15-869
Lecture 2
Virtualizing Reality

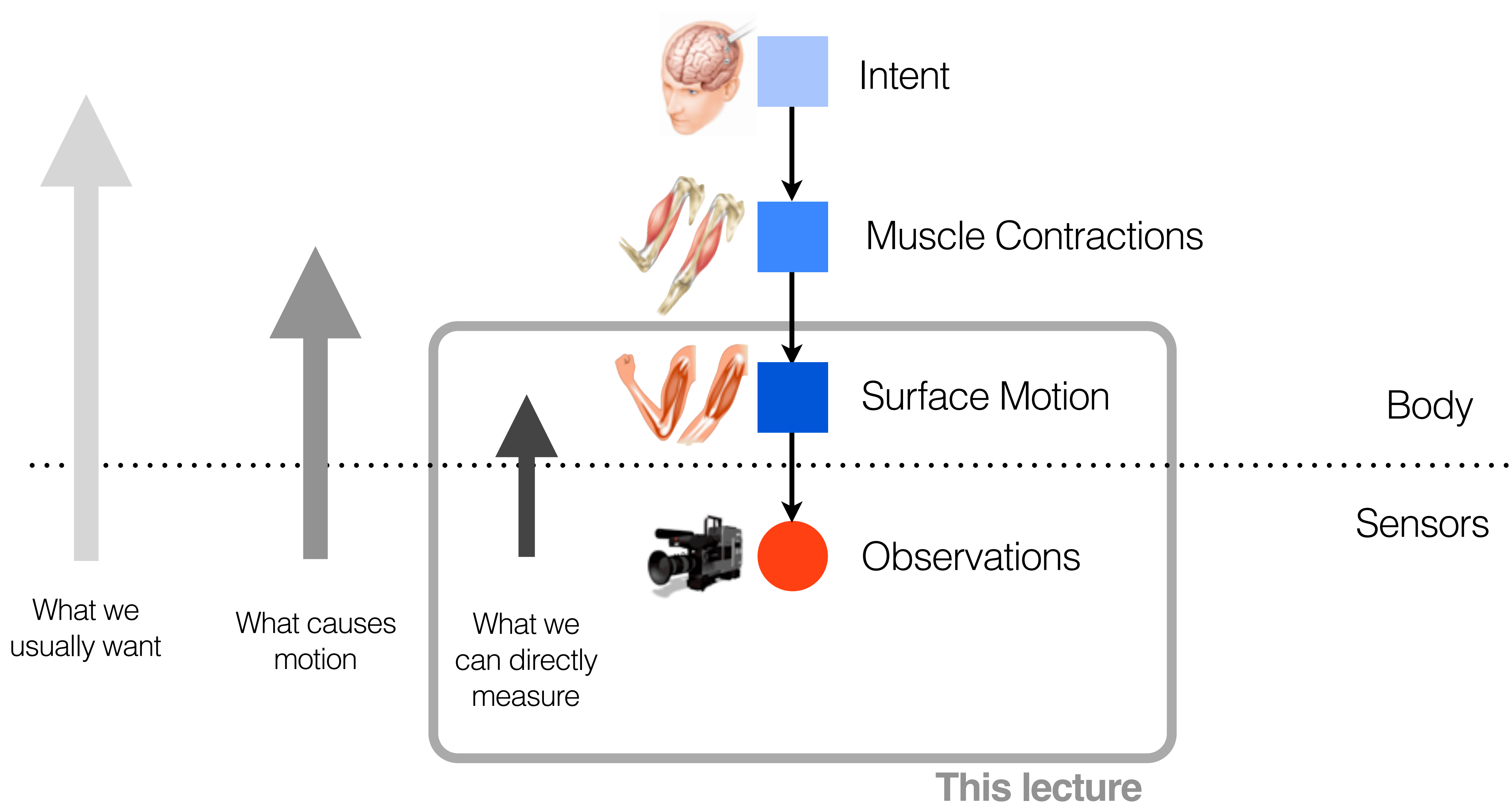
Yaser Sheikh

Human Motion Modeling and Analysis

Fall 2012

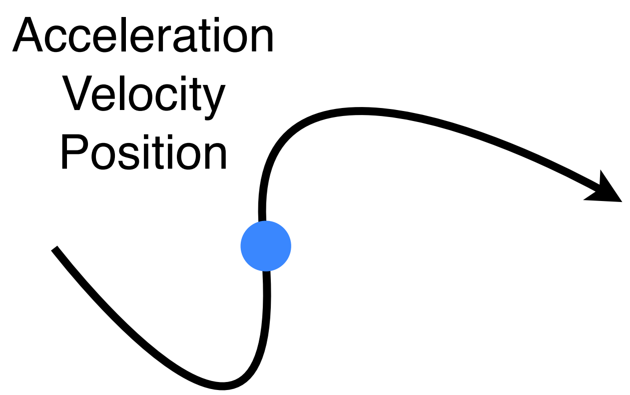
What is Human Motion?

What makes Human Motion Hard to Analyze?

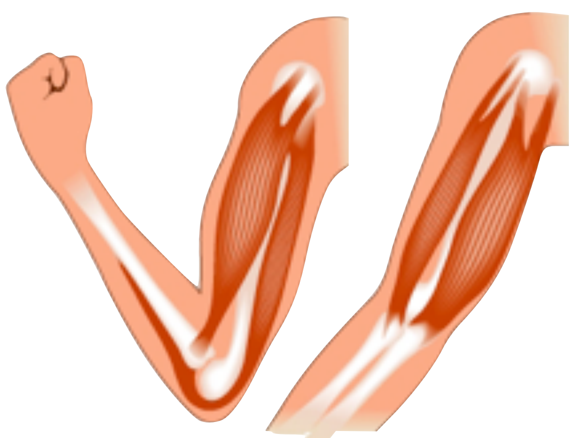


It's impossible to kiss your elbow

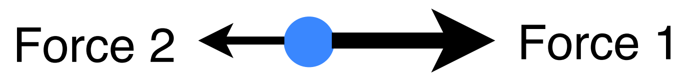
Kinematics vs Dynamics



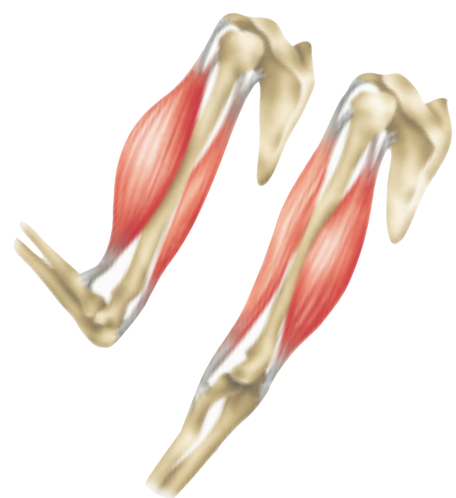
Kinematics: Geometry of Motion
(Motion without Cause)



This lecture



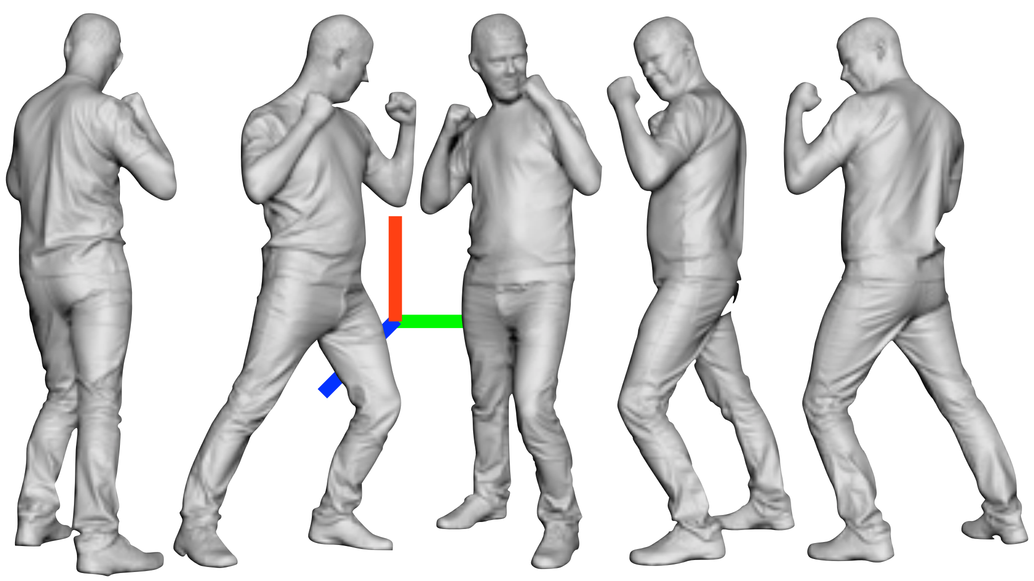
Dynamics: Physics of Motion
(Motion with Cause)



Capturing Human Motion

Holy Grail: Single Video Camera

Cameras are ubiquitous, cheap, and passive



3D Structure

3D Motion

Illustration from IR

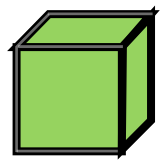
This Lecture...

3D Dynamic Surface Reconstruction using Passive Sensing

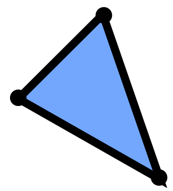
- How should we represent human body surfaces?
- What can we extract from images?
- A Brief History of Virtualizing Reality
- Volumetric and Point-based 3D Reconstruction Algorithms
- Tour of the Virtualizing Studio 4.0

How do we Represent the Body Surface?

