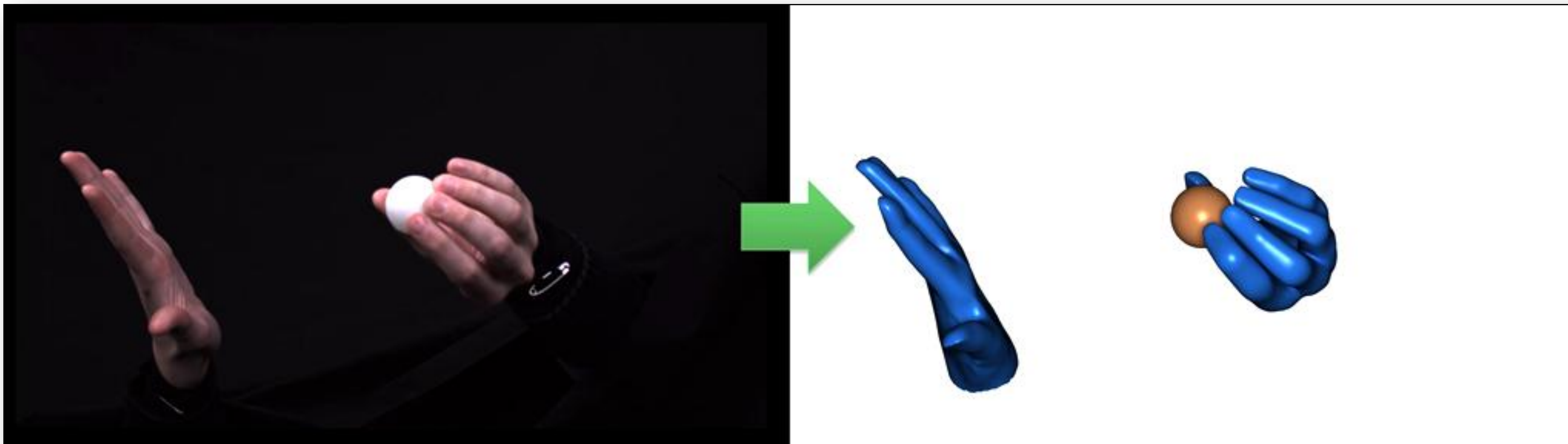
Output:



Scene Motion
(angles and positions)

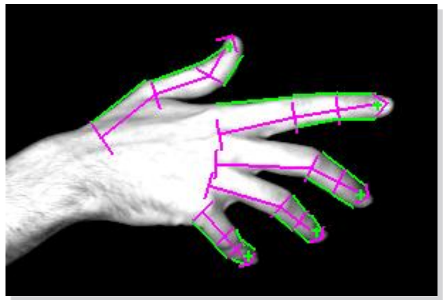
Goal: Markerless Motion Capture

Input: Output:

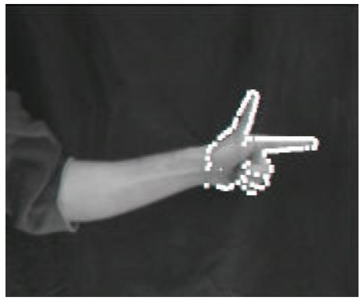


Full 3D Geometry
of the Scene

Related Work: Hand Motion Capture

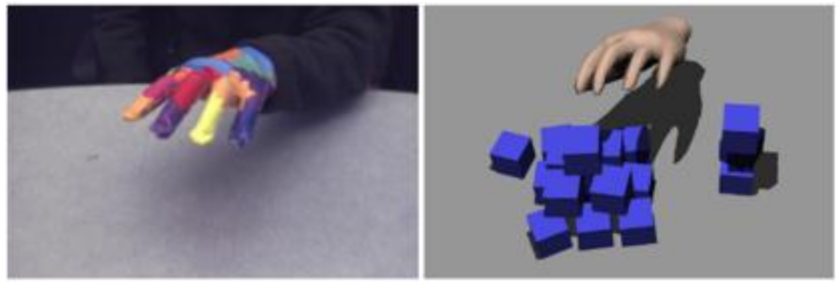


[Rehg et al. '94]

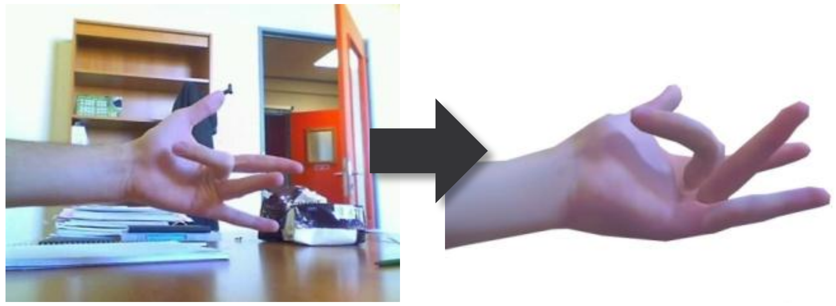


[Stenger et al. '01]

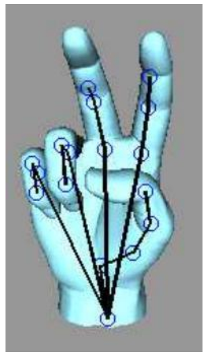
[Lu et al. '03]



[Wang et al. '09]



[de La Gorce et al. '11]

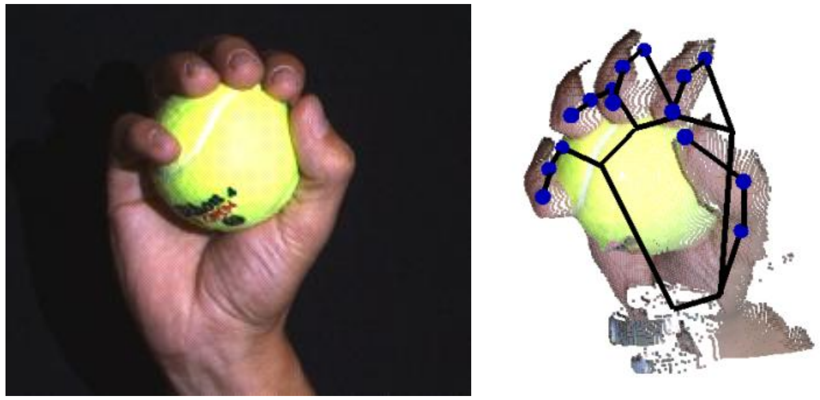


[Salzmann et al. '11]

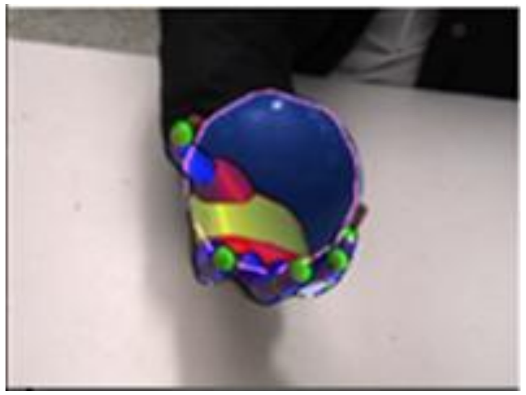
…

Single hand in isolation

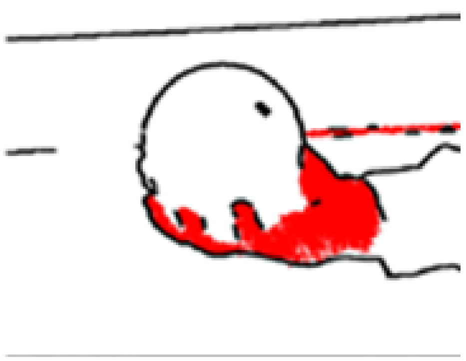
Related Work: Hand Motion Capture



[Hamer et al. '09]



[Oikonomidis et al. '11]



Hand interacting with an object

Assumption: Hand can be segmented from the object based on color

Pose estimation simpler!!

The Problem

Hands cannot be distinguished based on color

- Collisions
- Multiple occlusions
- Self-similarities

How do we deal with this?

Our Approach: The Scene Model

Assumption: Each trackable element of the scene can be modelled as an **articulated deformable object**

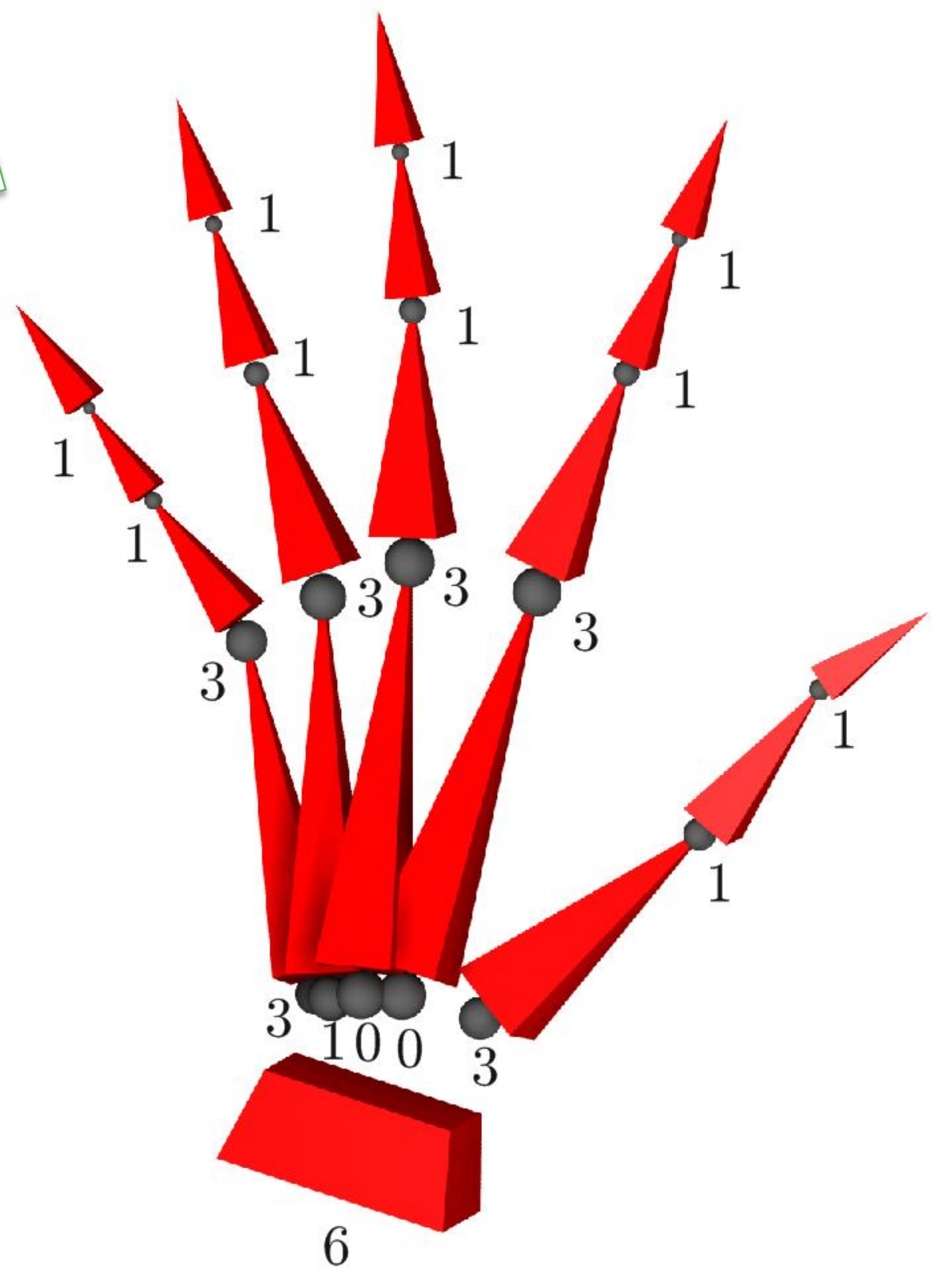
Our Approach: The Scene Model

Assumption: Each trackable element of the scene can be modelled as an **articulated deformable object**

Exponential map

ξ

Pose



20 bones

35 degrees of freedom

Our Approach: The Scene Model

Mesh representing the object at a reference pose

Multiview Stereo

[Geiger et al. 10]

[Ballan et al. 06]

Kinect

[Izadi et al. 11]

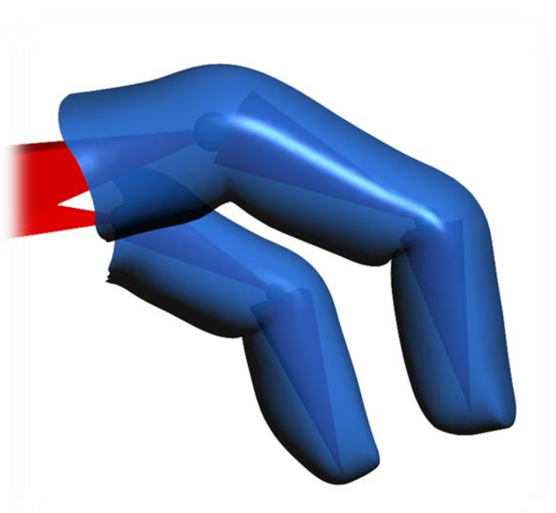
Our Approach: The Scene Model

Mesh representing the object at a reference pose

Linear Blend Skinning:

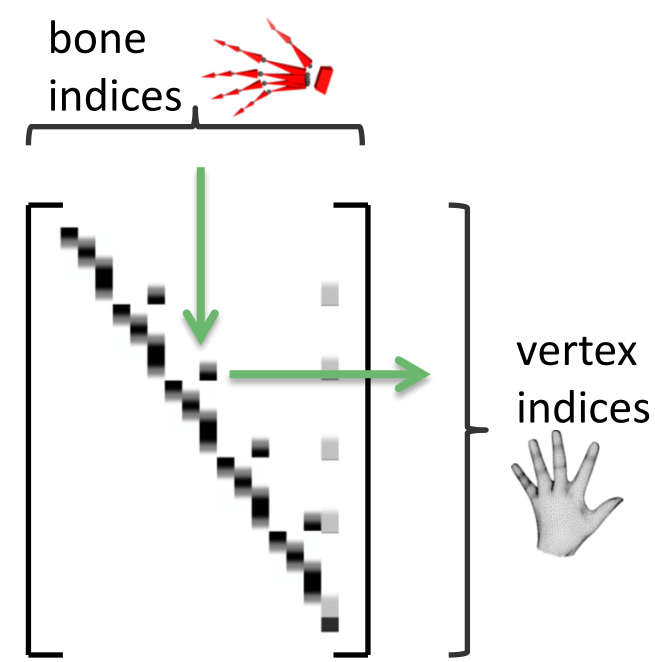
$$v_k(\theta) = \sum_{j=1}^{m} \alpha_{k,j} T_j(\theta) T_j(0)^{-1} v_k(0)$$

the motion of a vertex → the linear combination of all the motions that the vertex would undergo if rigidly attached to every bone, one at a time



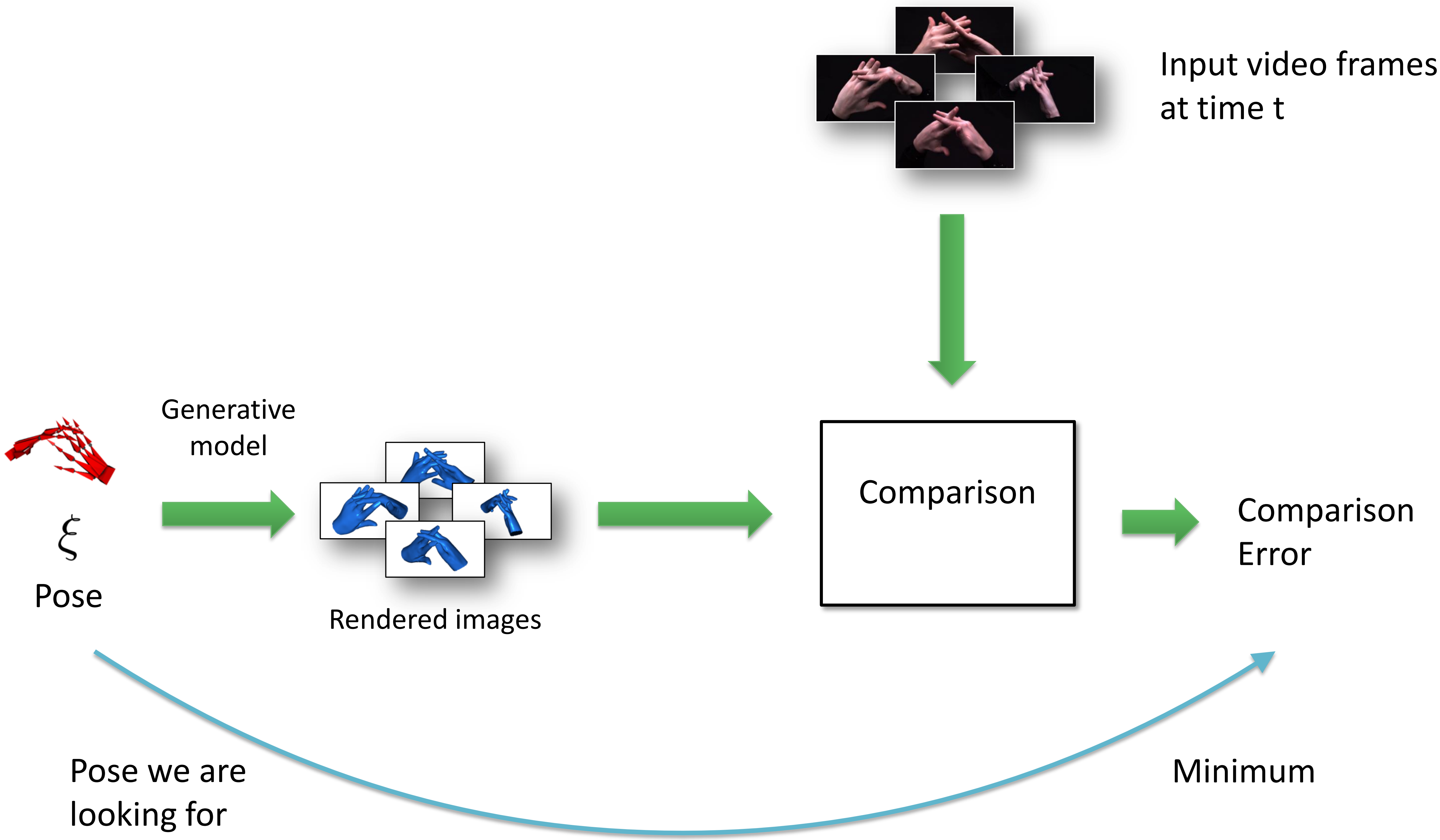
Smooth deformation of the surface

$$(\alpha_{k,j})_{k,j} =$$

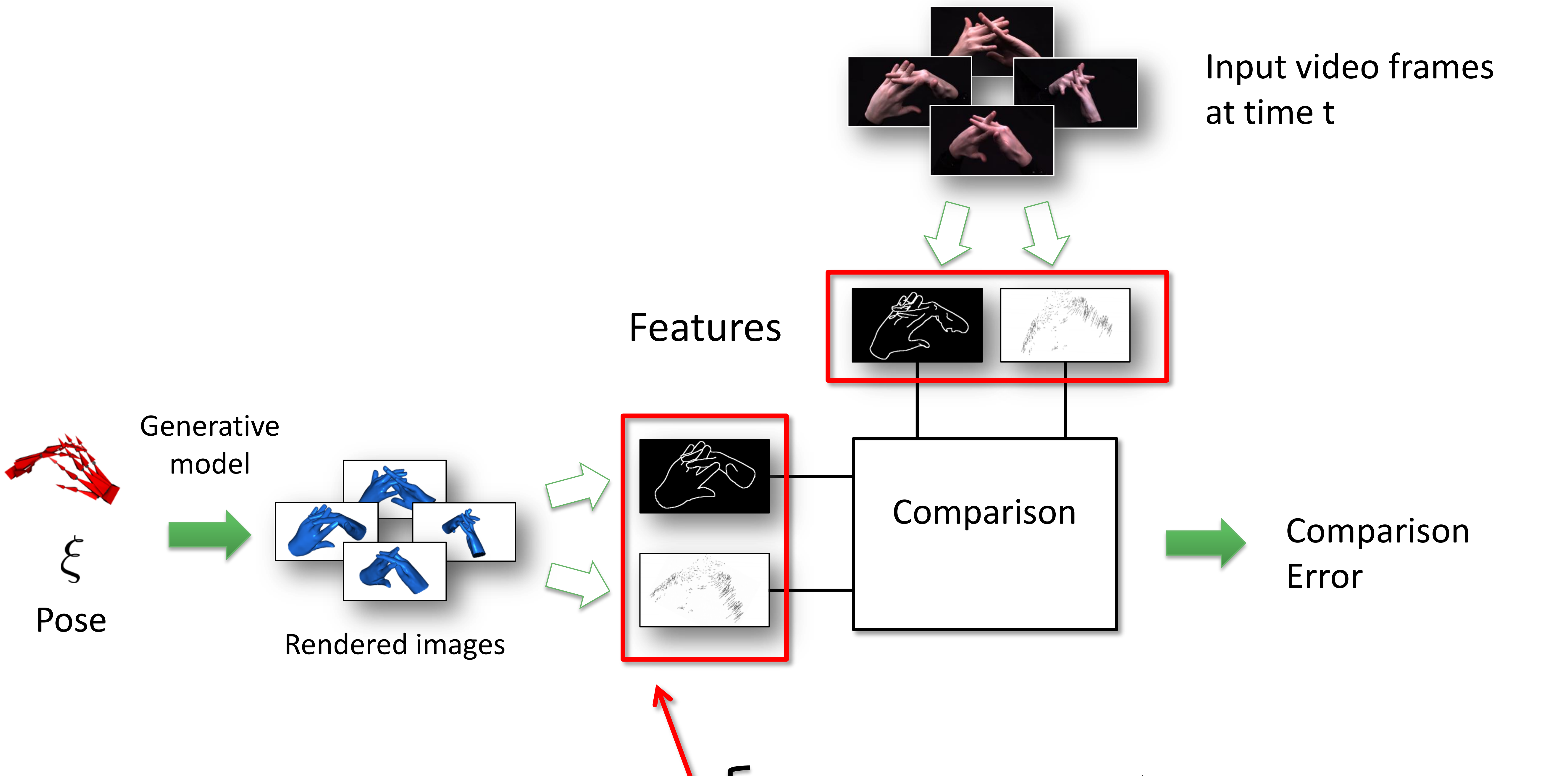


[Baran et al '07]