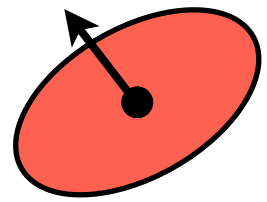
Representation Primitives



Voxel



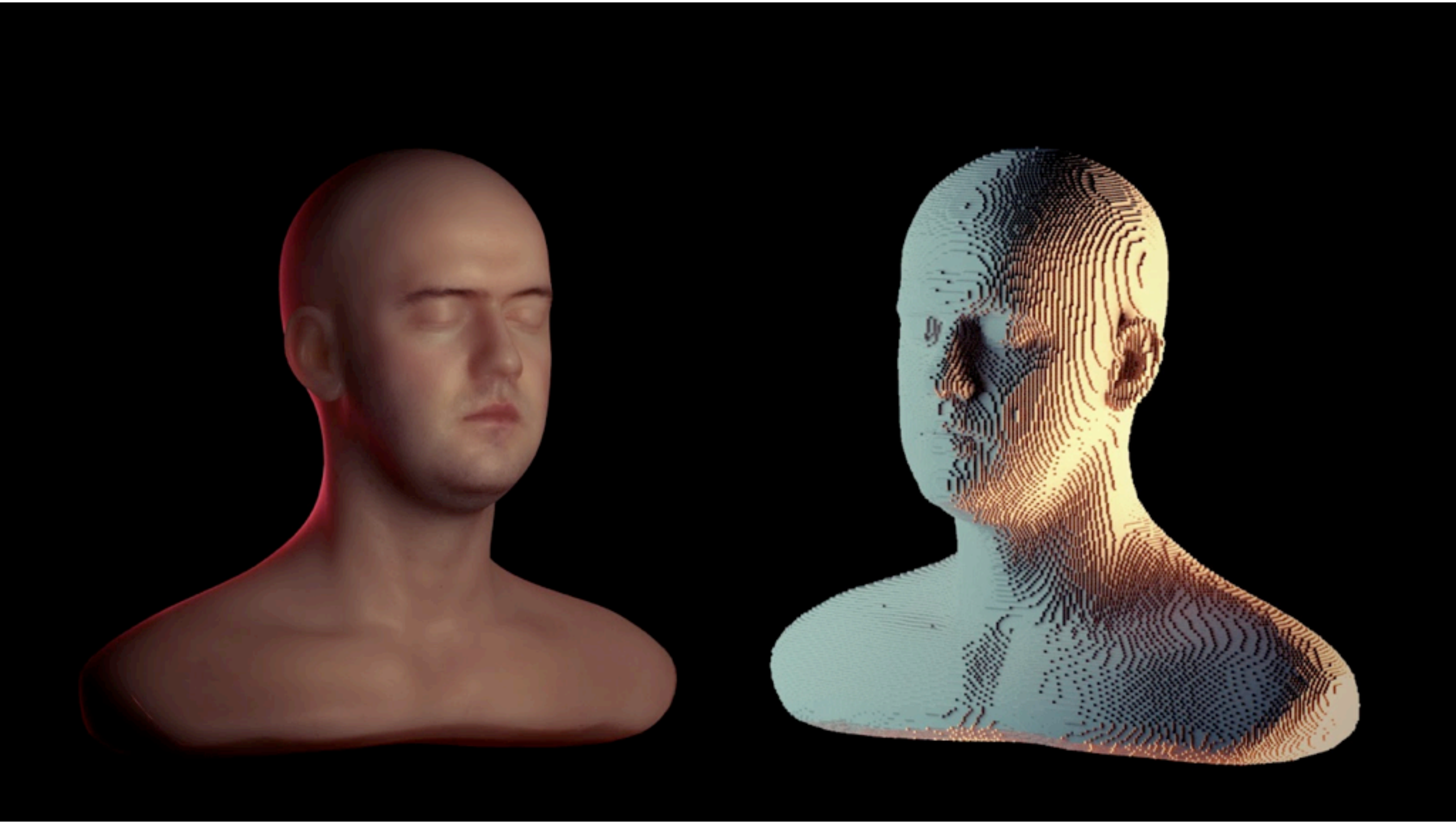
Mesh

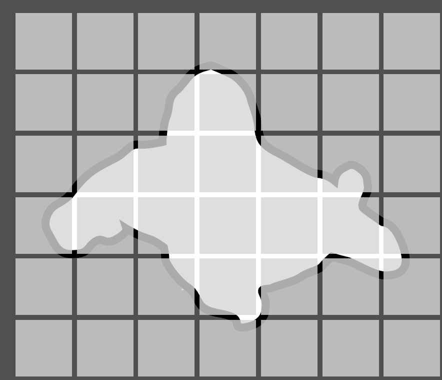


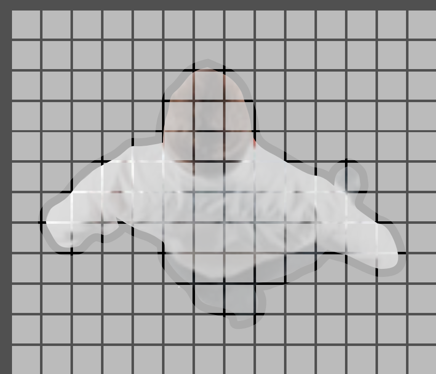
Surfel

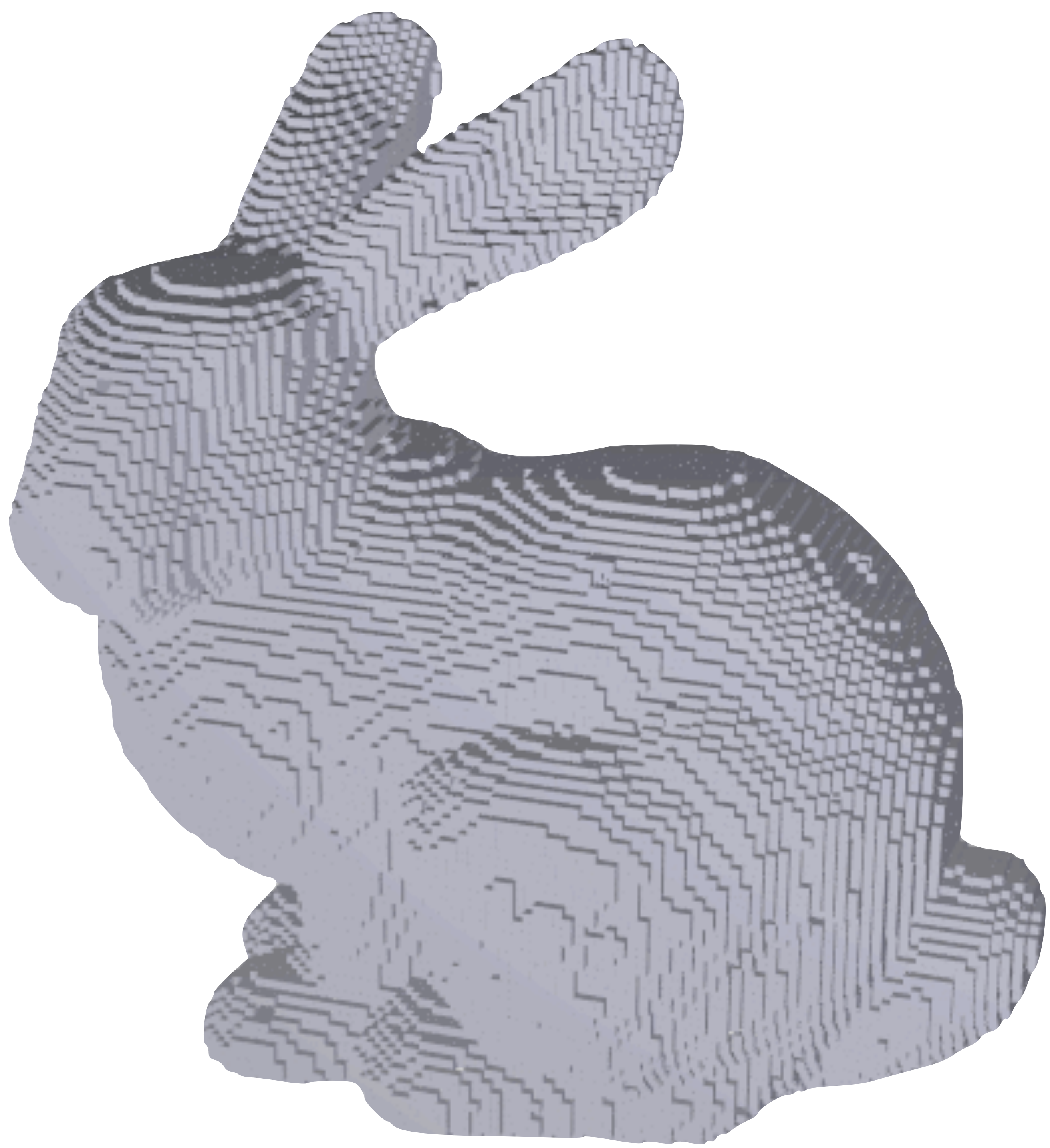
Voxels

Volumetric Picture Element









Voxels

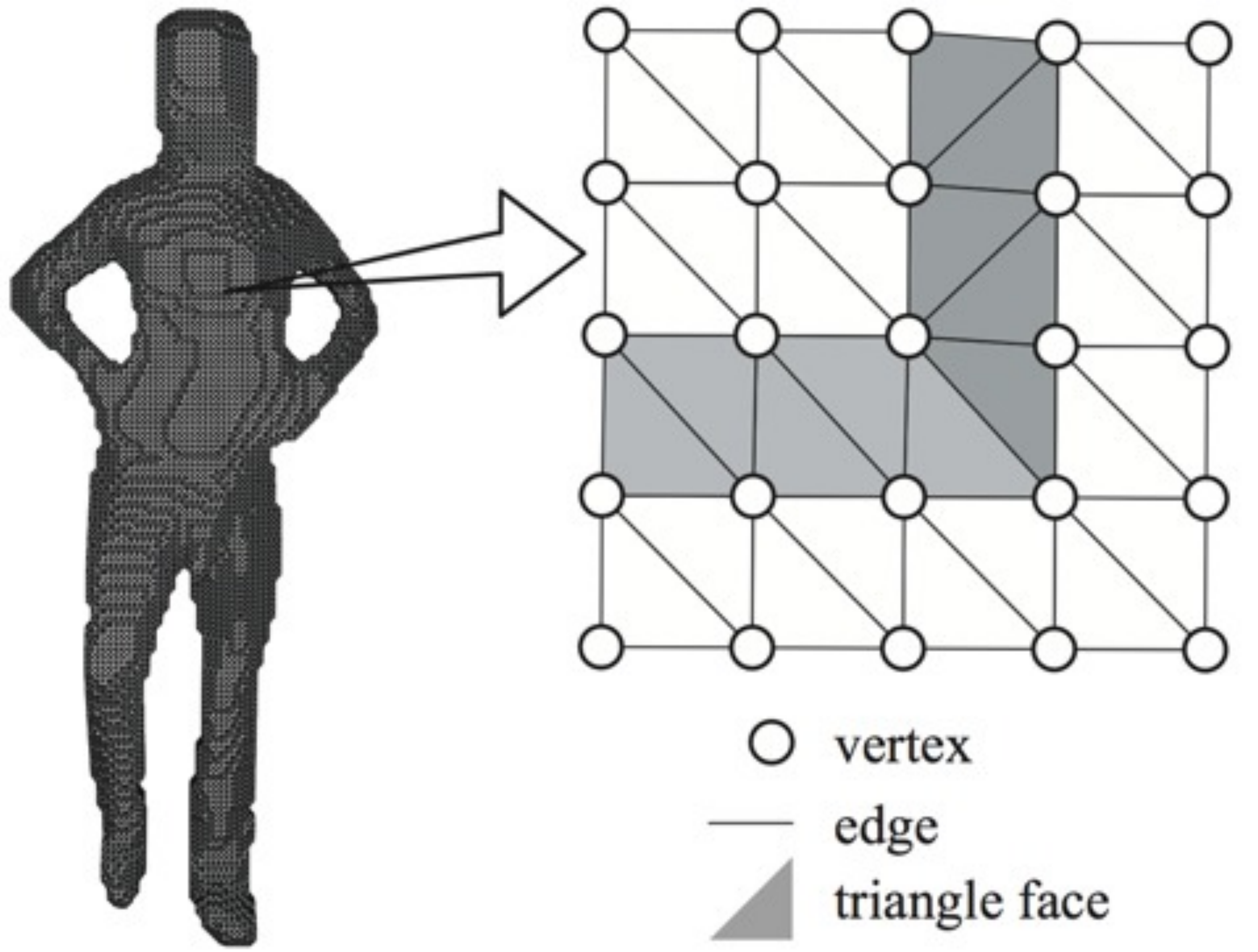
Volumetric Picture Elements

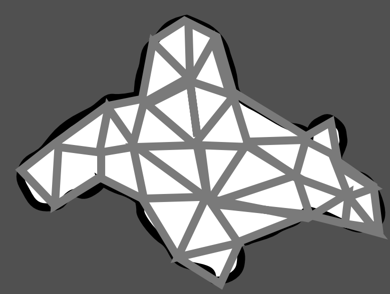
- Dynamic Voxels (doxels): Spacetime grid (e.g., 100 cm x 100 cm x 100 cm x 100 sec).
- Memory intensive (if used trivially)
- **Example**: 1 minute capture at 30 frames per second of 10 meter cubed space at centimeter resolution

$$\underbrace{60}_{\text{seconds}} \times \underbrace{30}_{\text{frames per second}} \times (\underbrace{100}_{\text{centimeters per meter}} \times \underbrace{10}_{\text{meters}})^3 = \underbrace{1,800,000,000,000}_{\text{number of voxels}}$$

Mesh

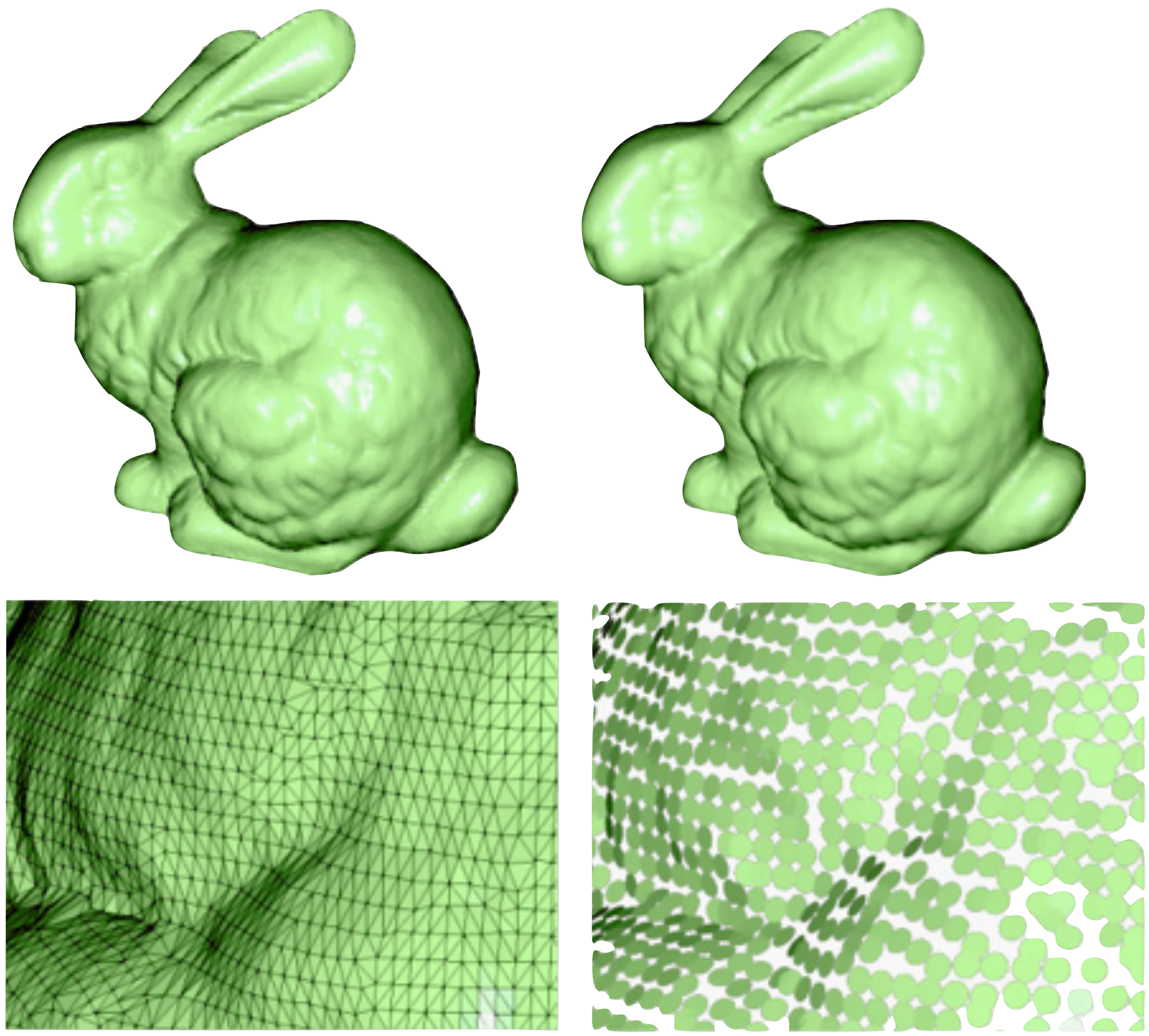
- Continuity constraint embedding
- Limited memory consumption
- Fixed topology



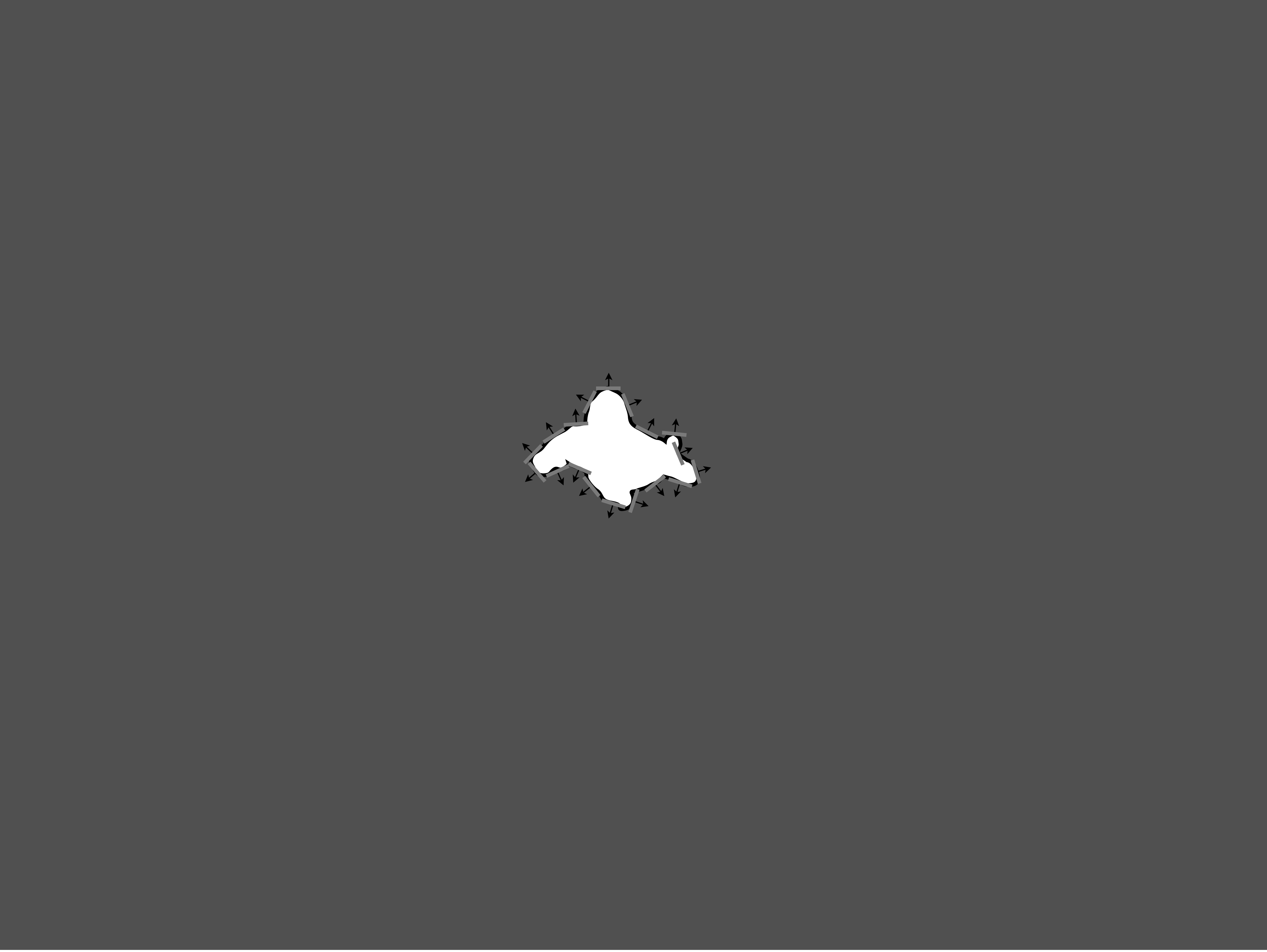


Surfels

Surface Elements



Pfister et al., Surfels: Surface Elements as Rendering Primitives, SIGGRAPH 2000.



Representation

Reconstructing 3D Body Shape and Motion

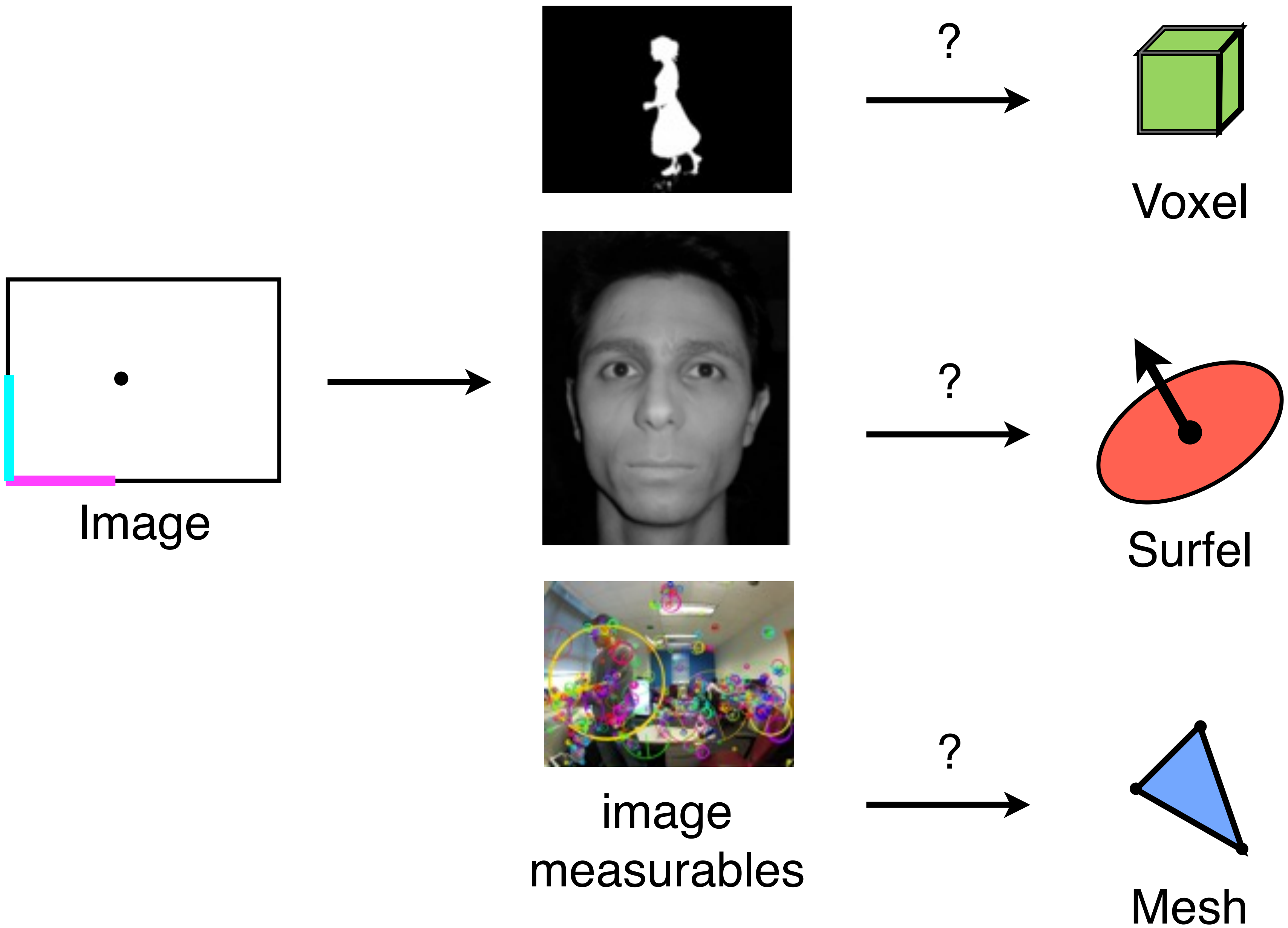


Image Information

Measurables

Silhouettes

Correspondences

Shading

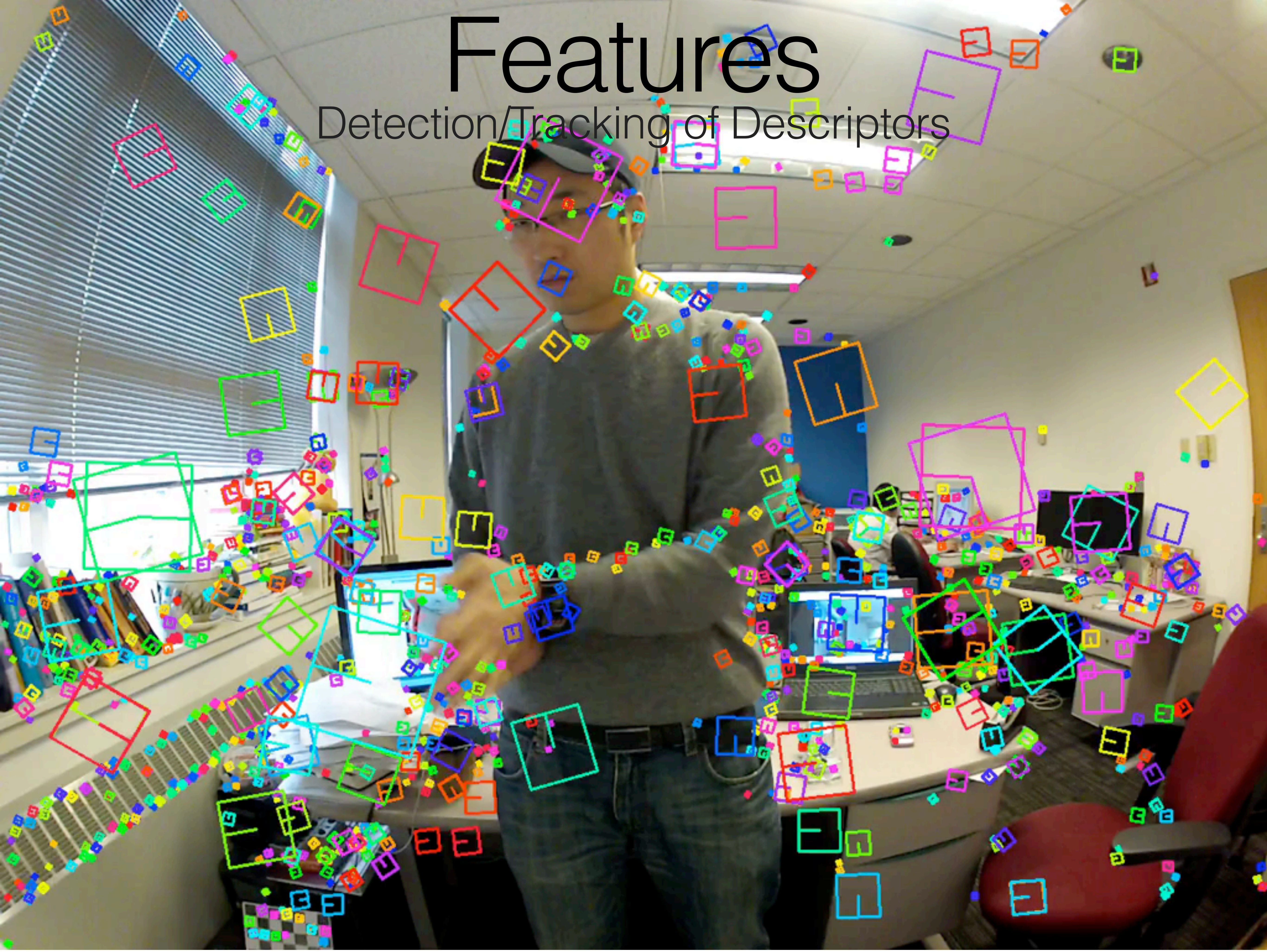
There is also shading, texture, and other cues. See Shape from X (Marr)