Pose Estimation

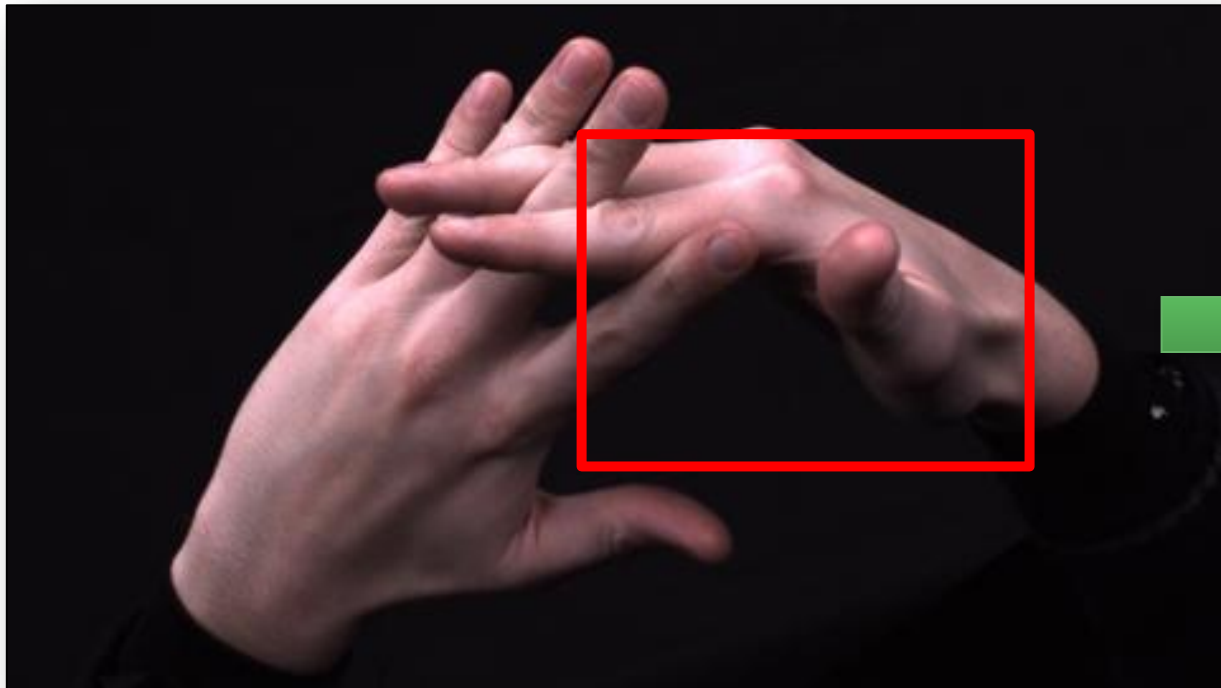


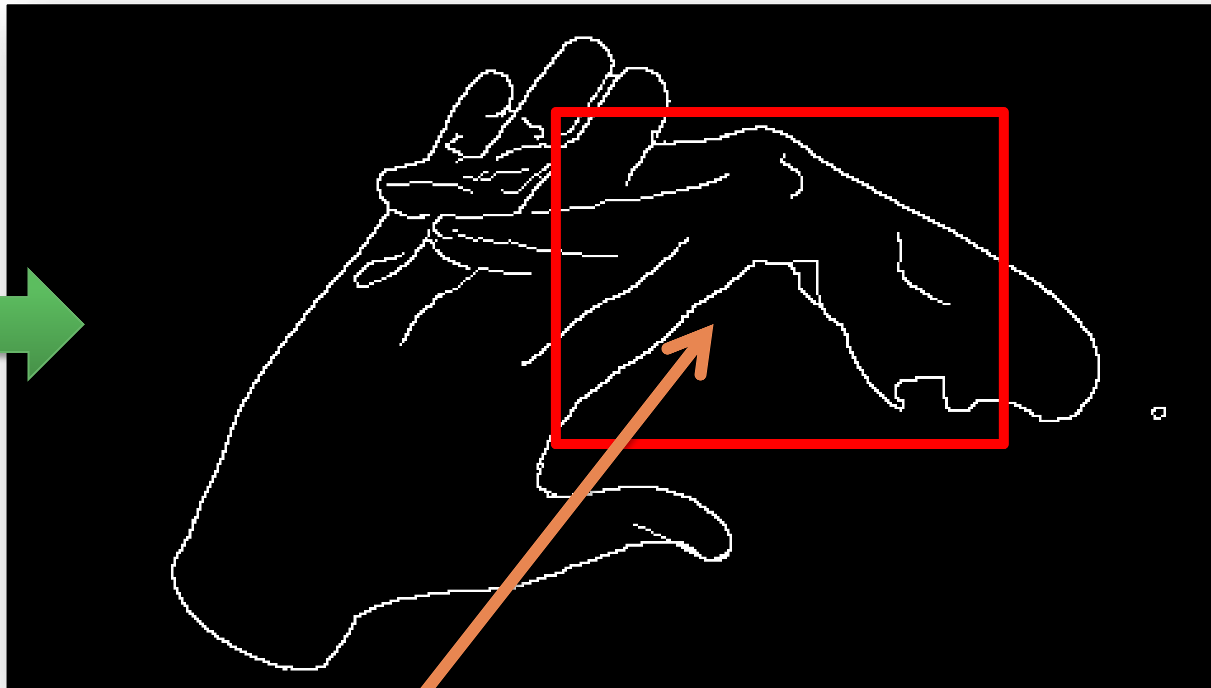
Pose Estimation



- Edges ⇨ Cue on the displacement of the hand
- Optical flow ⇨ Temporal continuity

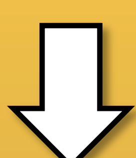
Pose Estimation





NOT sufficient !

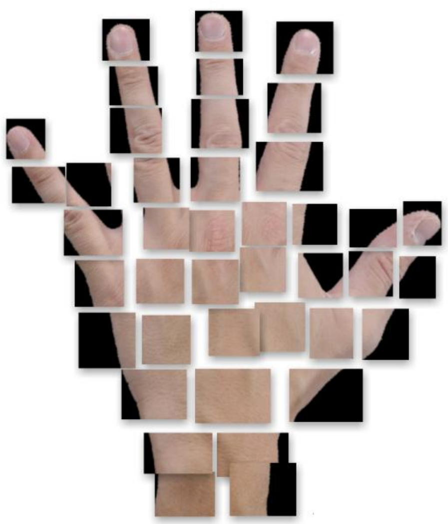
An additional stronger cue needs to be used!



- Edges might disappear due to color similarities
- Optical flow might not be able to compensate

Salient Points

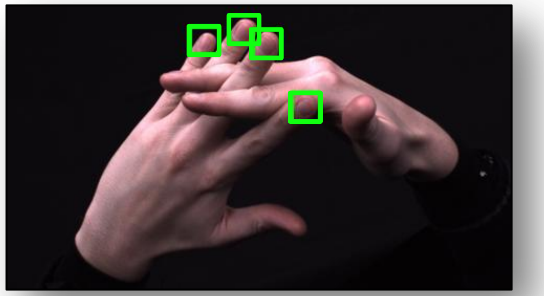
Learn the appearance of
some characteristic features



Classifier

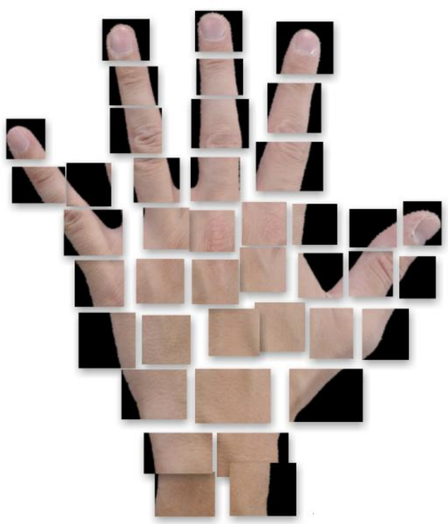
[Gall et al. '11]

Detect them on the videos



Salient Points

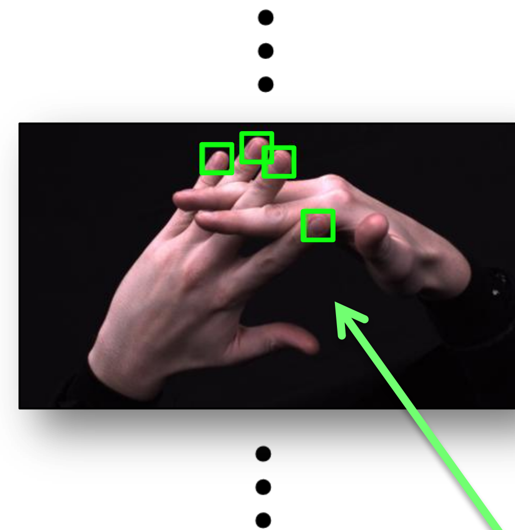
Learn the appearance of some characteristic features



Classifier

[Gall et al. ‘11]

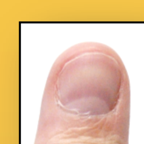
Detect them on the videos



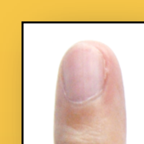
Finger nails

Self-similarities

⇒



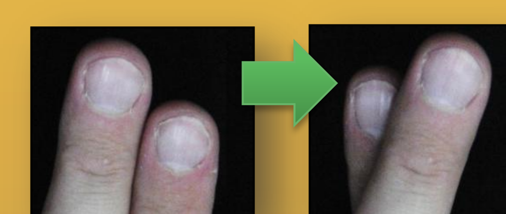
=



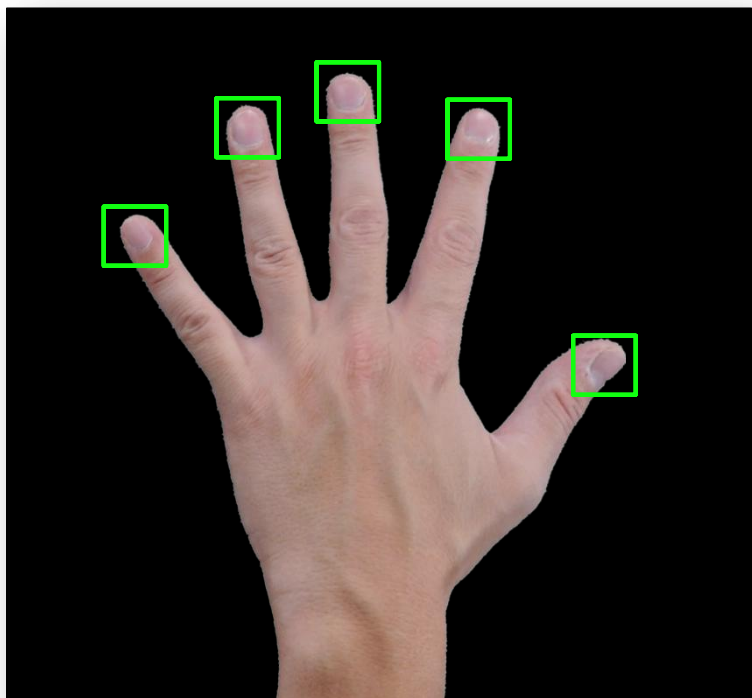
Thumb Nail

Pinky Nail

- Cannot discriminate between nails of different fingers
- Tracking does not help due to the frequent occlusions

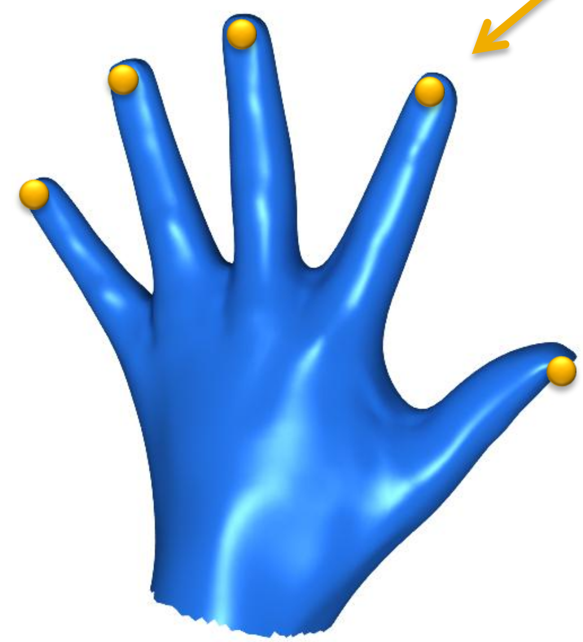


Salient Points



Video frame

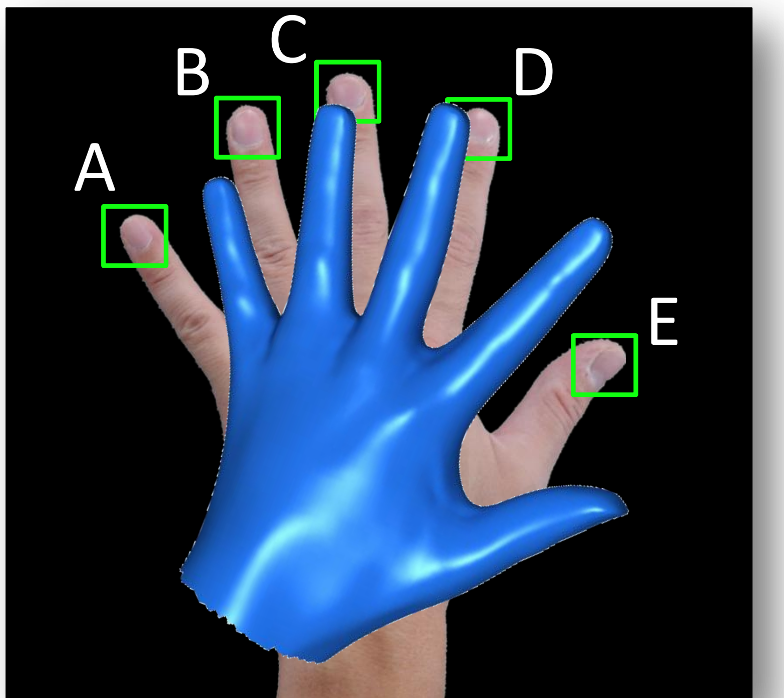
1-to-1 mapping



Vertices corresponding to the finger nails

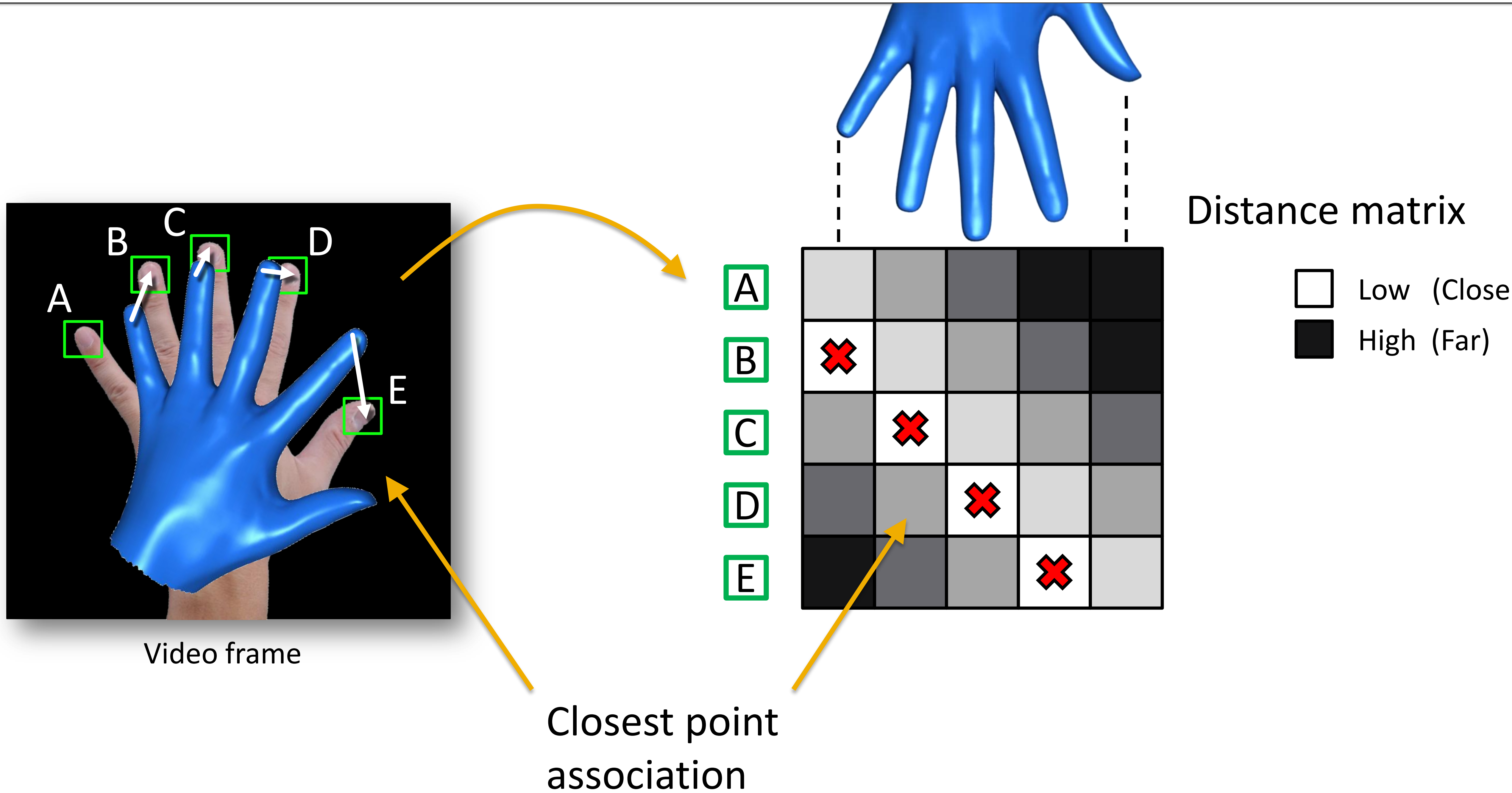
Hand model

Finding the Mapping

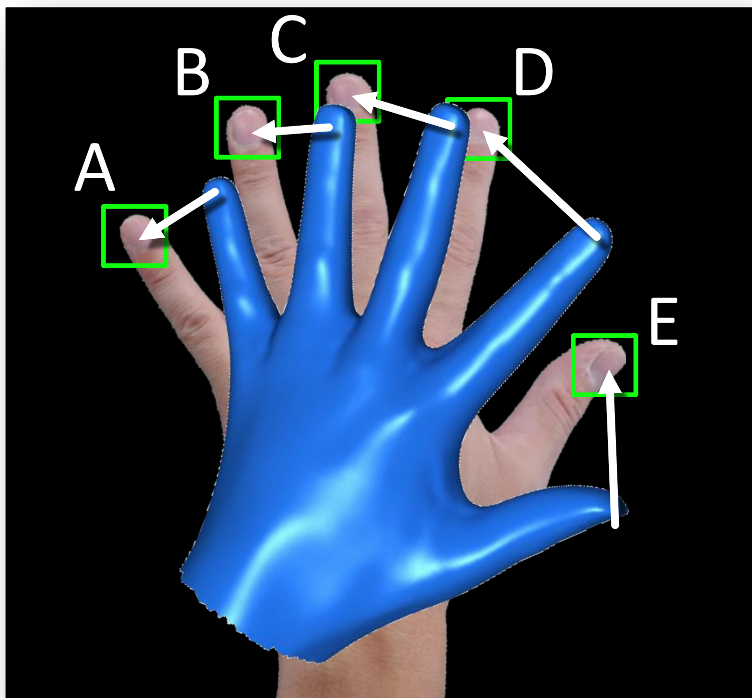


Video frame

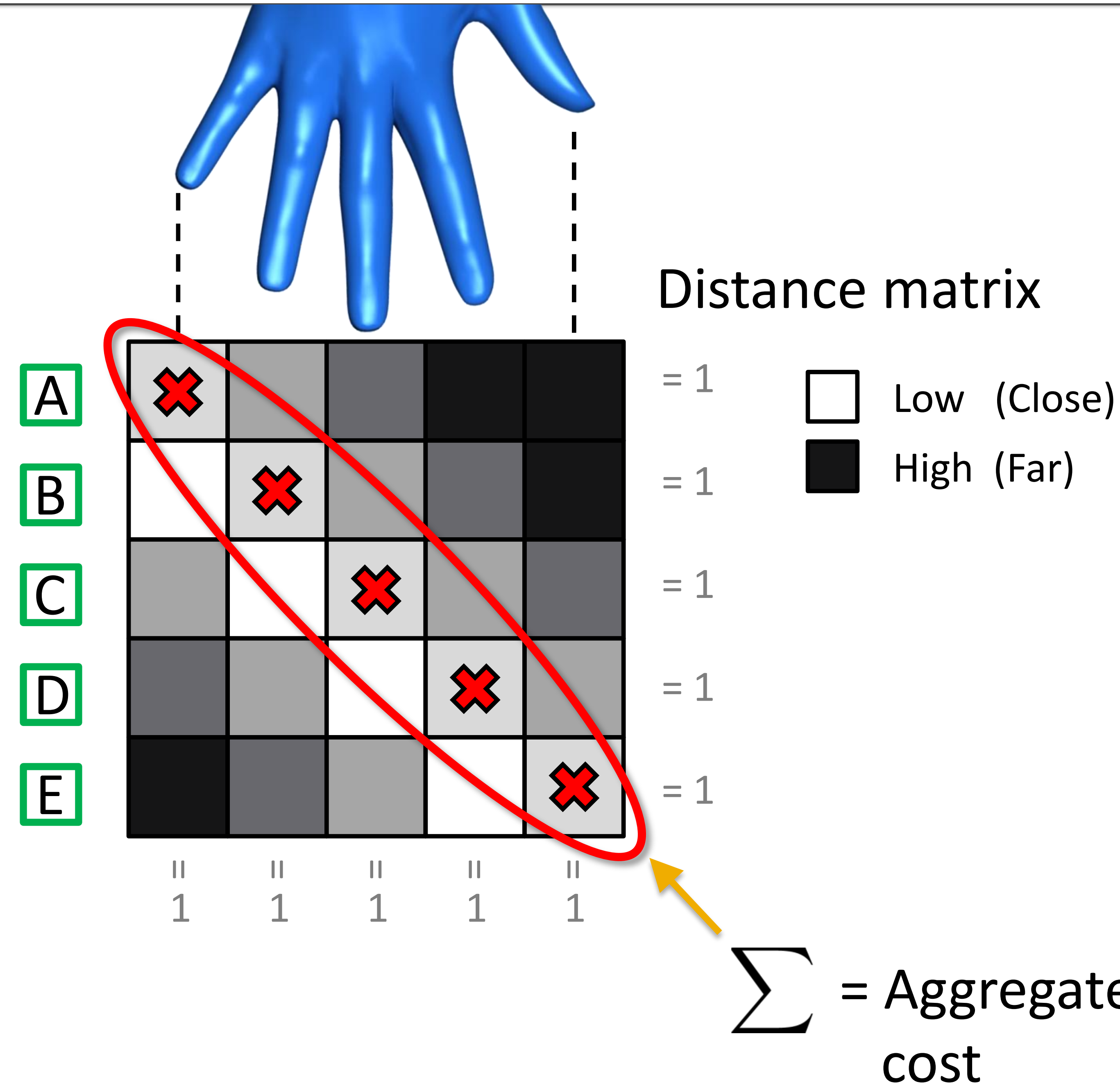
Finding the Mapping



Finding the Mapping



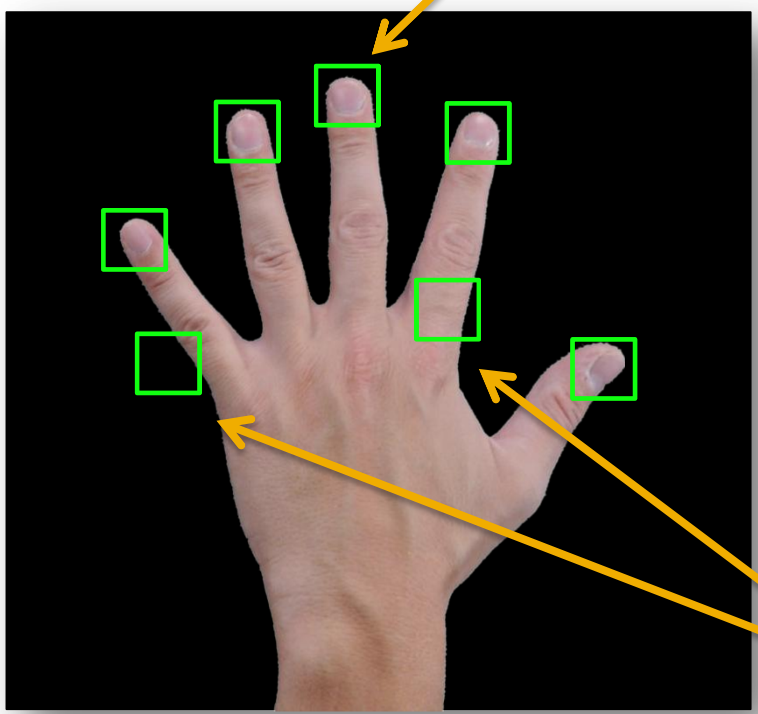
Video frame



Finding the Mapping

Missing detections

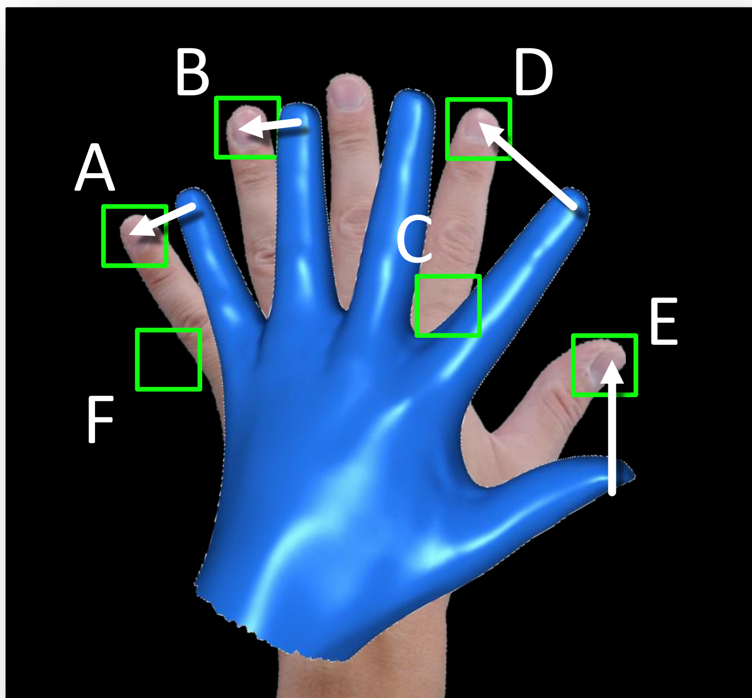
Classifier is not perfect!!