Shading

Surface normals from shading information

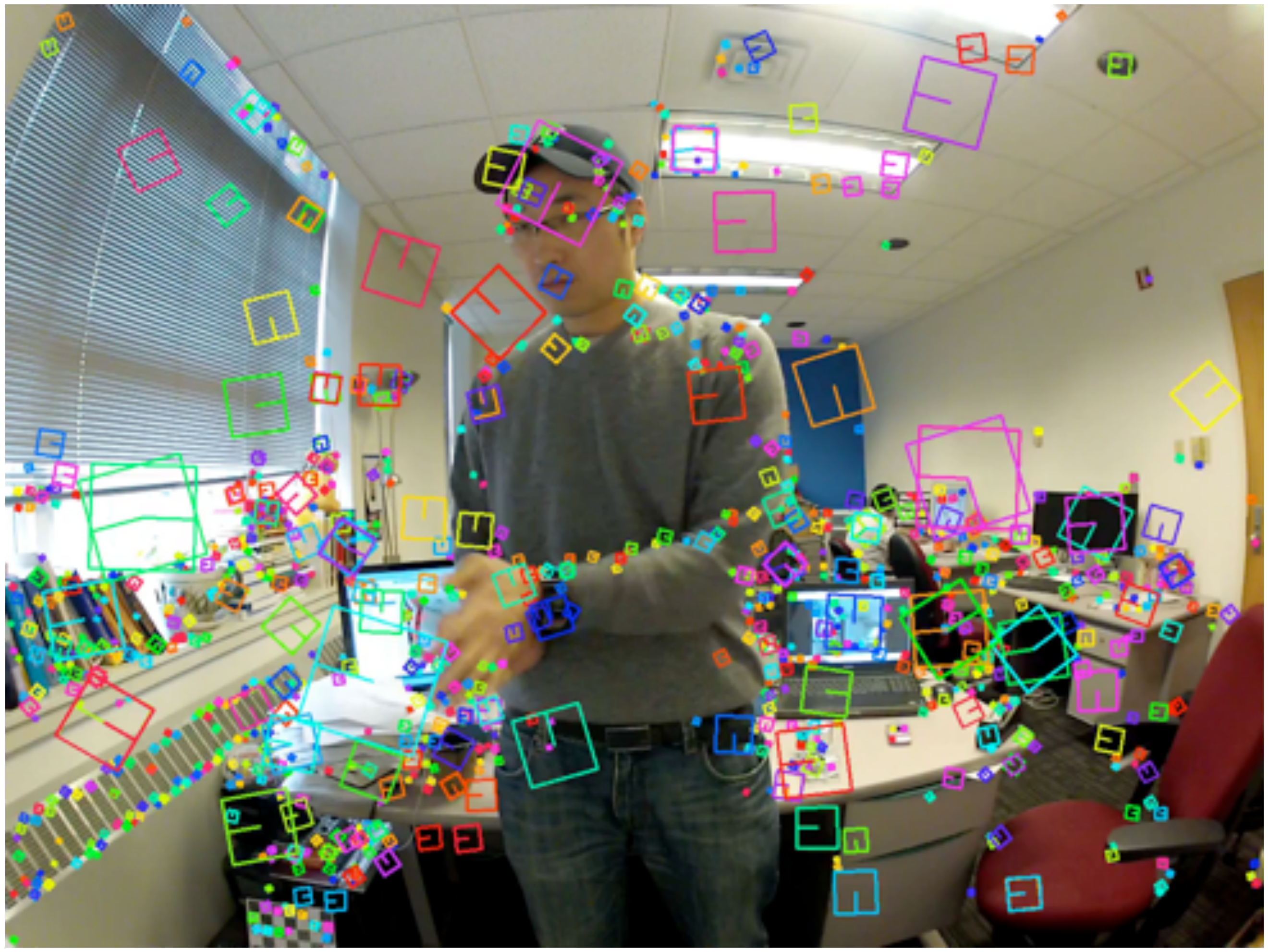
Features

Detection/Tracking of Descriptors



Correspondences

Feature-based Matching



Correspondences

Feature-based Matching

Silhouettes

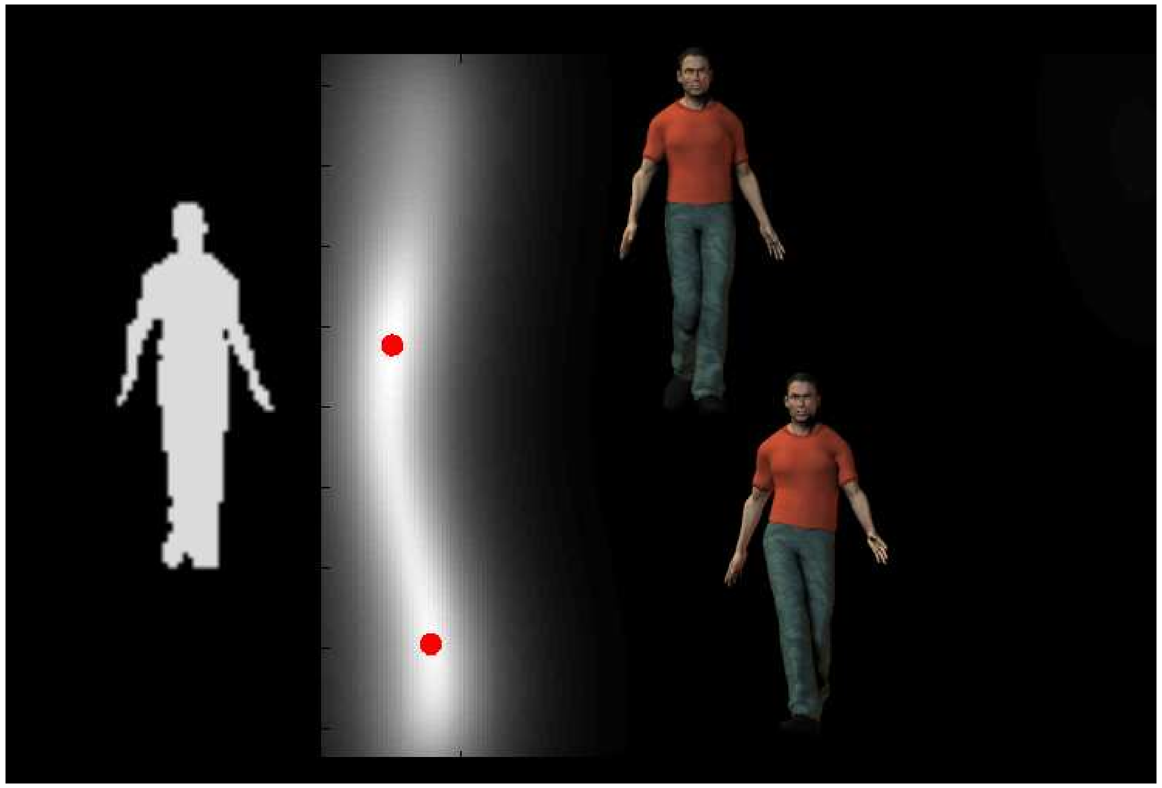
Background subtraction

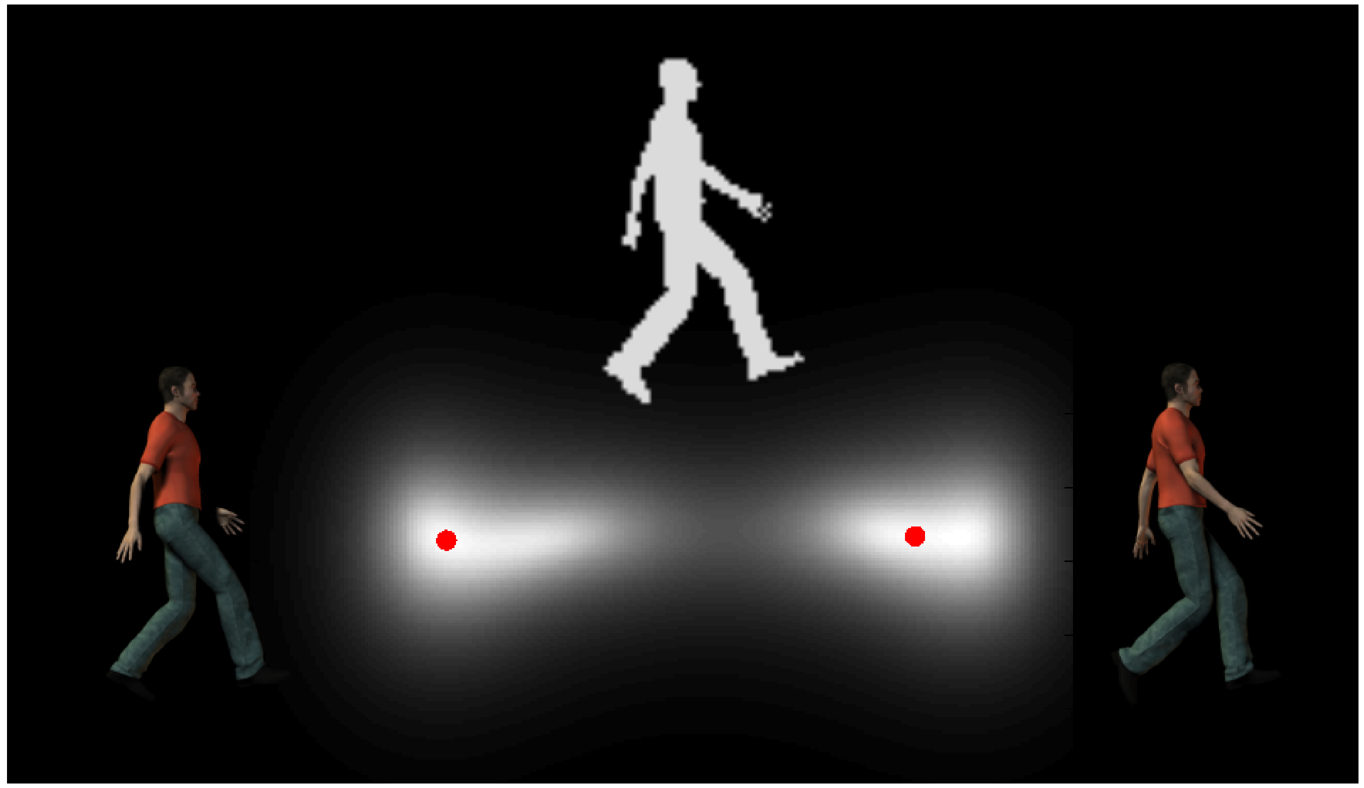
de Aguiar et al. Performance Capture from Sparse Multi-view Video, SIGGRAPH 2008.

Silhouettes

Holy Grail: Single Video Camera

Problem is unsolved. *Very* unsolved.

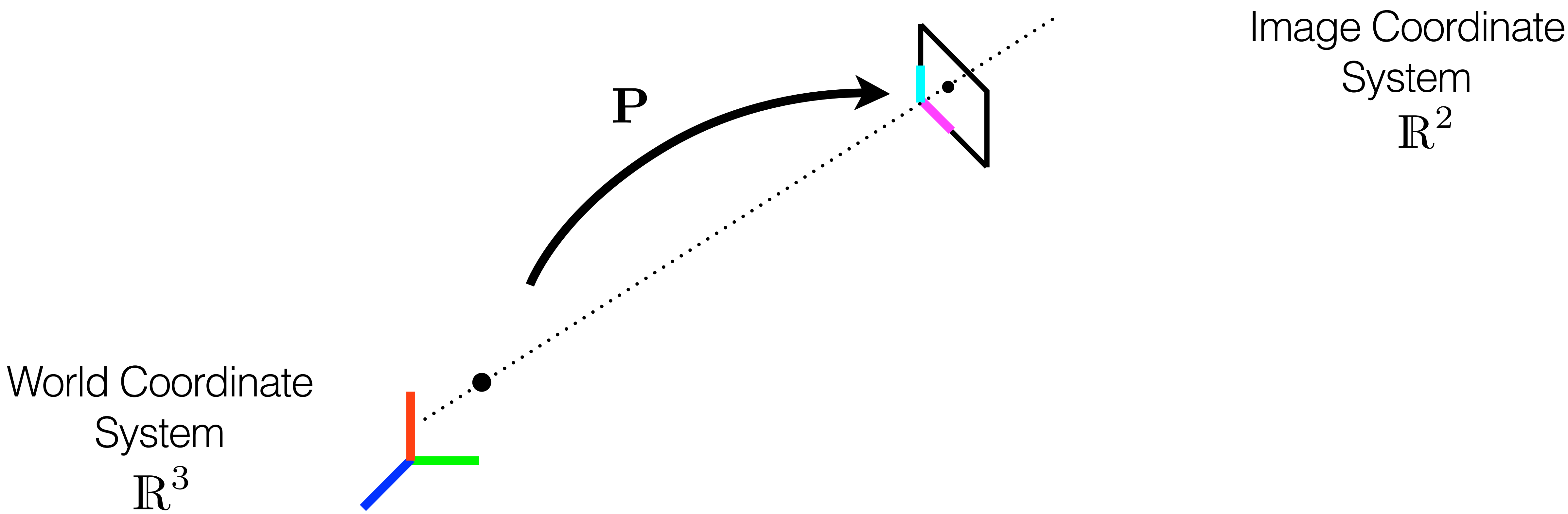




Based on slide by Raquel Urtasun

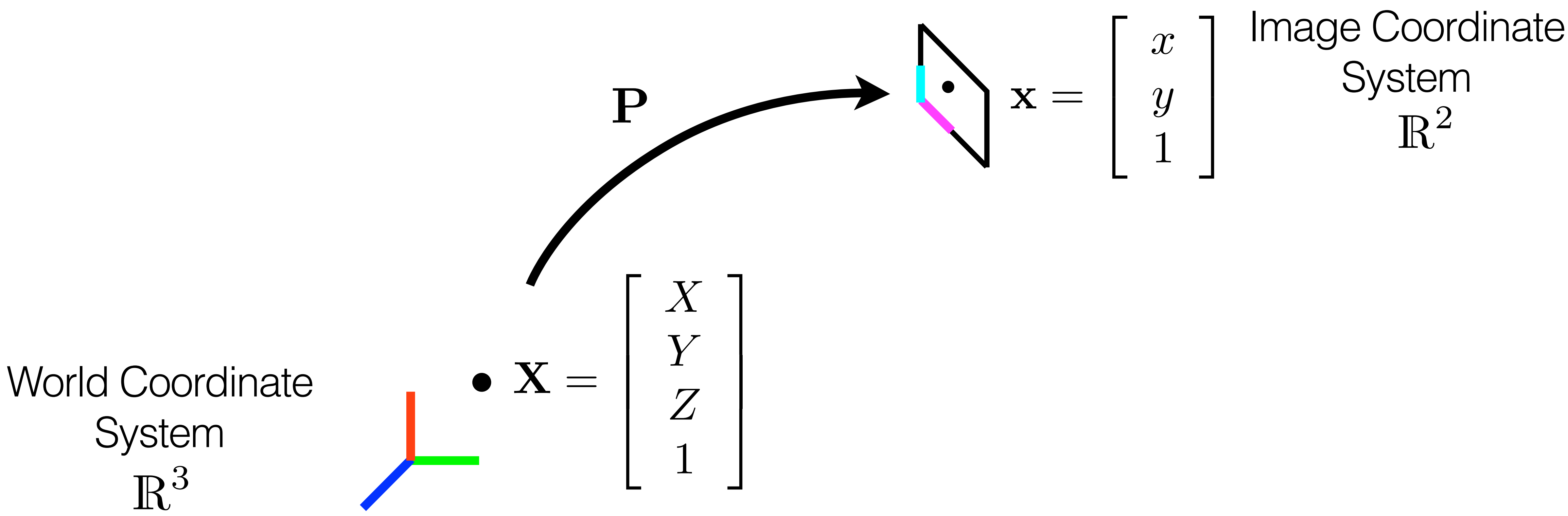
3D-2D Projection

How are images formed?