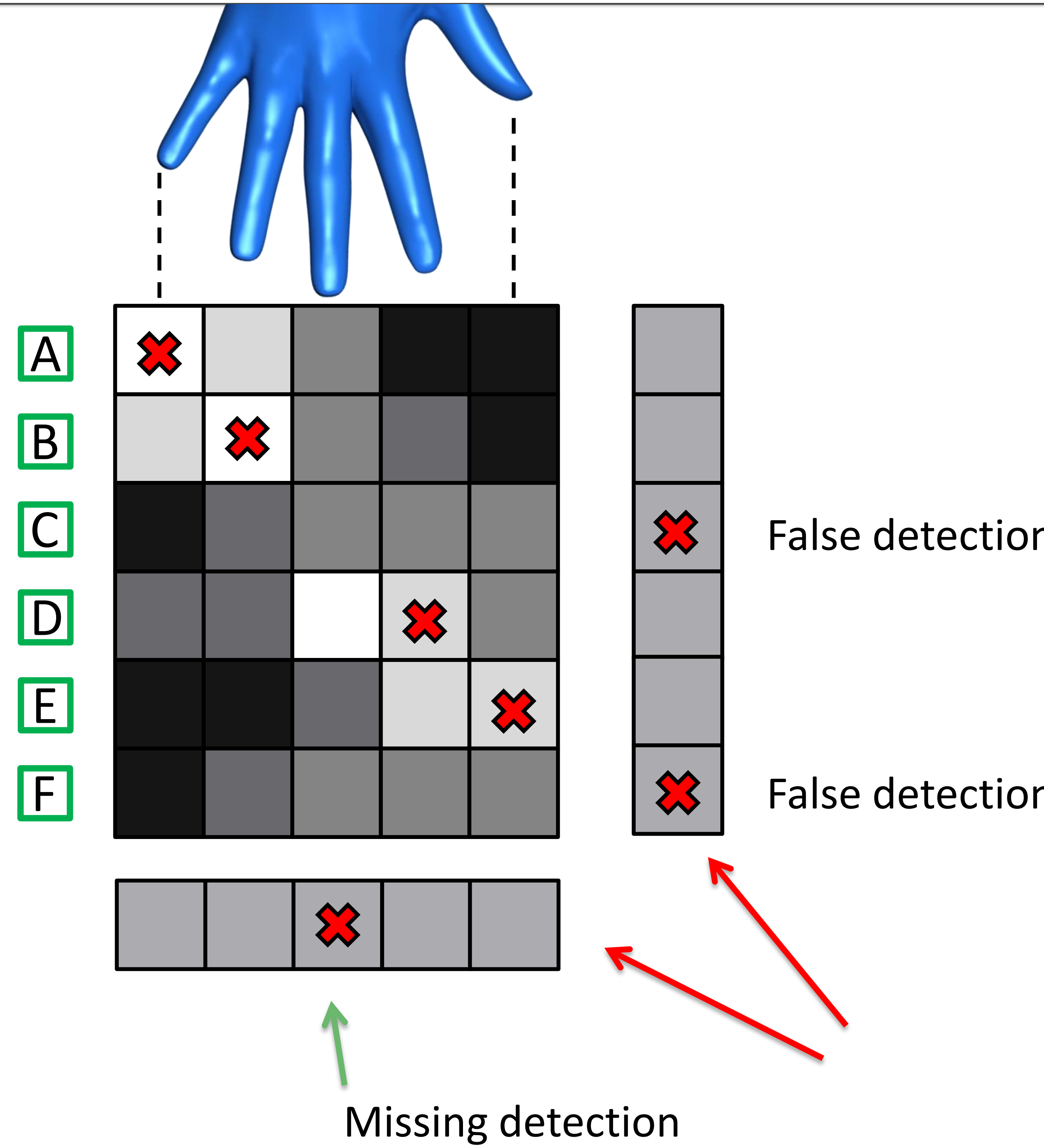
Video frame

False detections

Finding the Mapping



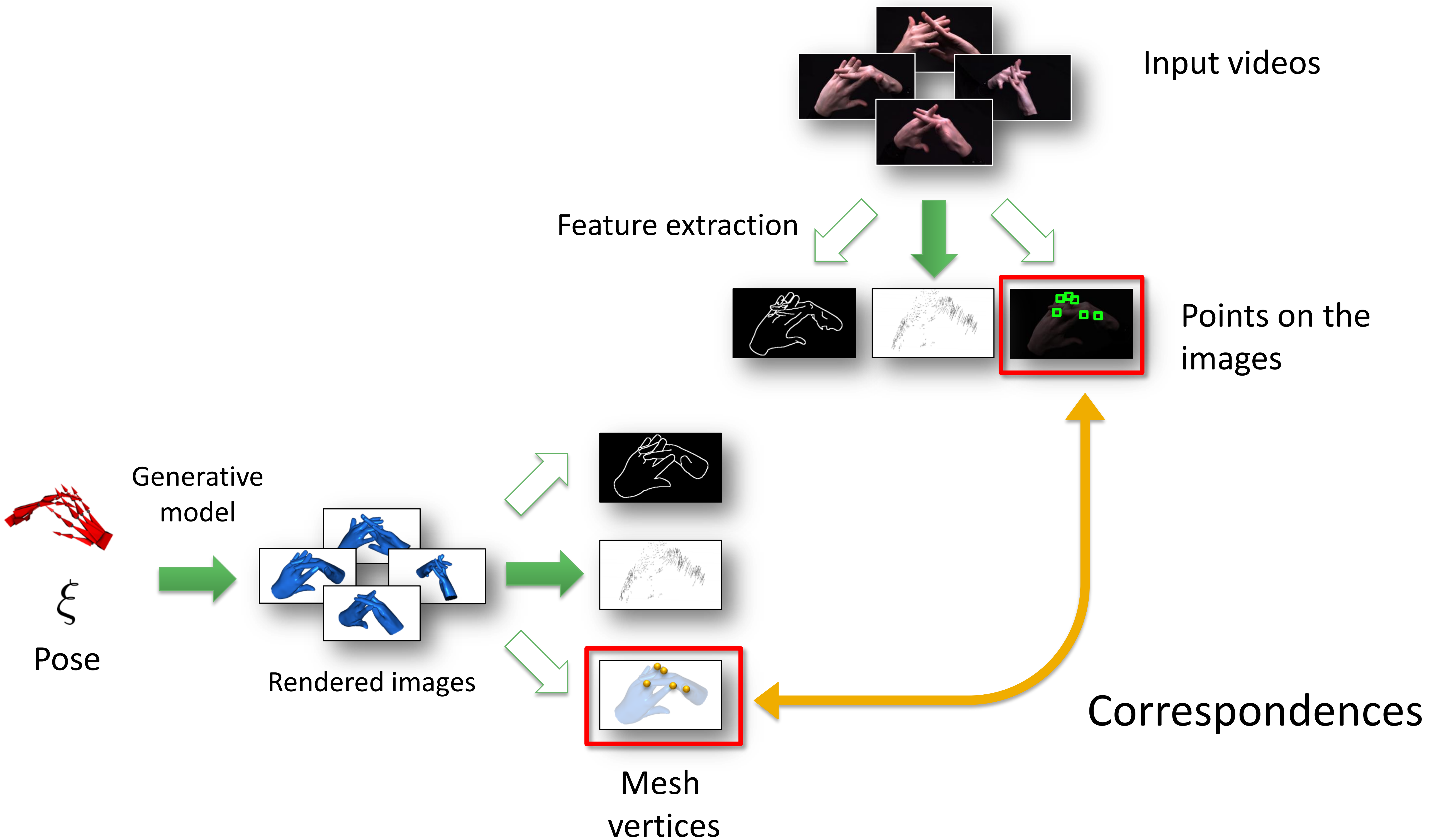
Video frame



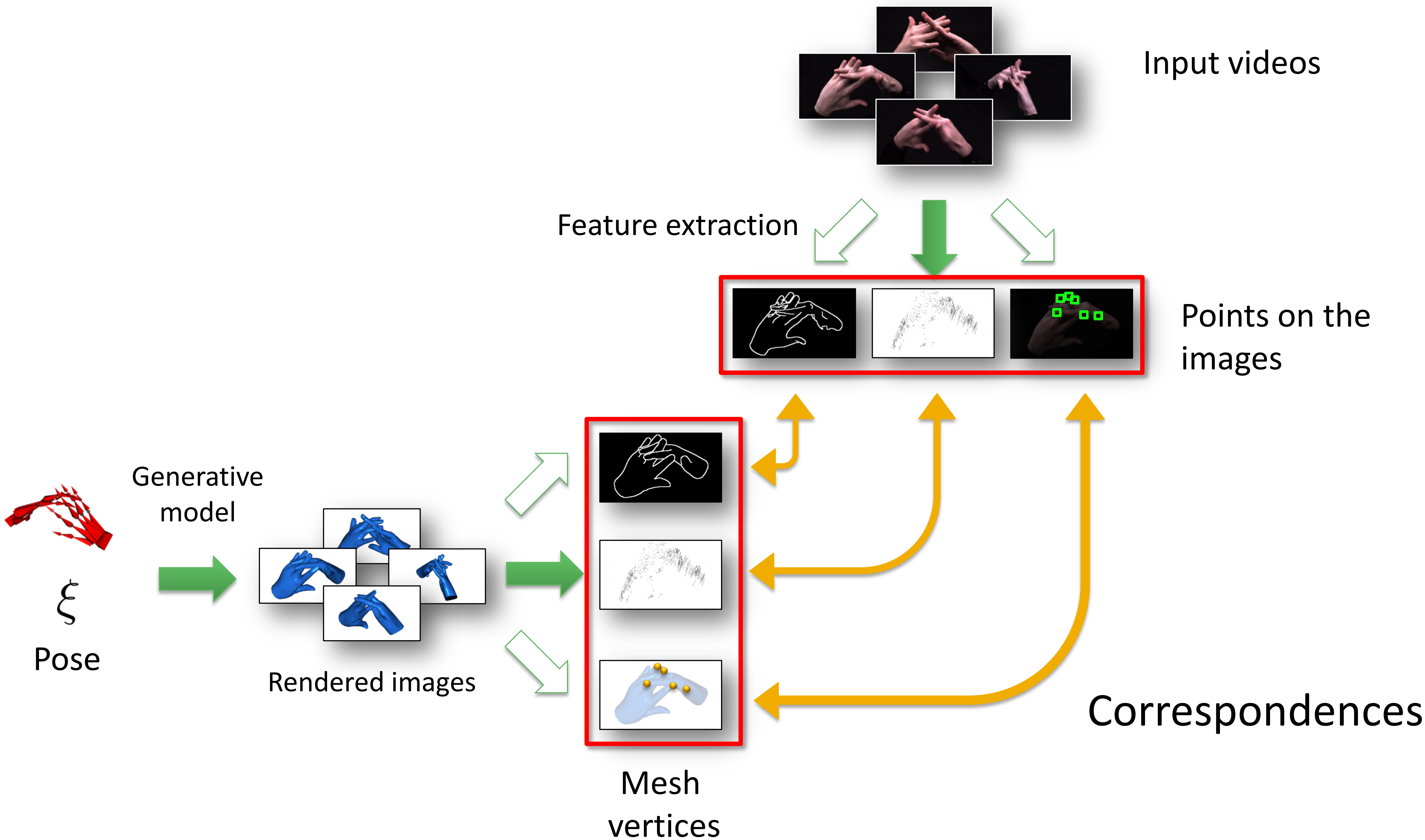
Bipartite Graph Matching Problem
with outliers
(solution in polynomial time)

[Belongie et al. '02]

Pose Estimation



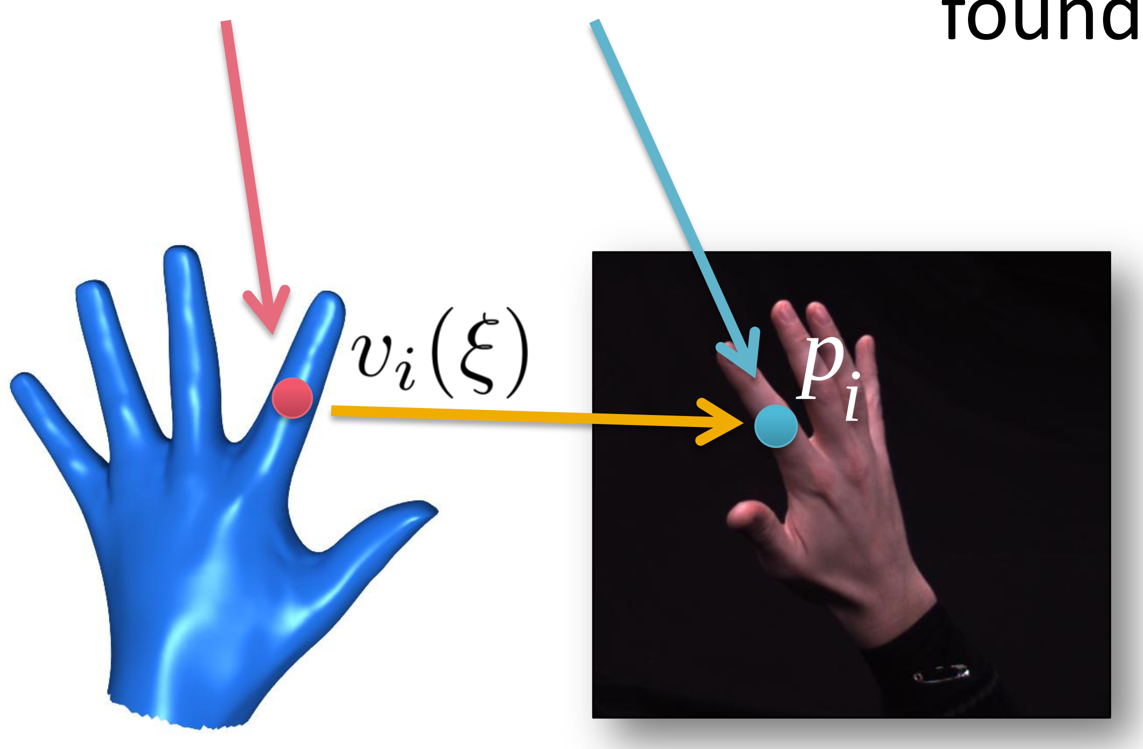
Pose Estimation



Pose Estimation

$$\underset{\xi}{\operatorname{argmin}} \sum \left\| \operatorname{Proj}(v_i(\xi)) - p_i \right\|^2$$

Reprojection error of the founded correspondences