3D-2D Projection

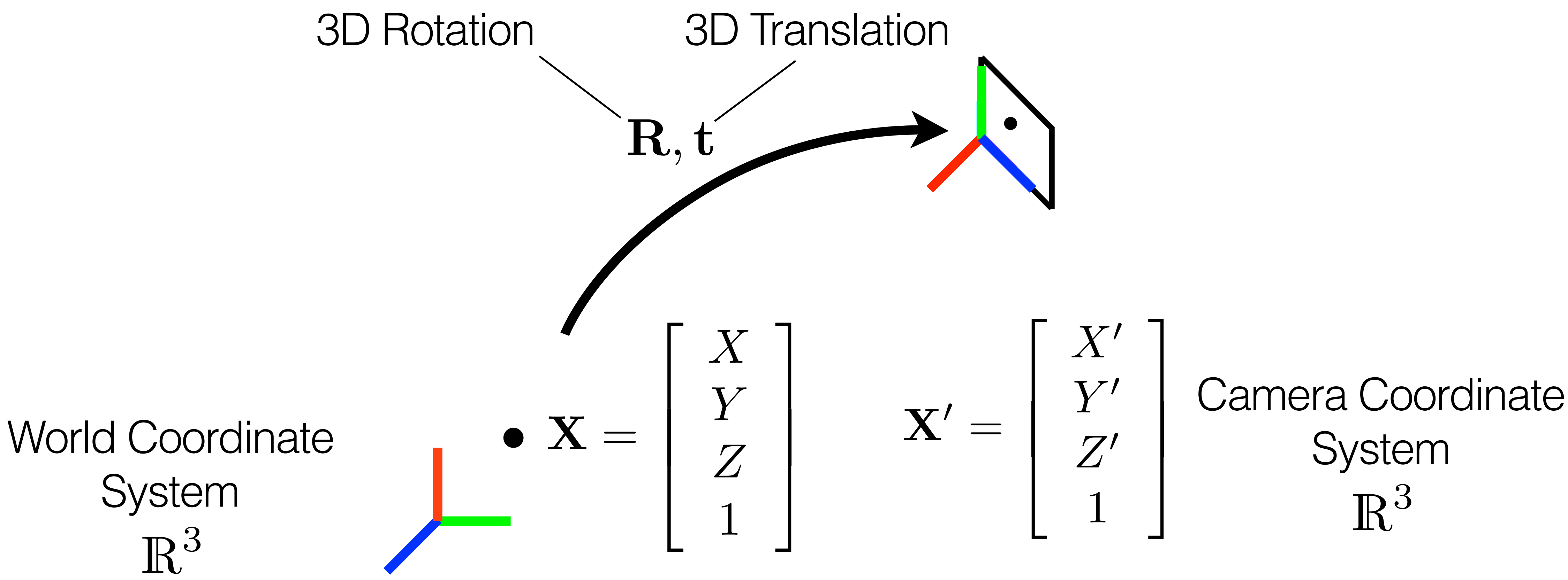
How are images formed?



$$\mathbf{X} \rightarrow \mathbf{x}$$

3D-3D Transformation

World Coordinate to Camera Coordinate

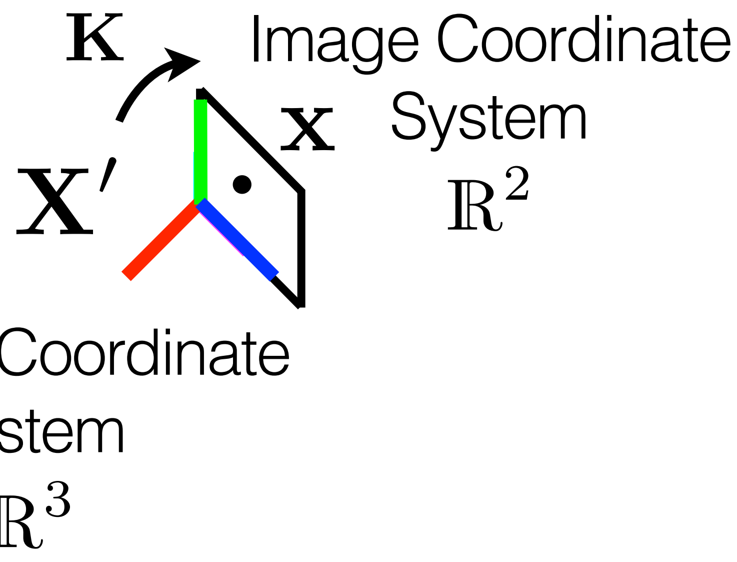


Point in Camera Coordinates — Point in World Coordinates

$$\begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{3\times3} & \mathbf{t}_{3\times1} \\ \mathbf{0} & 1 \end{bmatrix}_{4\times4} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

Intrinsic Matrix

Camera Coordinate to Image Coordinate



$$\begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix} = \mathbf{K}_{3\times 3} \left[\; \mathbf{I}_{3\times 3} \;\middle|\; \mathbf{0}_{3\times 1} \; \right]_{3\times 4} \begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix}$$

focal length

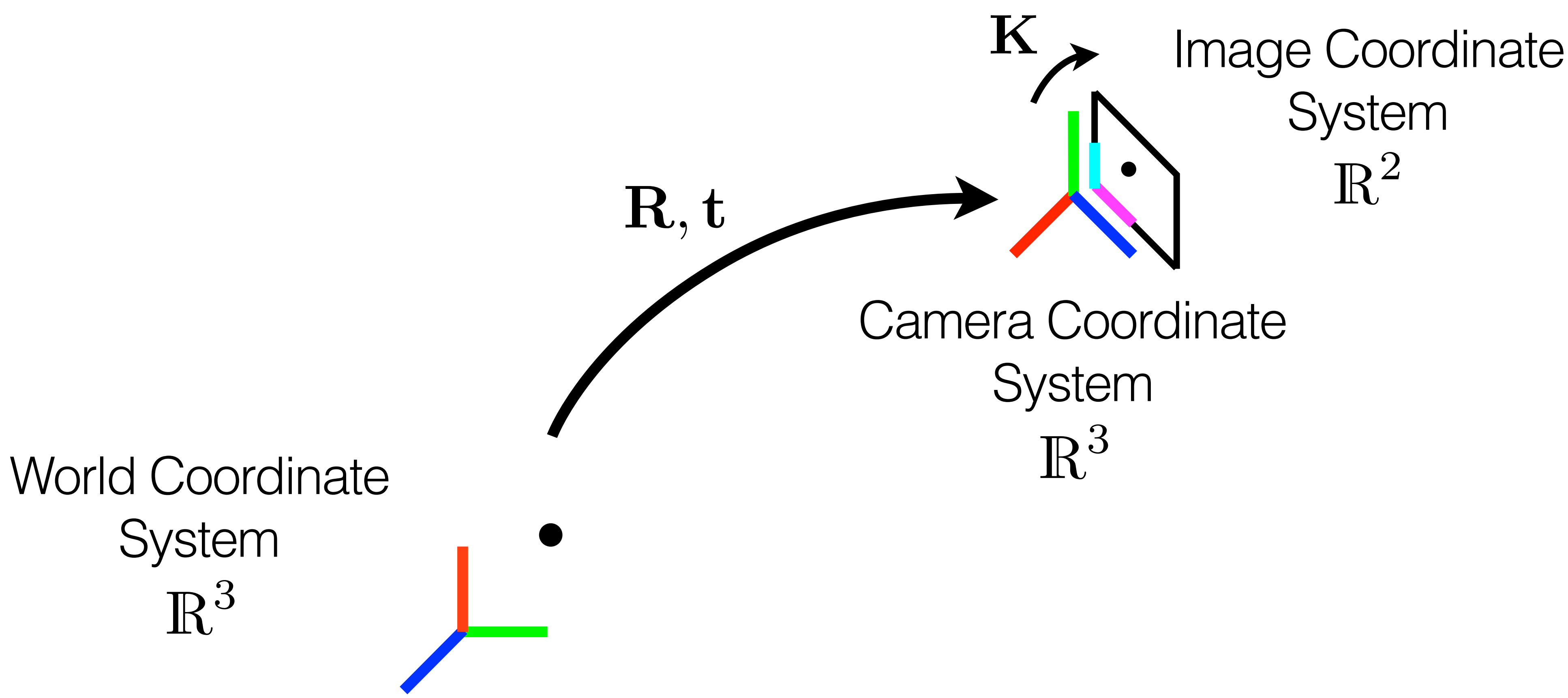
pixel scaling factors

Principal offset

$$\mathbf{K} = \begin{bmatrix} s_x f & 0 & p_x \\ 0 & s_y f & p_y \\ 0 & 0 & 1 \end{bmatrix}$$

3D-2D Projection

World to **Camera** to **Image** Coordinate



$$\begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix} = \mathbf{K}_{3\times 3} \left[\begin{array}{c|c} \mathbf{I}_{3\times 3} & \mathbf{0}_{3\times 1} \end{array} \right]_{3\times 4} \begin{bmatrix} \mathbf{R}_{3\times 3} & \mathbf{t}_{3\times 1} \\ \mathbf{0} & 1 \end{bmatrix}_{4\times 4} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

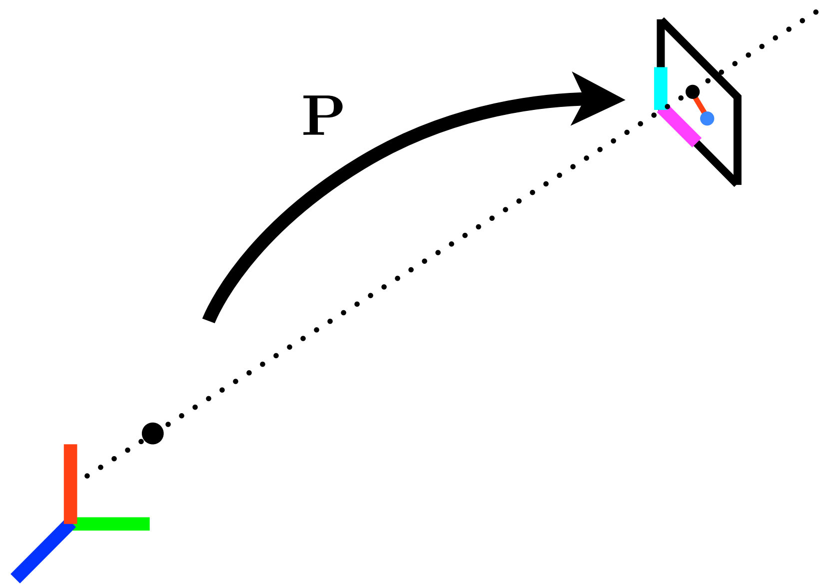
$$\mathbf{x} \cong \mathbf{K}[\mathbf{R}|\mathbf{t}]\mathbf{X}$$

$$\mathbf{x} \cong \mathbf{P}_{3\times 4}\mathbf{X}$$

Find **P** using camera calibration
http://www.vision.caltech.edu/bouguetj/calib_doc/

$$\|\cdot\|_d$$

Normalized Distance in the presence of noise



$$\mathbf{x} \cong \mathbf{P}_{3\times 4}\mathbf{X}$$

“equal up to scale” not “equal”

$$\begin{bmatrix} \lambda x \\ \lambda y \\ \lambda \end{bmatrix} = \mathbf{P} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

$$\mathbf{x} = \lambda \mathbf{P}\mathbf{X}$$

$$\|\mathbf{x} - \lambda \mathbf{P}\mathbf{X}\|_2 = \|\mathbf{x}, \mathbf{P}\mathbf{X}\|_d$$

Measure of Goodness
Maximum Likelihood Objective
(under Gaussian Noise)

Single Image Projection

Invertible?

$$\mathbf{x} \longleftrightarrow \mathbf{P}_{3\times 4}\mathbf{X}$$

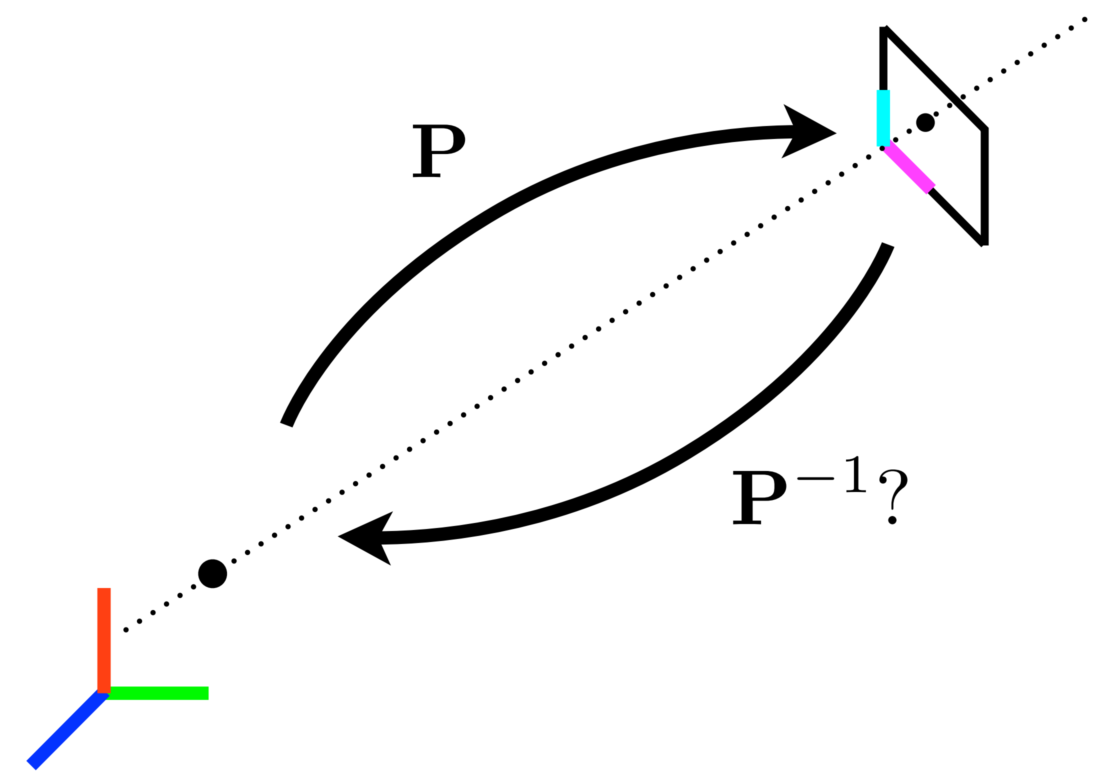


Image Coordinate System
$\mathbb{R}^2$

World Coordinate System
$\mathbb{R}^3$

$$\min_{\mathbf{X}} \|\mathbf{x}, \mathbf{PX}\|_d$$

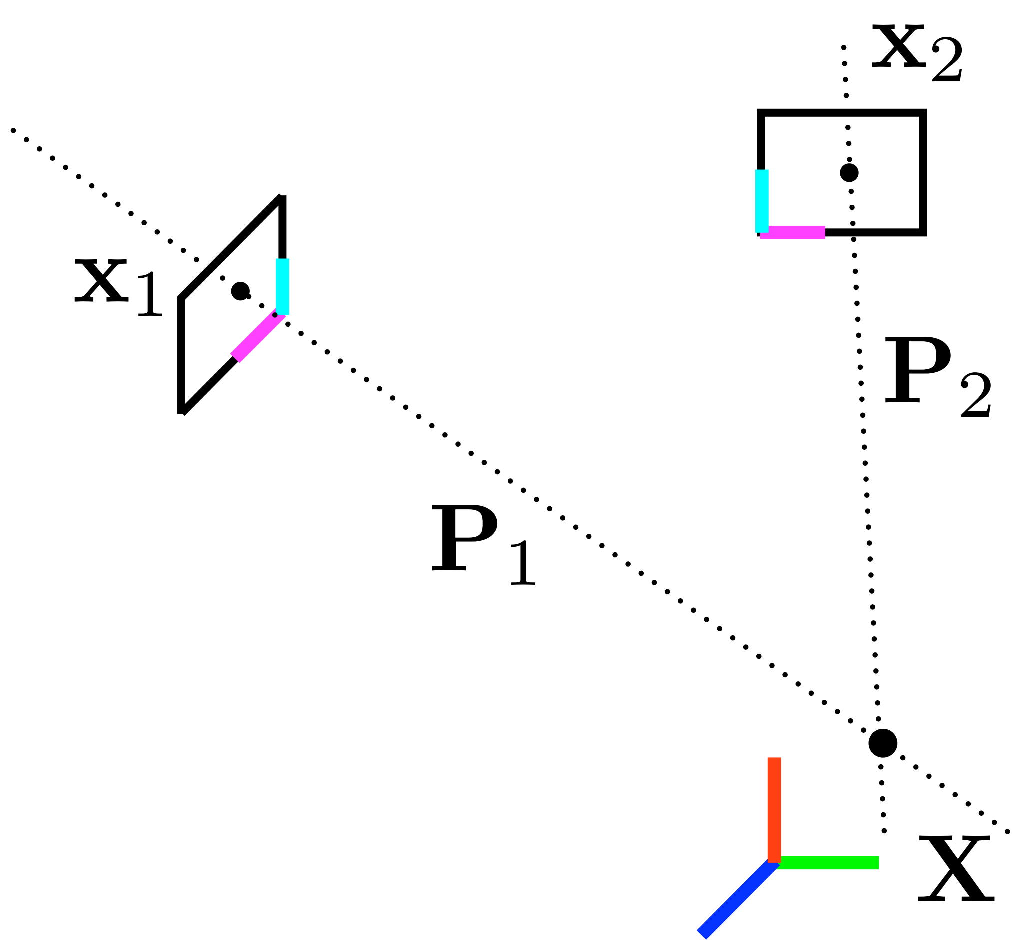
Ill-posed: 3 unknown, 2 constraints

Reconstruct me! :)

How do we resolve this?

Multiple Views!

$$\mathbf{x} \longrightarrow \mathbf{P}_{3\times 4}\mathbf{X}$$

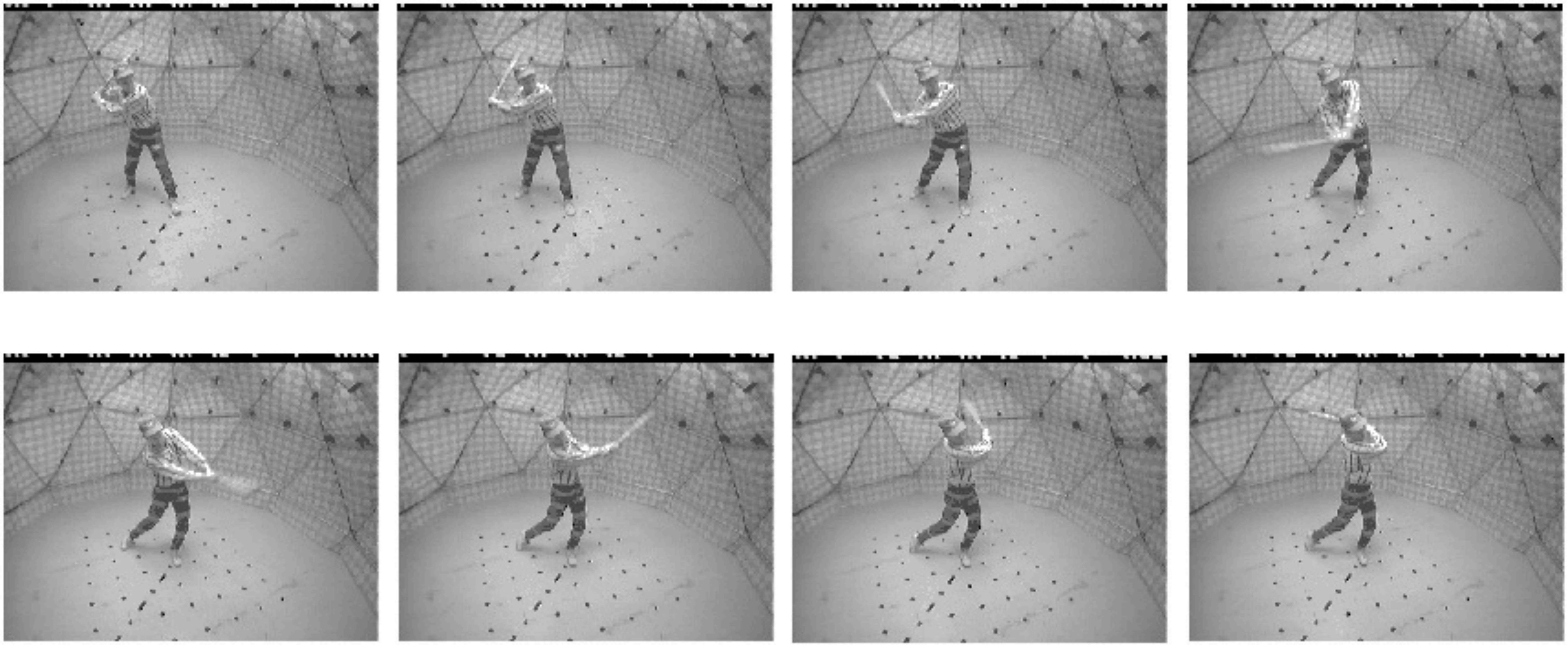


Virtualized Reality™

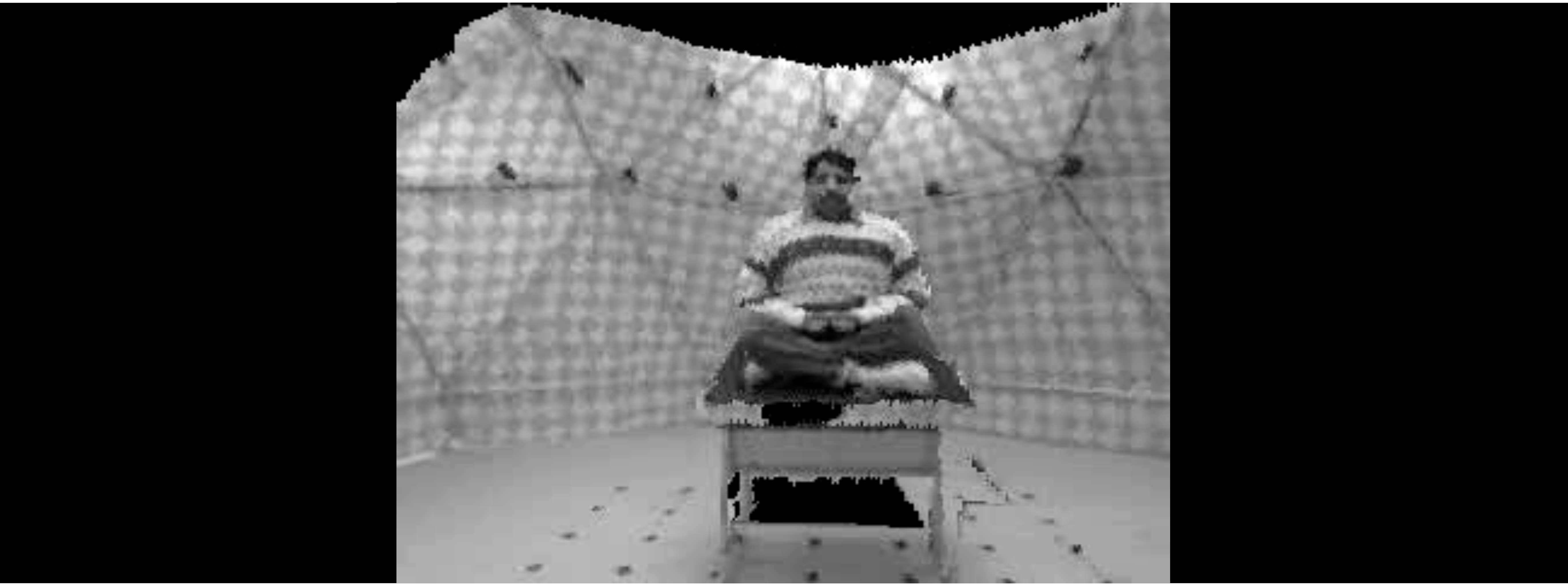
Takeo Kanade

Virtualizing Studio

Kanade, Narayanan, Rander (1995)

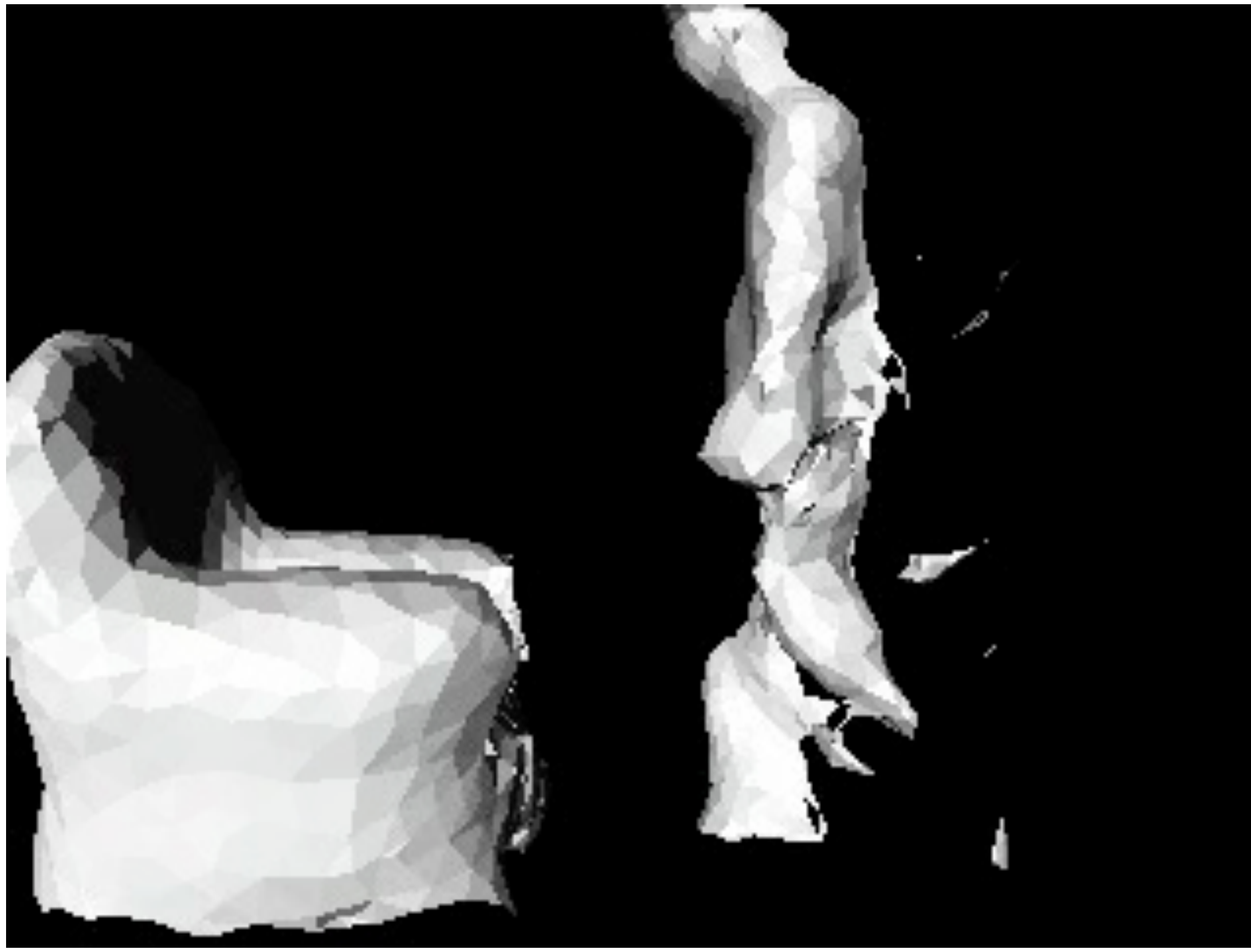


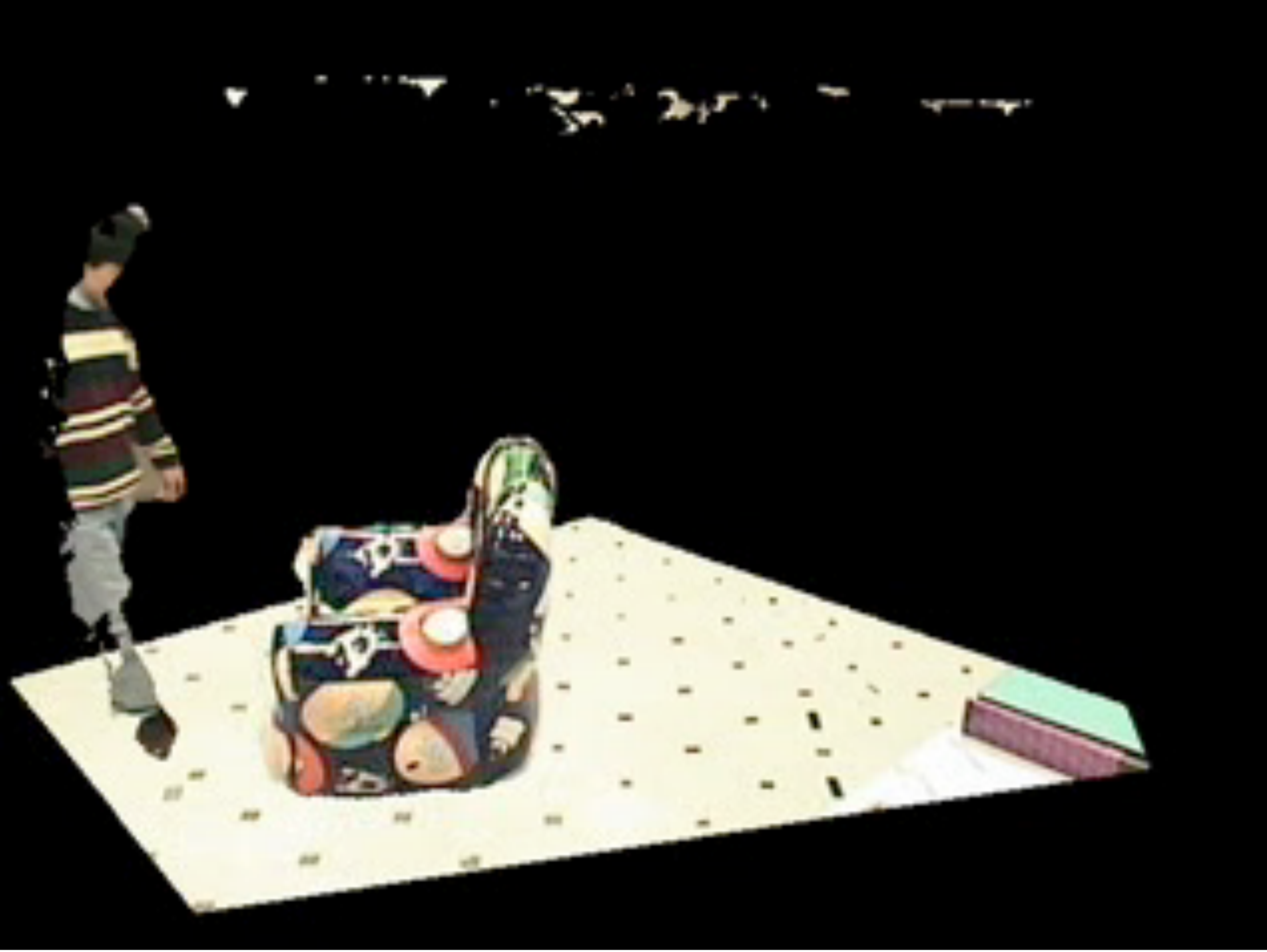
Virtualizing Studio



Virtualizing Studio

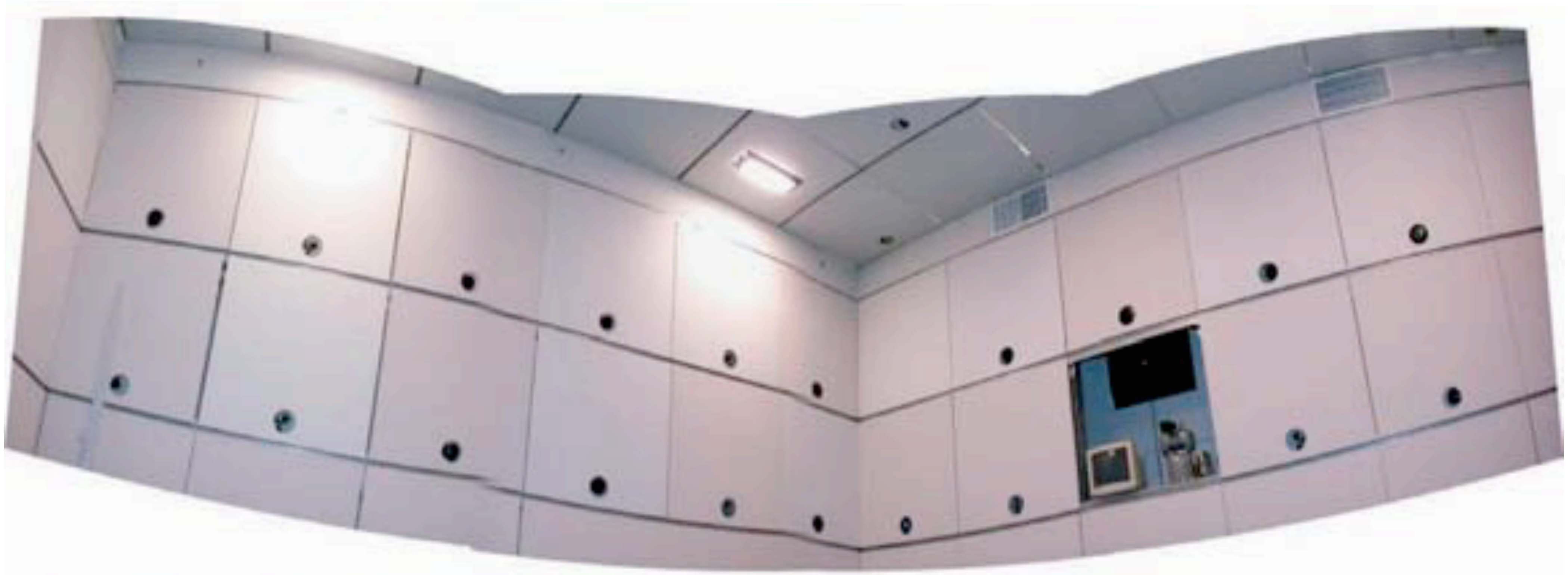
Vedula, Saito, Kanade (1998)

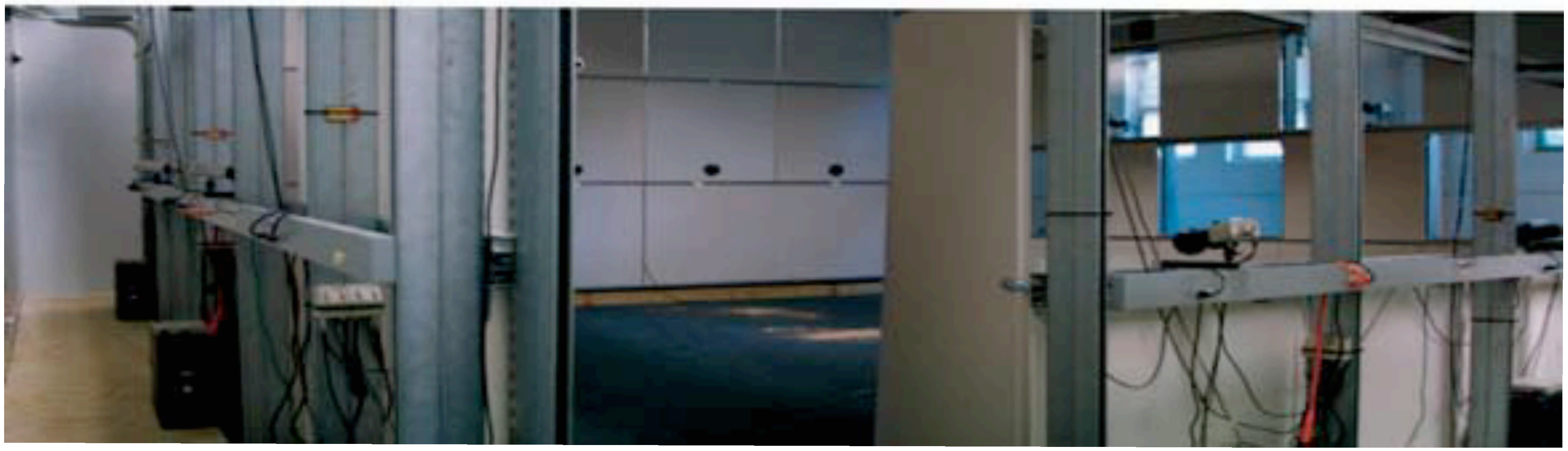



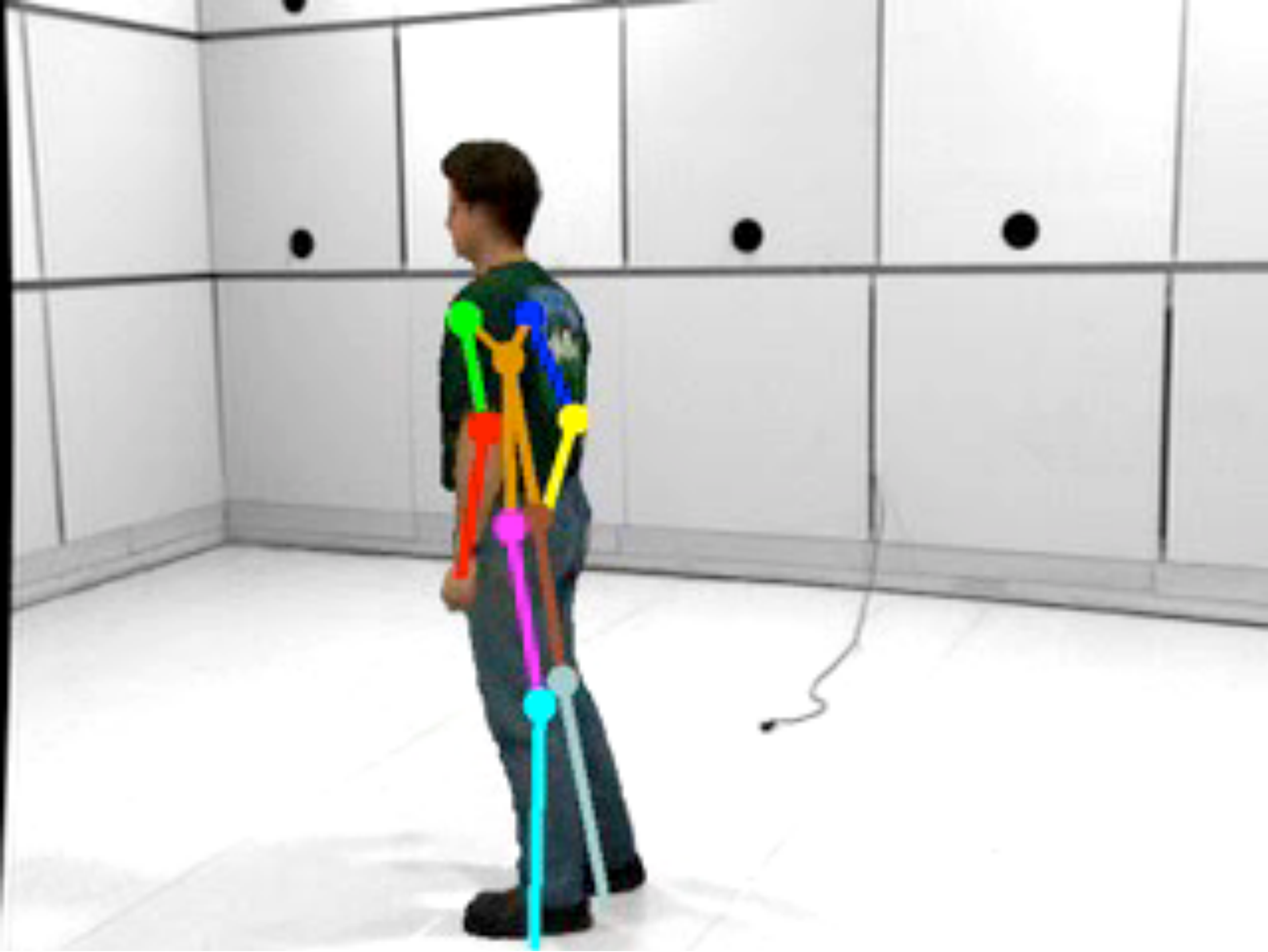


Virtualizing Studio

Matthews, Baker, Gross, Kanade (2002)