- Non-Linear Least Square
- Differentiable

Minimization using Levenberg Marquardt

Optimization

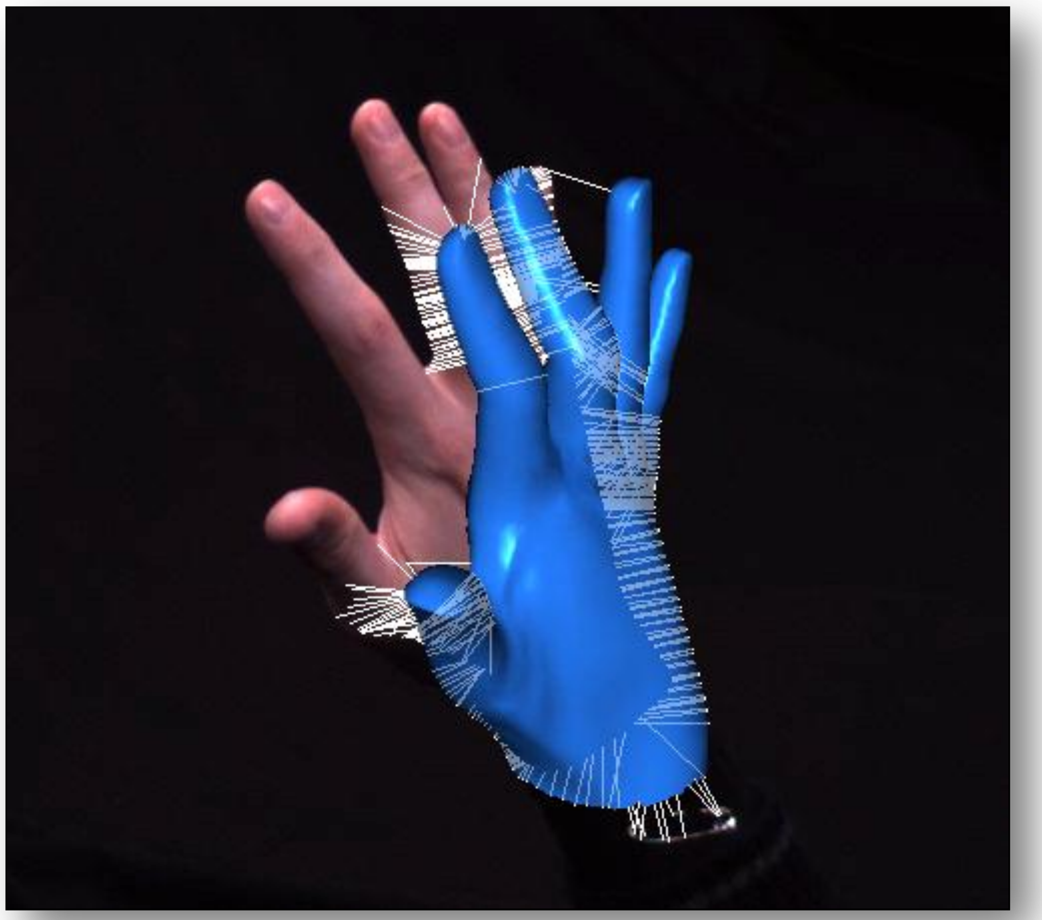
Alternating optimization scheme

- Generate mesh at pose ξ
- Solve for the correspondences
- Solve for the pose

$$\sum \|\mathrm{Proj}(v_i(\xi)) - p_i\|^2$$

Optimization

Alternating optimization scheme



- Generate mesh at pose ξ
- Solve for the correspondences
- Solve for the pose

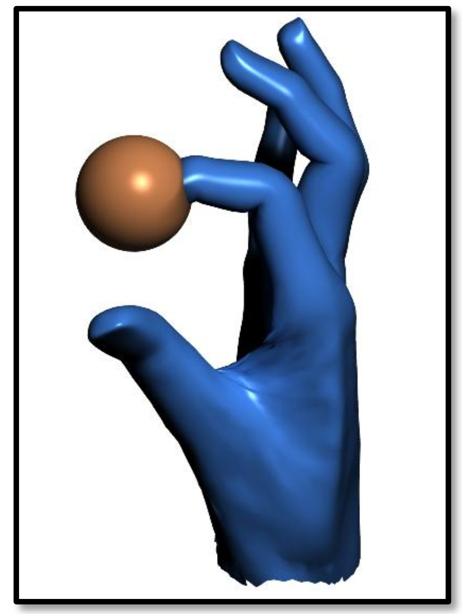
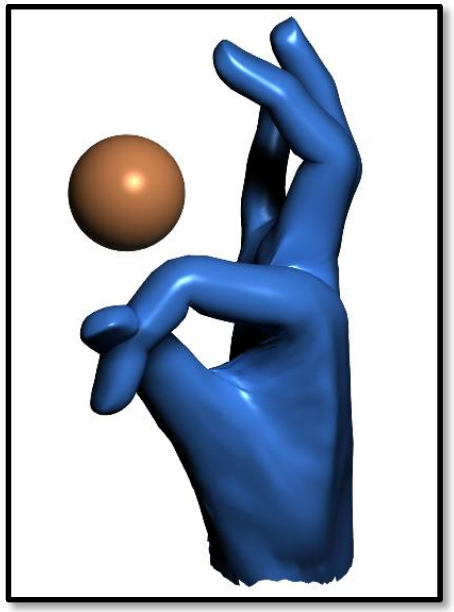
$$\sum \|\mathrm{Proj}(v_i(\xi)) - p_i\|^2$$

Local minima → Re-initialization using simulated annealing

(needed twice in all our experiments)

Collisions and Self-Intersections

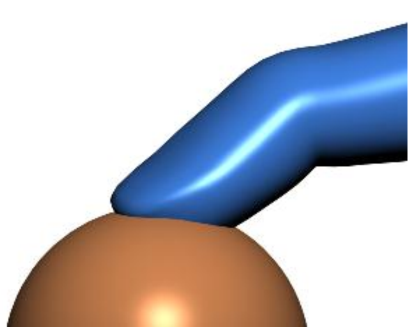
- Lack of information
- Overfitting

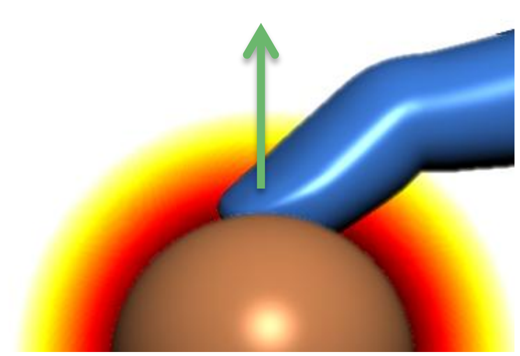
Collisions

Self-Intersections

$$\operatorname*{argmin}_{\xi} \sum \left\|\mathrm{Proj}(v_i(\xi)) - p_i\right\|^2 \boxed{+ \ \Gamma(\xi)}$$



Colliding faces



Local Distance Field

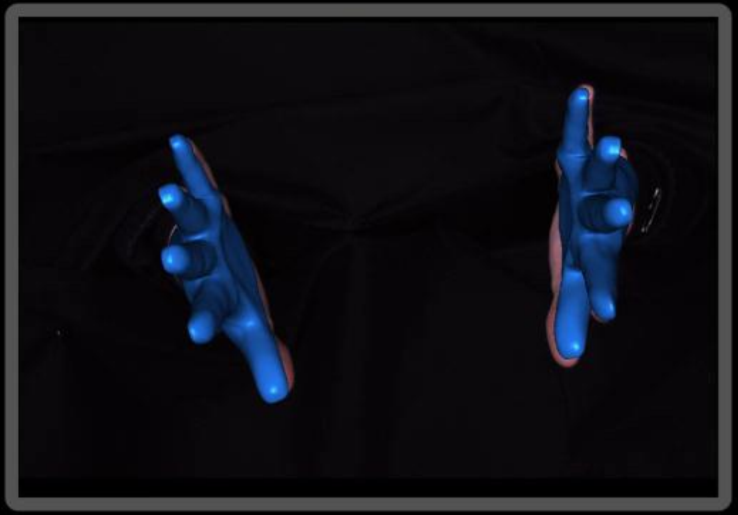
Results


Cam #4
(Input video)
Cam #4
(Result)
Cam #6
(Input video)
Cam #6
(Result)
Cam #7
(Input video)
Cam #7
(Result)
Cam #7
(Input video)
Cam #7
(Result)
Cam #7
(Input video)
Cam #7
(Result)
Cam #4
(Input video)
Cam #4
(Result)

Finger Tips Touching

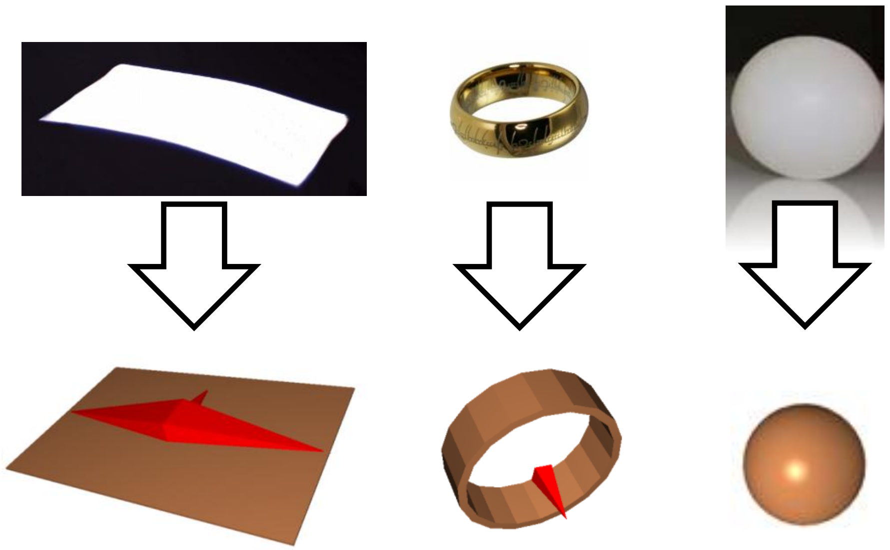
Cam #4
(Input video)

Cam #4
(Result)



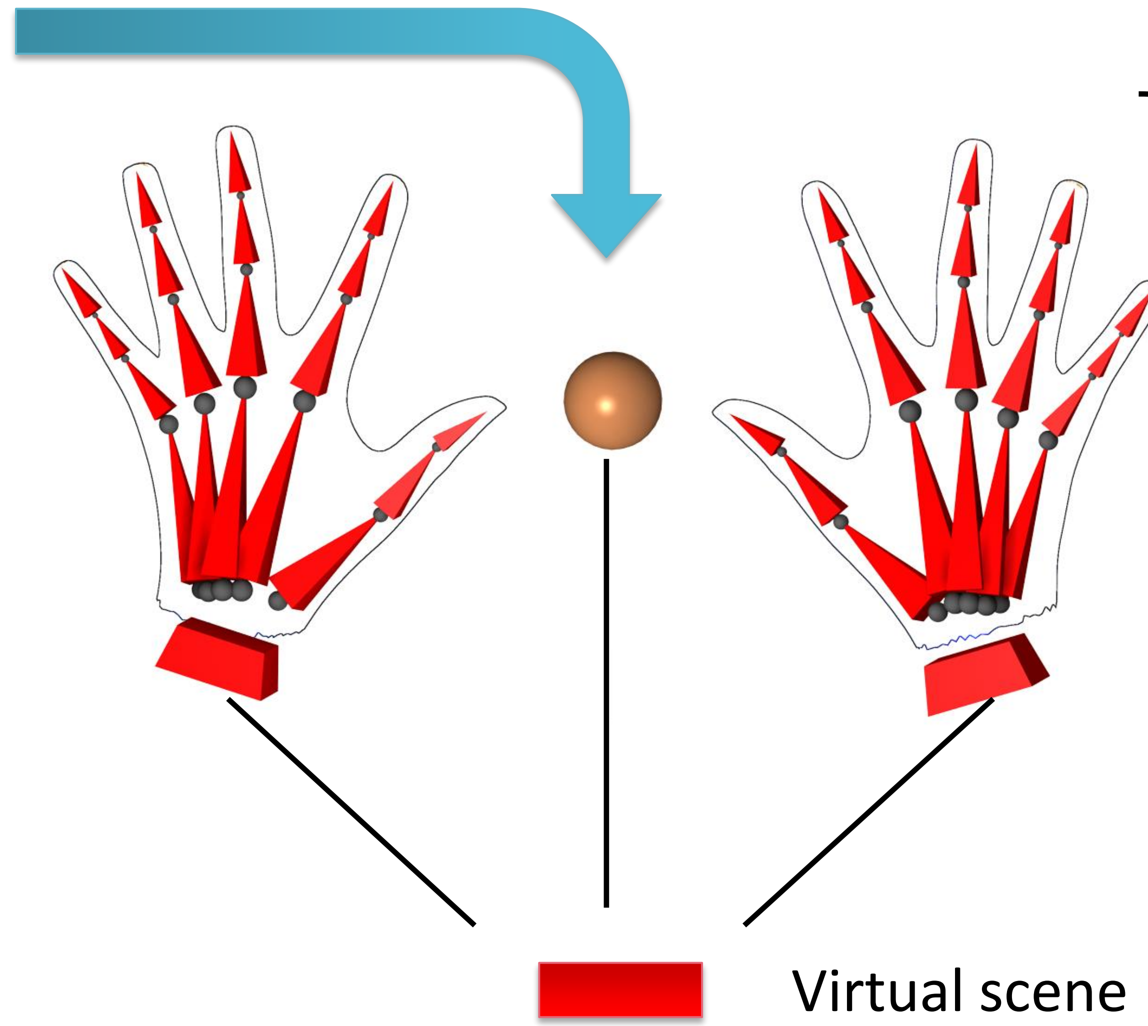
Cam #6
(Result overlaid on Input Video)

How to handle additional Objects



Scan the object

3D models



Virtual scene bone

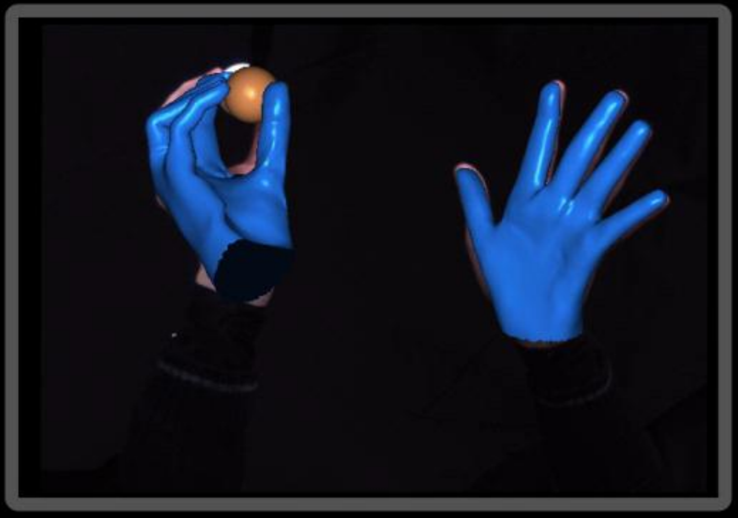
Unique object

Use the algorithm as it is!!