Blue-C

Gross et al. (2003)

16 cameras

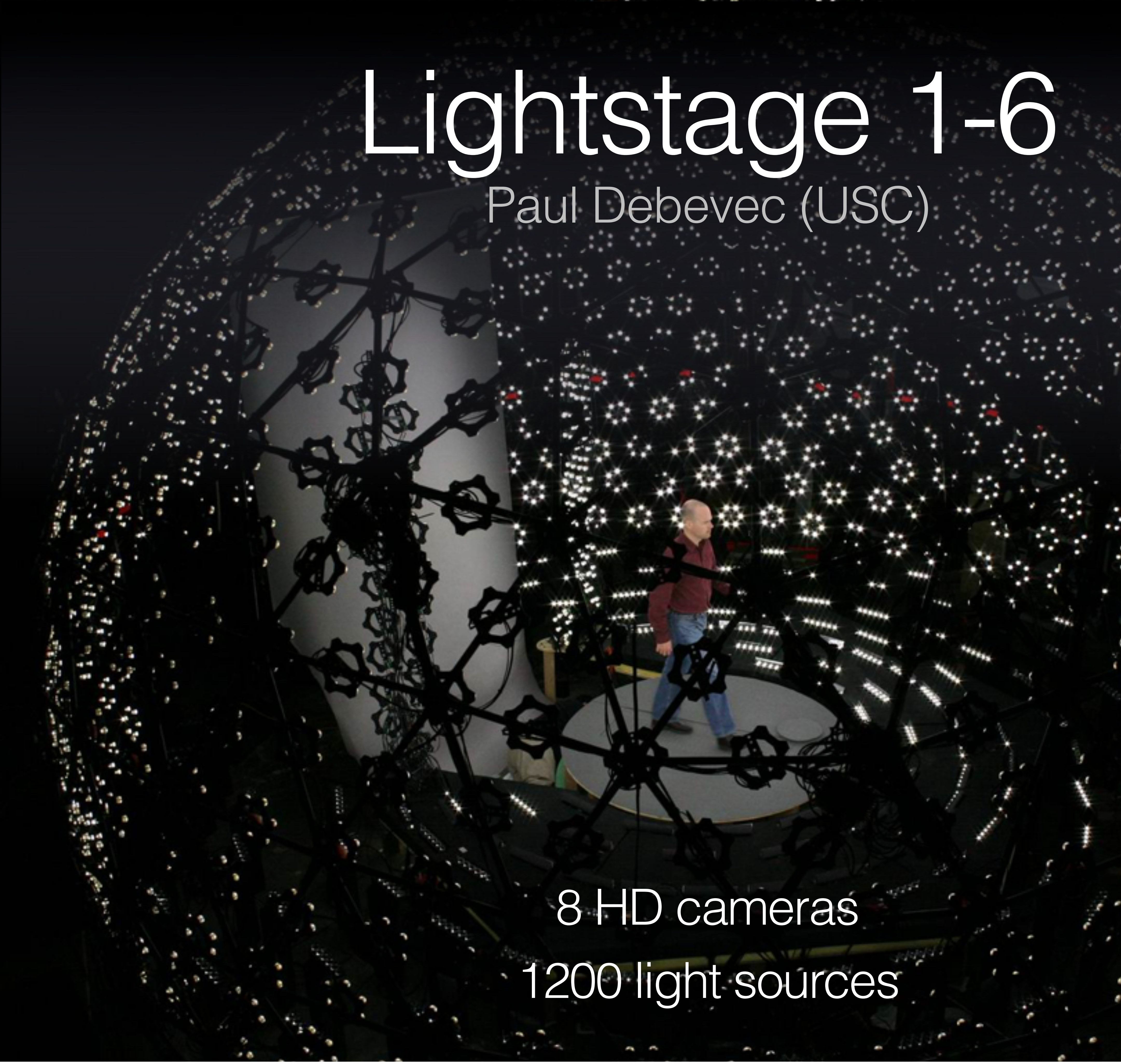
Stanford Multicamera Array

Levoy et al. (2005)

100 VGA Cameras

Lightstage 1-6

Paul Debevec (USC)



8 HD cameras

1200 light sources

Onsite 3D Video Capture

Nobuhara et al. (2009)

16 UXGA cameras

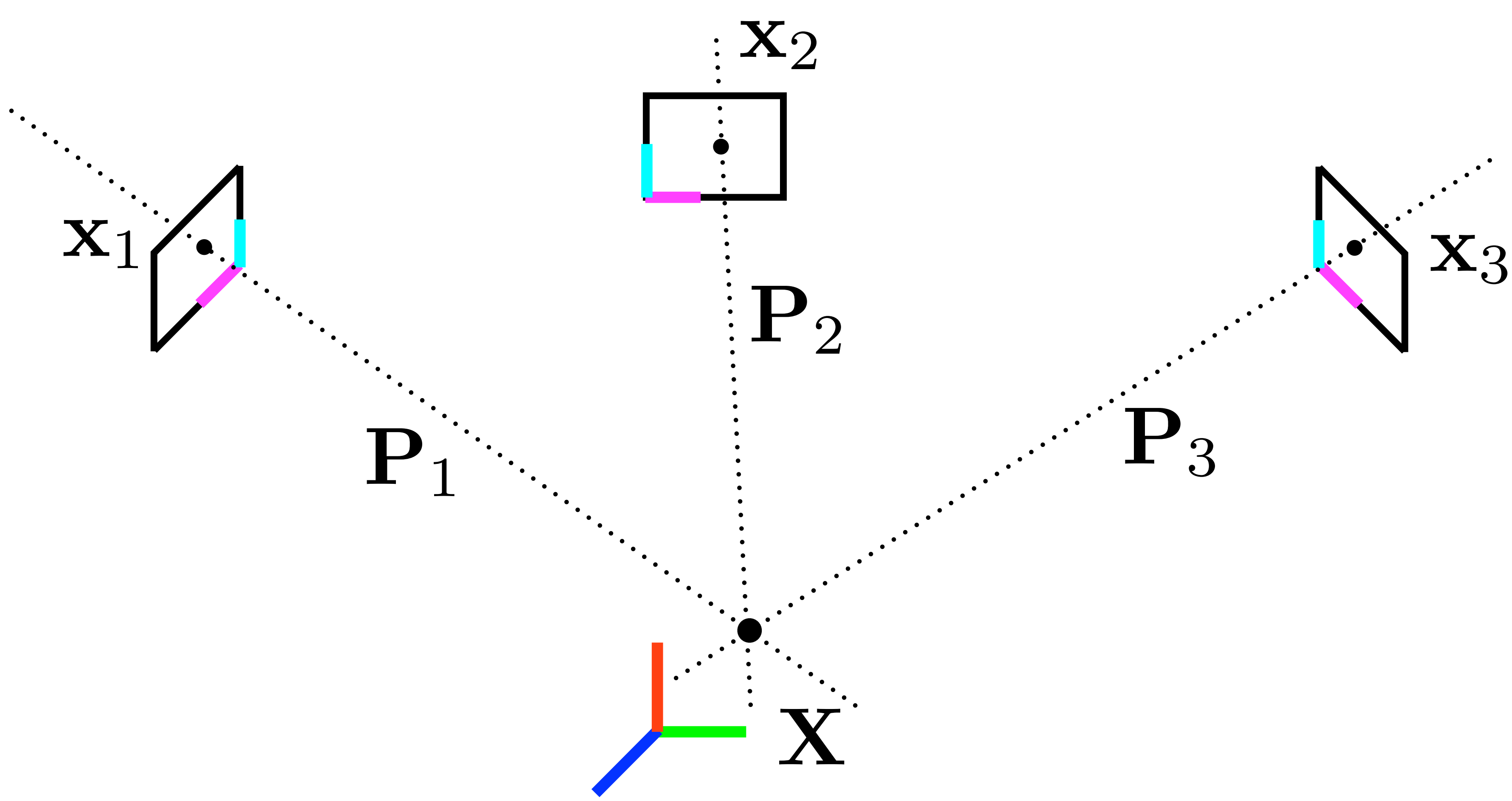
Video courtesy of Shohei Nobuhara

Discussion

Multiple View Reconstruction

Resolve ambiguity

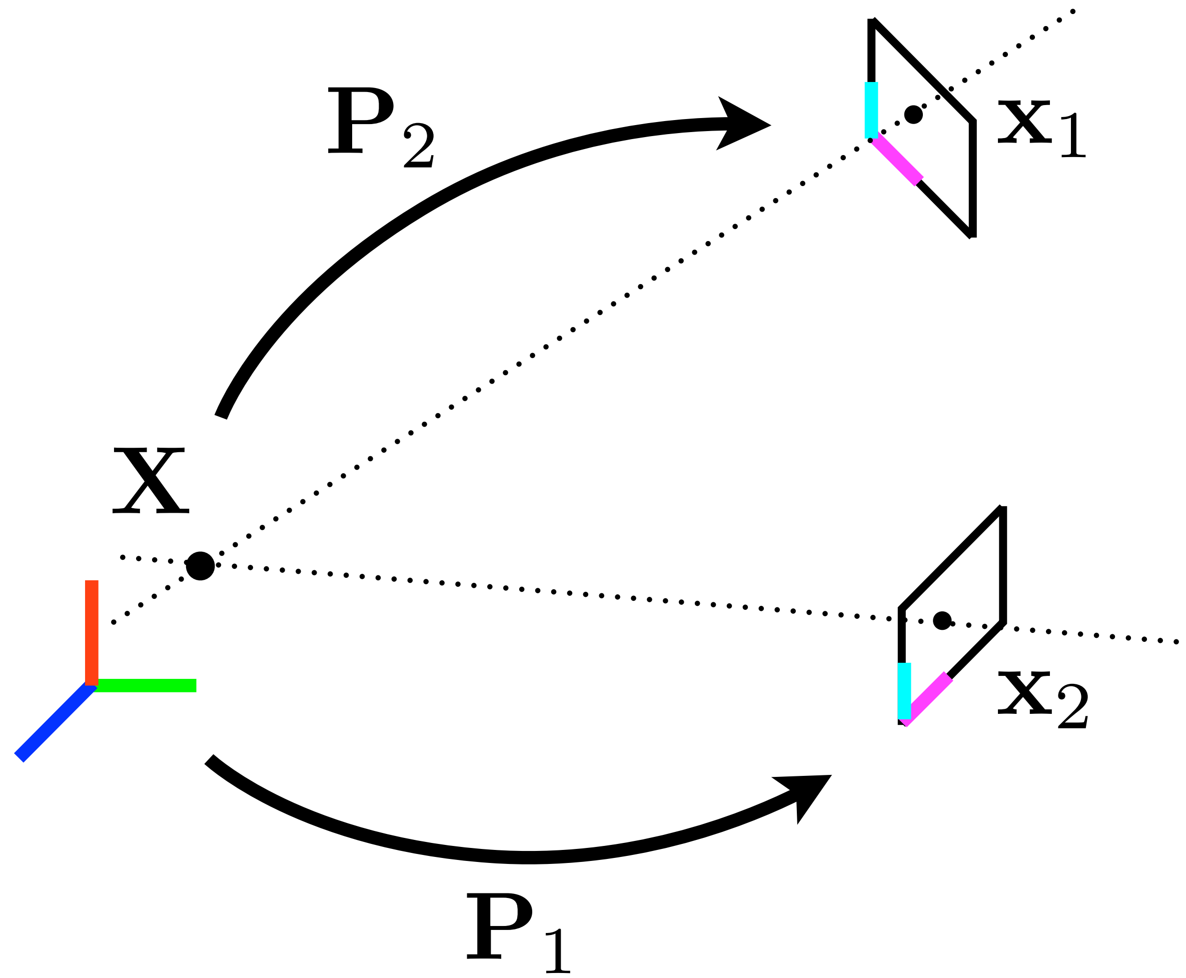
$$\mathbf{x} \longleftrightarrow \mathbf{P}_{3\times 4}\mathbf{X}$$



$$\min_{\mathbf{X}} \sum_i \|\mathbf{x}_i, \mathbf{P}_i\mathbf{X}\|_d$$

Stereoscopic 3D Reconstruction

Correspondence-based



$$\min_{\mathbf{X}} \|\mathbf{x}_1, \mathbf{P}_1\mathbf{X}\|_d + \|\mathbf{x}_2, \mathbf{P}_2\mathbf{X}\|_d$$

Nonlinear least squares

Initialization

Direct Linear Transform Algorithm

$$\mathbf{x} \cong \mathbf{P}_{3\times 4}\mathbf{X}$$

Projection Equation --- Equal up to scale

$$\|\mathbf{x} - \lambda \mathbf{P}\mathbf{X}\|_2 = \|\mathbf{x}, \mathbf{P}\mathbf{X}\|_d$$

Normalized Distance

$$\mathbf{x} \times \lambda \mathbf{P}\mathbf{X} = 0$$

Cross product:

$$\mathbf{x} \times \mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\| \sin(\theta)\ \mathbf{n}$$

$$\mathbf{x} \times \mathbf{P}\mathbf{X} = 0$$

Cross product of a vector and a scaled version of itself is zero

Function of **x** and **P**

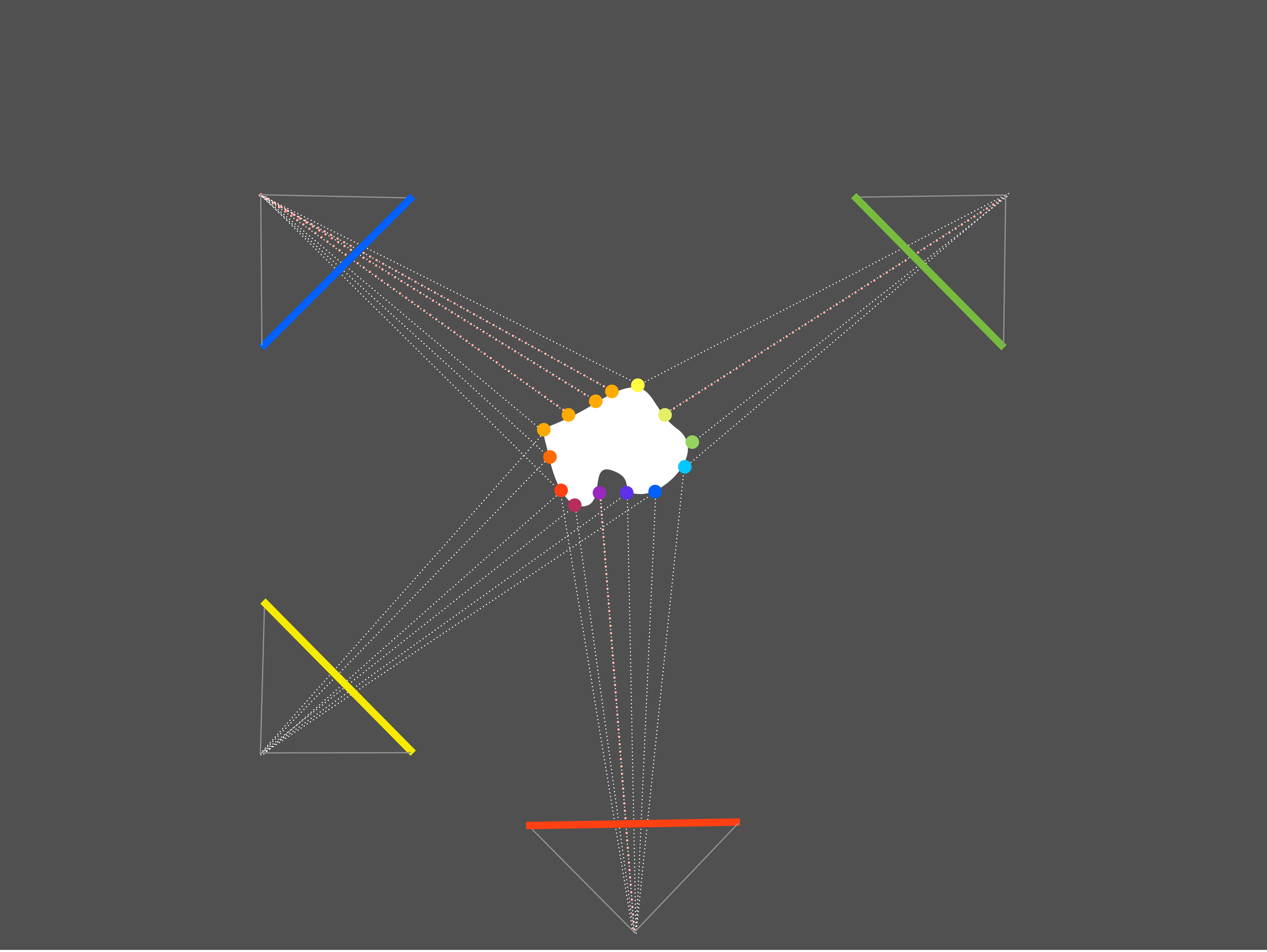
$$\mathbf{A}_{2\times 4}\mathbf{X}_{4\times 1} = 0$$

Underconstrained Homogeneous System

From camera 1
⋮
From camera *F*

$$\begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_F \end{bmatrix} \mathbf{X} = 0$$

Homogeneous System --- Solve using SVD



Challenge

Correspondence

The three most important problems in computer vision are registration, registration, registration!