- Collisions
- Occlusions

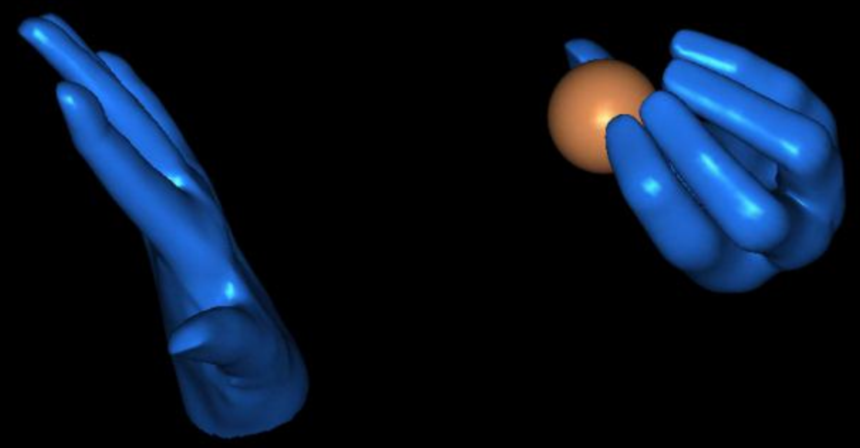
Handled transparently

Holding and Passing a Ball



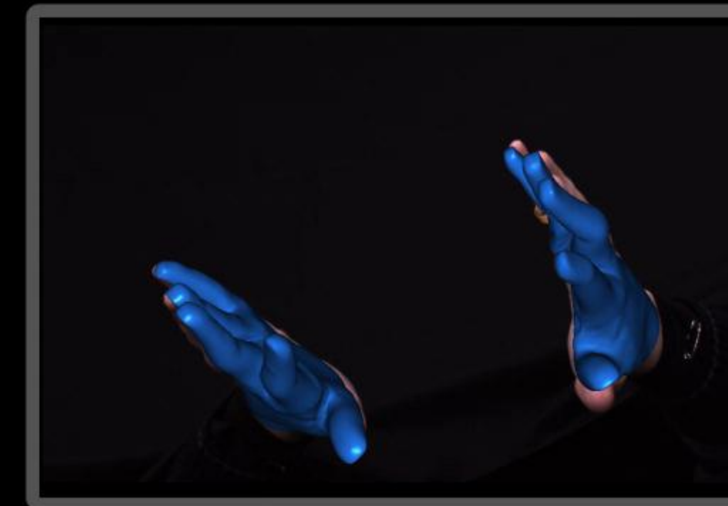
Cam #3
(Result overlaid on Input Video)

Cam #5
(Input video)

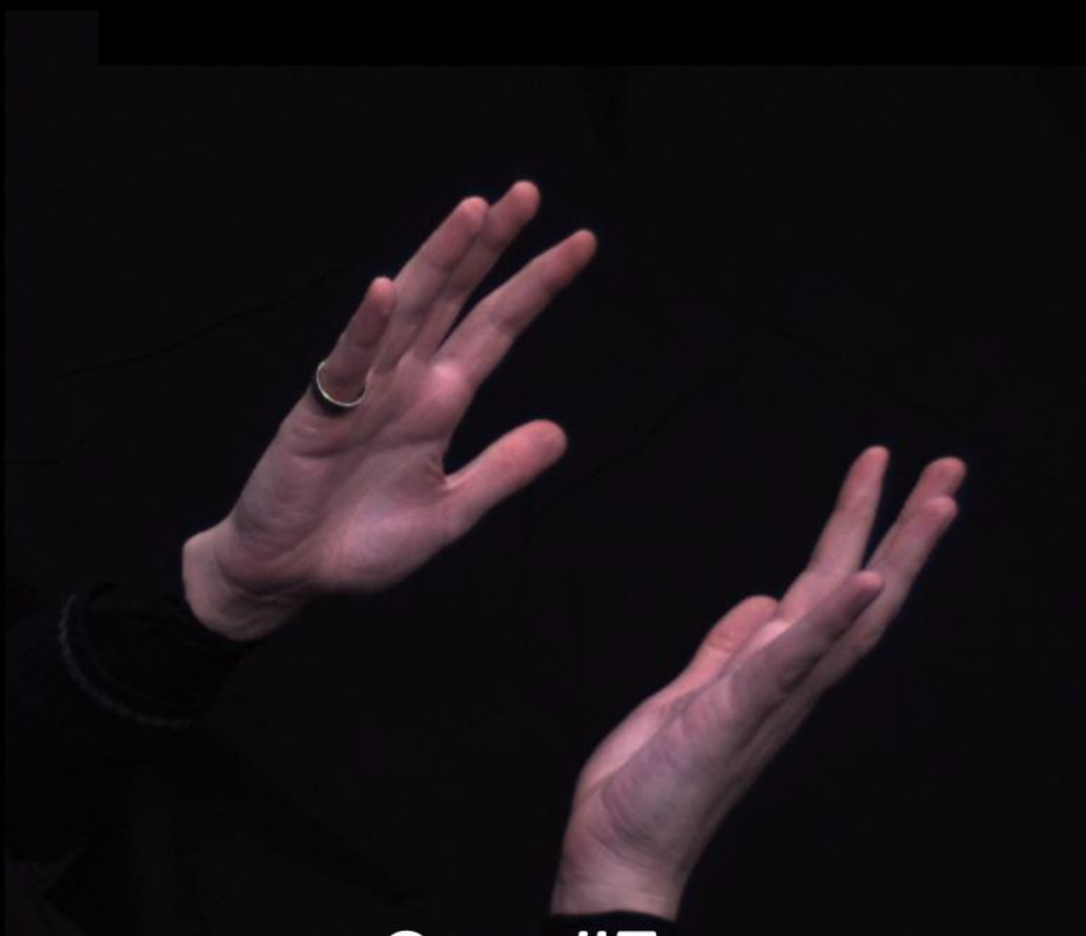


Cam #5
(Result)

Taking off a Ring



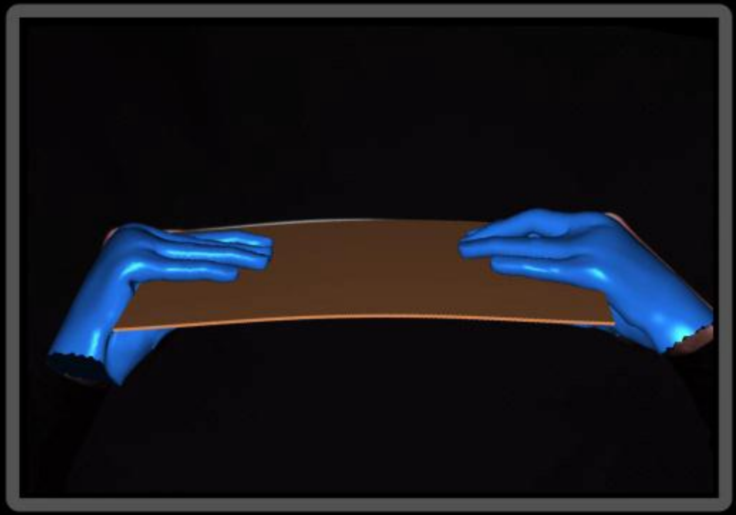
Cam #5
(Result overlaid on Input Video)



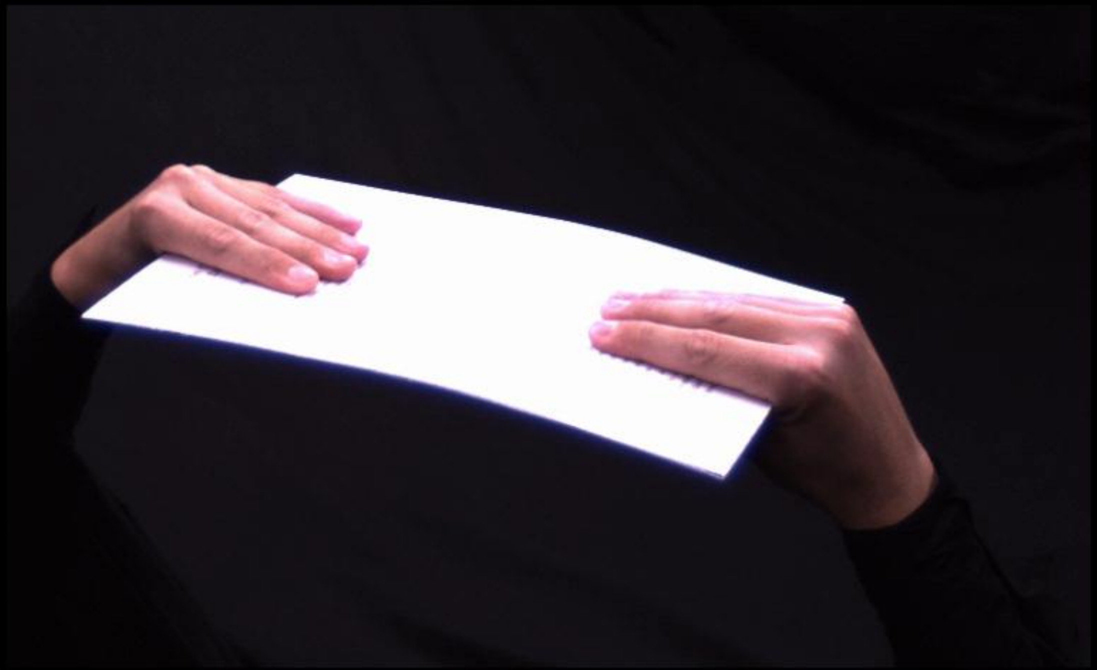
Cam #7
(Input video)

Cam #7
(Result)

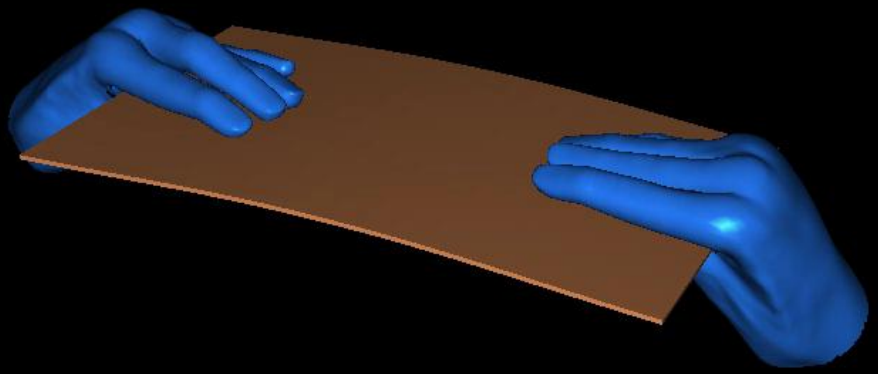
Paper Folding



Cam #7
(Result overlaid on Input Video)



Cam #8
(Input video)



Cam #8
(Result)

Conclusions

- We proposed a method to estimate the articulated motion of hands interacting with objects
  - many DOF (up to 78)
occlusions → usage of multiple cues
(edges, optical flow, **salient points**)
  - collisions
self-intersections → **Distance fields**
  - self-similarities → Solve the association problem as a **Bipartite Graph Matching** problem

Quantitative evaluation:

	Joints position error $\|\cdot\|_1$
[Our Approach]	1.5mm
[Oikonomidis et al. '11]	4.7mm

3x more accurate than the state of the art

Luca Ballan *Aparna Taneja* *Jürgen Gall* *Luc Van Gool* *Marc Pollefeys*

Thank you!

Datasets and 3D models available at
http://cvg.ethz.ch/research/ih-mocap/

Supported by erc & FNSNF Microsoft