--- Takeo Kanade

Stereoscopic 3D Reconstruction

Pros	**Cons**

Stereoscopic 3D Reconstruction

Pros

Can provide temporal correspondence

High accuracy

Accuracy depends on the number of cameras

Can identify concavities

Cons

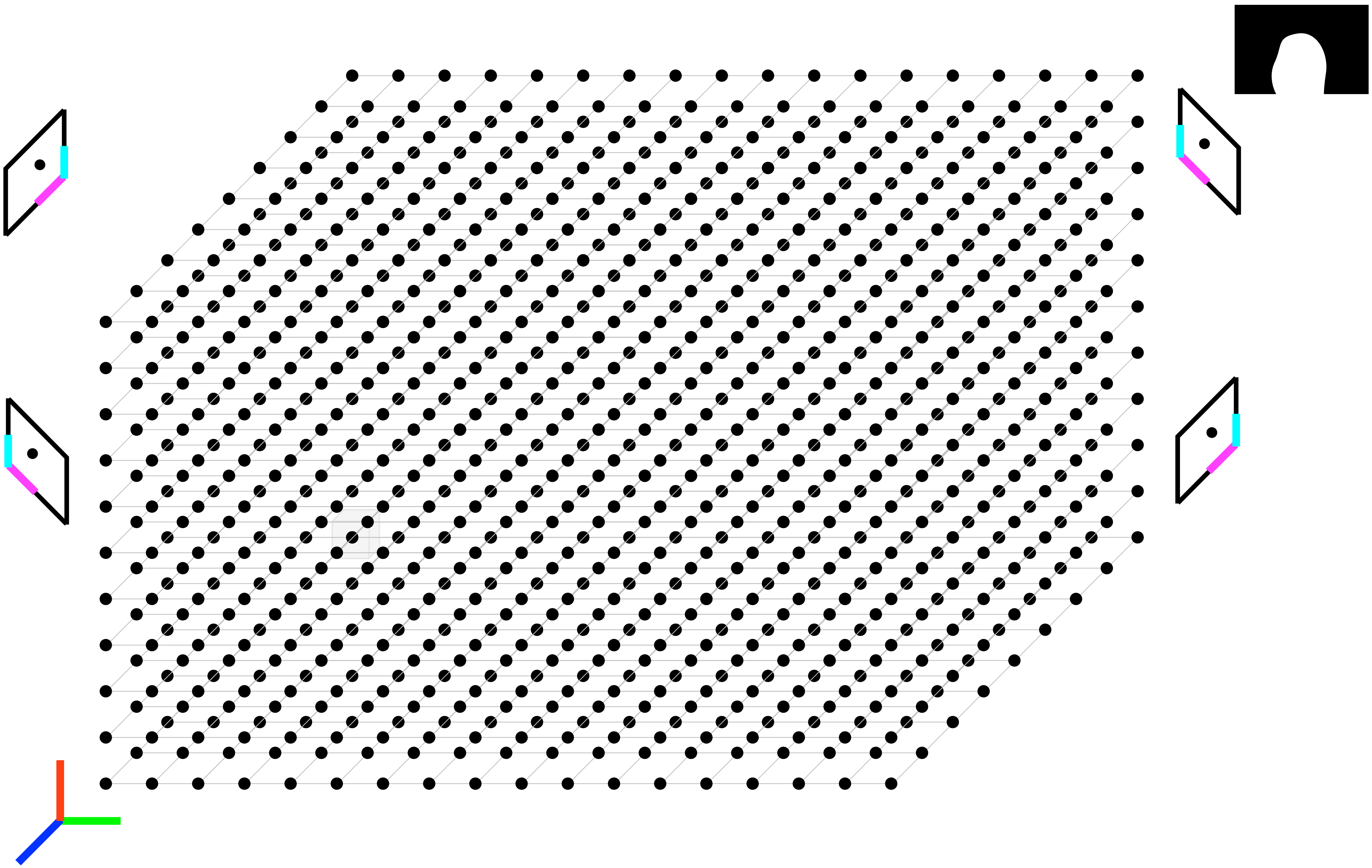
Requires accurate spatial correspondence

Sparse reconstruction

Does not provide normal information

Voxel Carving

Correspondence-free Reconstruction



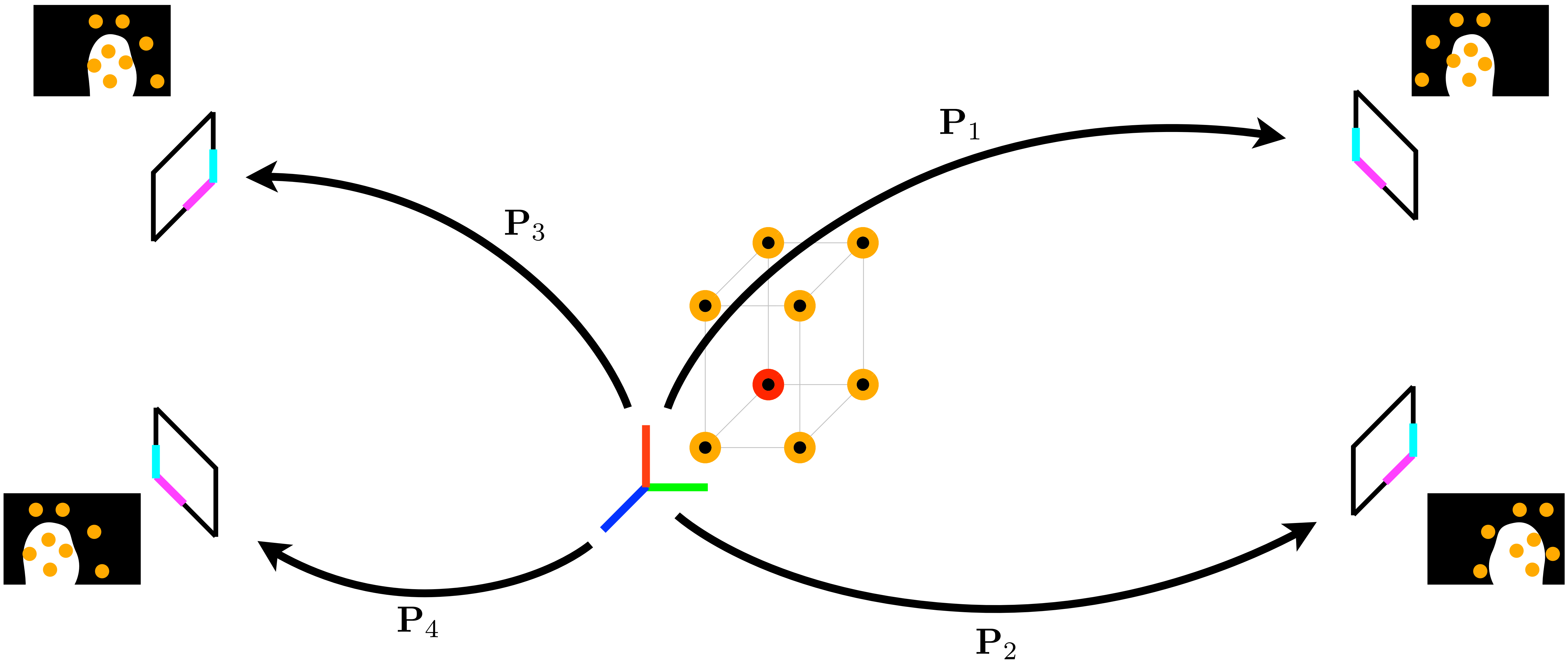
Silhouettes

Background subtraction

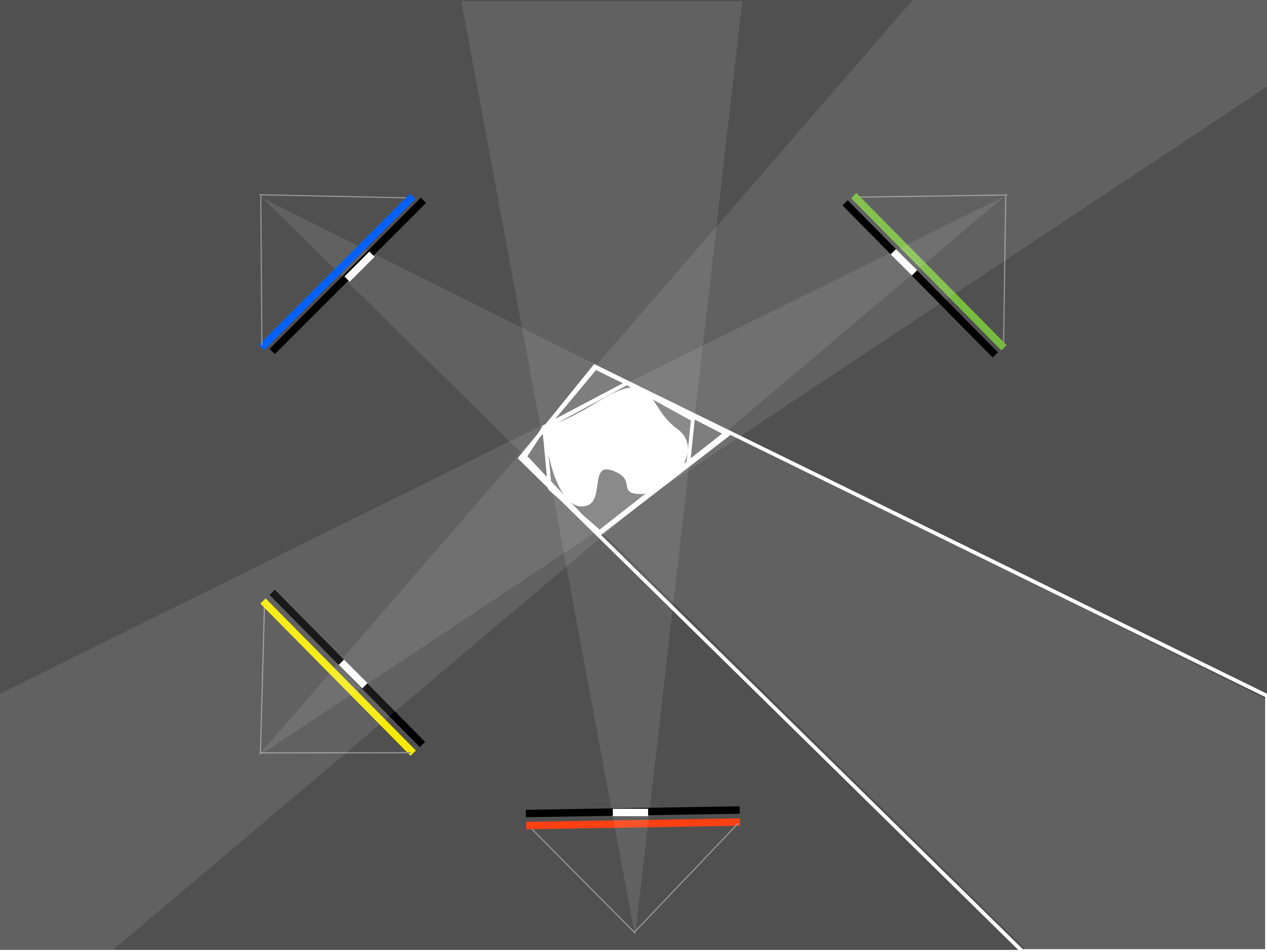
de Aguiar et al. Performance Capture from Sparse Multi-view Video, SIGGRAPH 2008.

Voxel Carving

Correspondence-free Reconstruction



$$\mathbf{x} \cong \mathbf{P}_{3\times 4}\mathbf{X}$$



Visual Hull

Gregor Miller, 2008

Voxel-Carving

Pros | **Cons**

Voxel-Carving

Pros

Does not require spatial correspondences

Trades off density with computation

Easy to code

Camera work with few cameras

Cons

Does not provide temporal correspondence

Redundant computation

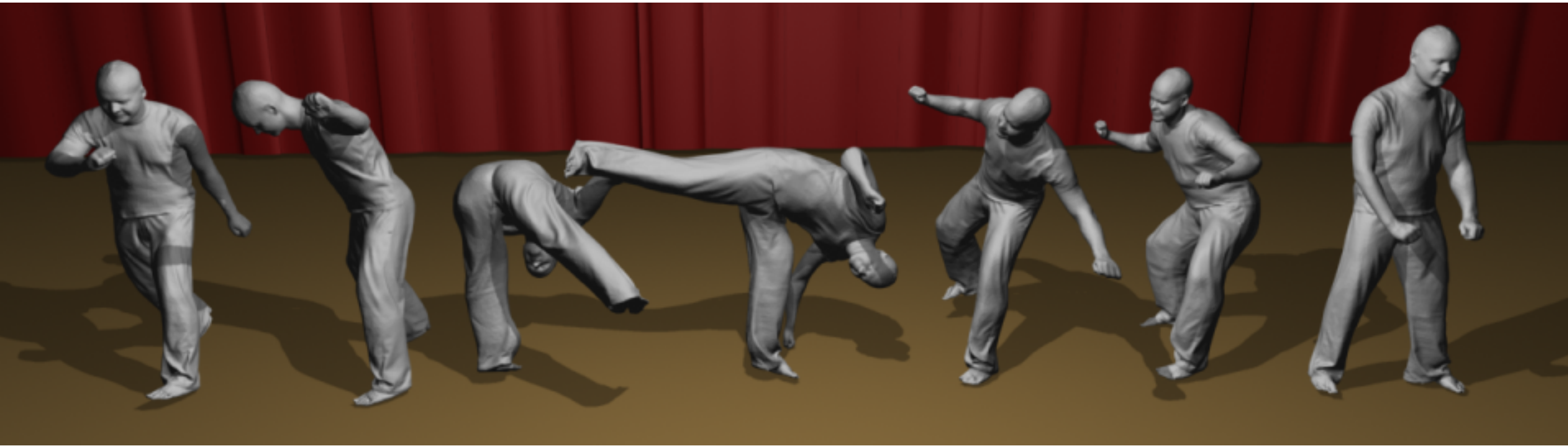
Requires accurate silhouettes

Does not provide normal information

Accuracy depends on the number of cameras

Convex Hull

Animating 3D Scans



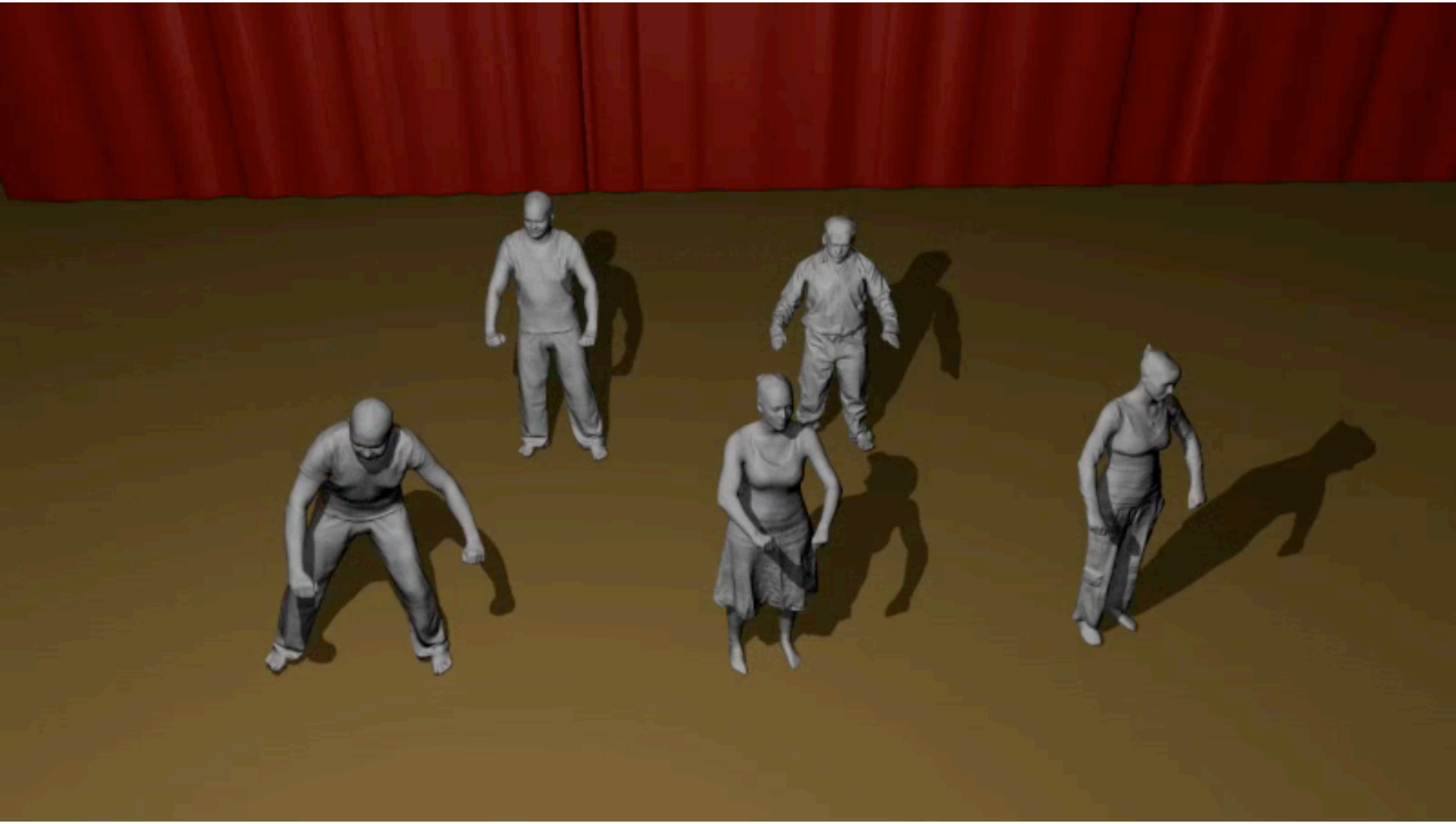
E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, S. Thrun, *Performance Capture from Sparse Multi-view Video*, in Proc. of SIGGRAPH 2008

SCAPE: Shape Completion

Anguelov (2005)

The Kitchen Sink

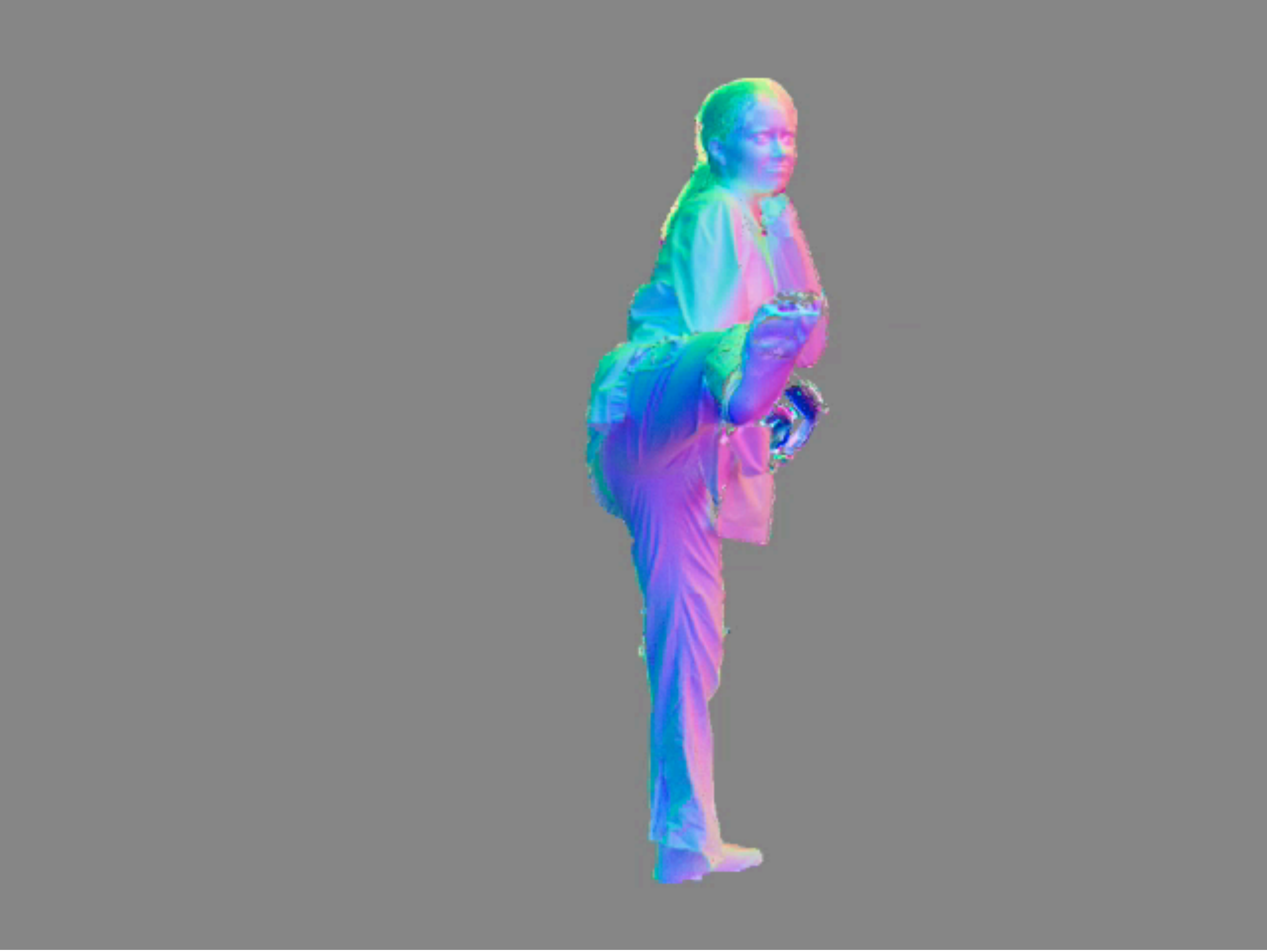
de Aguiar (2008)



Animating 3D Scans

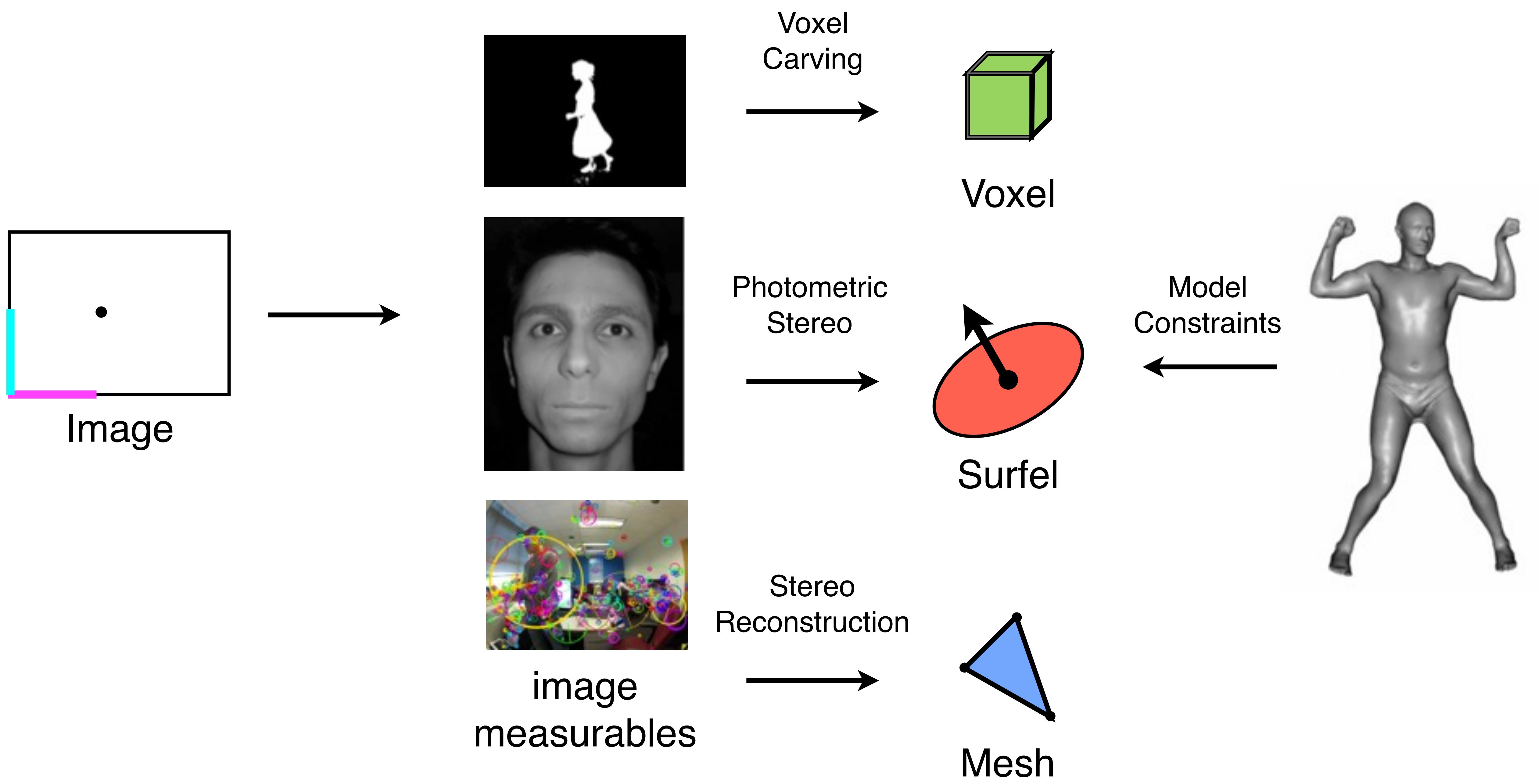
Pros and Cons

Pros	Cons
High resolution	Drift
Can fill missing data	Topology changes
Temporal continuity	Low detail (if generic models are used)
	Baked detail (if specific models are used)

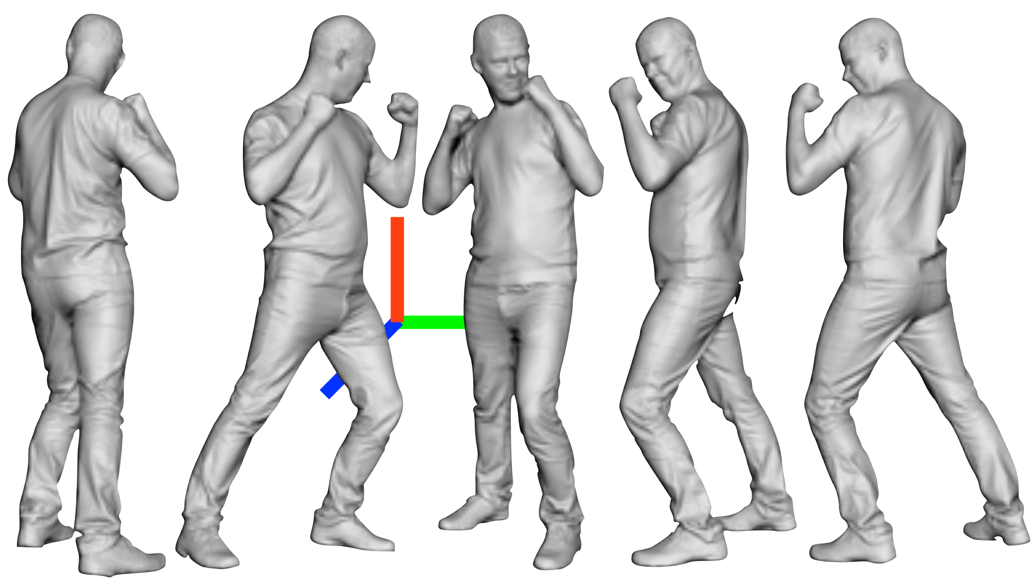


Representation

Reconstructing 3D Body Shape and Motion



Conclusion



3D Structure

3D Motion

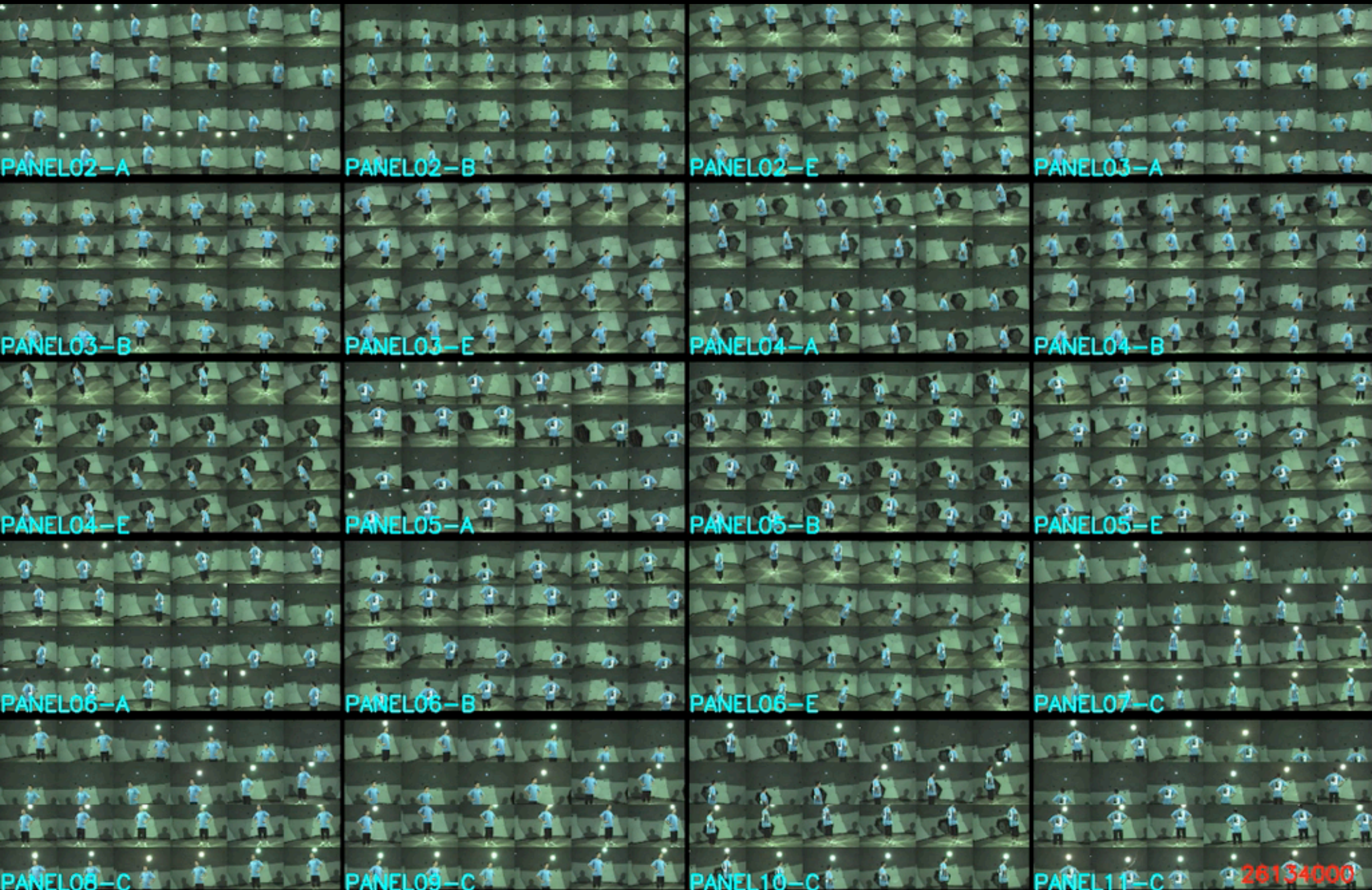
3D Structure reconstruction is maturing.
3D Motion estimation is primitive.

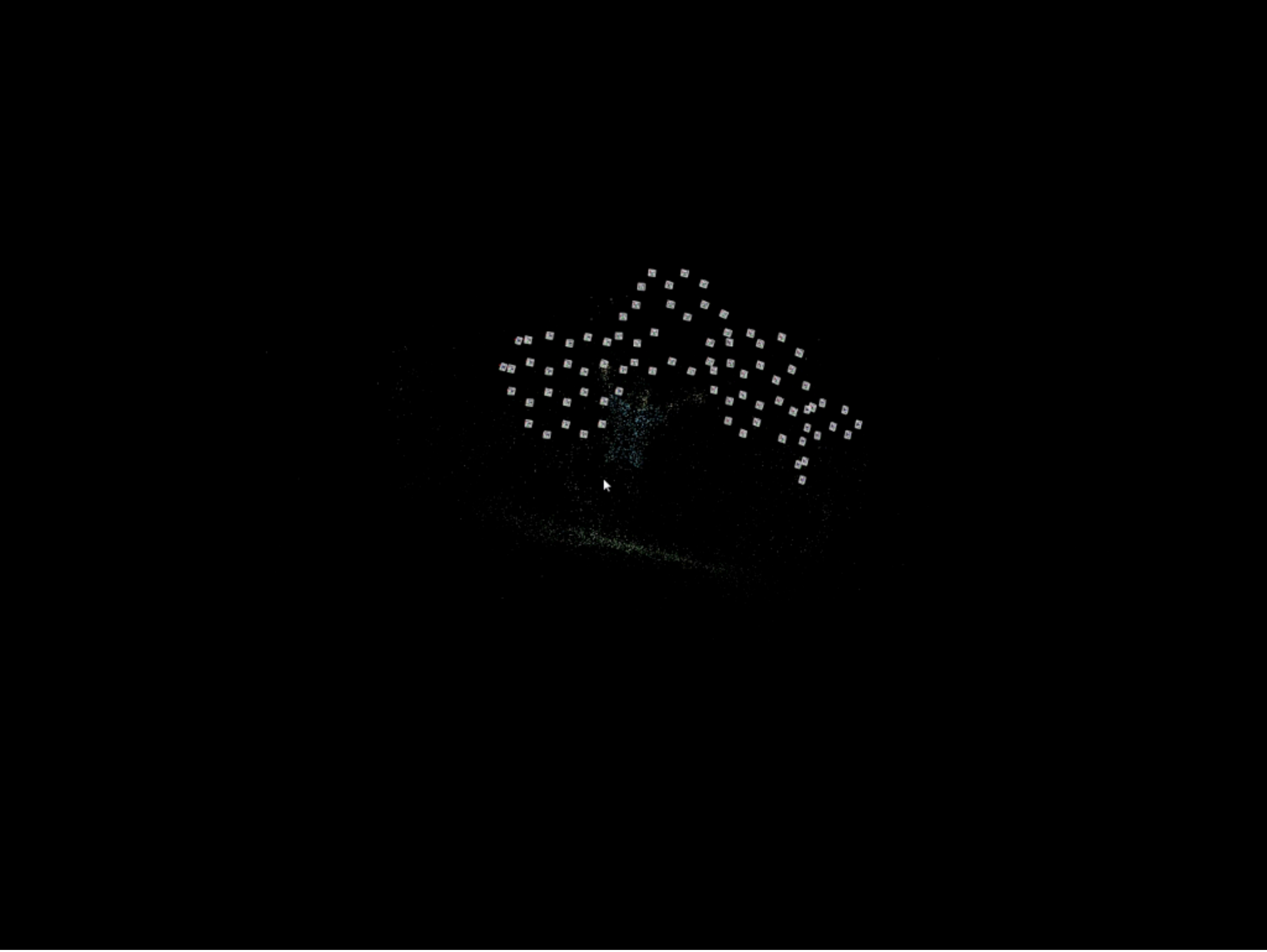
This Lecture...

3D Dynamic Surface Reconstruction using Passive Sensing

- How should we represent human body surfaces?
- What can we extract from images?
- A Brief History of Virtualizing Reality
- Volumetric and Point-based 3D Reconstruction Algorithms
- Tour of the Virtualizing Studio 4.0

Virtualizing Studio 4.0


PANEL02-A
PANEL02-B
PANEL02-E
PANEL03-A
PANEL03-B
PANEL03-E
PANEL04-A
PANEL04-B
PANEL04-E
PANEL05-A
PANEL05-B
PANEL05-E
PANEL06-A
PANEL06-B
PANEL06-E
PANEL07-C
PANEL08-C
PANEL09-C
PANEL10-C
PANEL11-C
26134000



Further Reading

Subtitle

- Matusik et al. Image-Based Visual Hulls, SIGGRAPH, 2000.
- de Aguiar et al. Performance Capture from Sparse Multi-view Video, SIGGRAPH 2008.
- Pfister et al., Surfels: Surface Elements as Rendering Primitives, SIGGRAPH 2000.
- Matsuyama et al., 3D Video and Its Applications, 2012.
- Vlasic et al., Dynamic Shape Capture using Multi-View Photometric Stereo, SIGGRAPH Asia, 2009

Demo!