Human shape and pose recognition

Hugo Abreu & Anita Durr

Artificial Vision

January 14, 2020

Table of contents

Introduction

What is it? What is it used for?

Pose / shape recognition

- Image of the human body
- Information on its spatial position / pose / morphology

Introduction

What is it? What is it used for?

Pose / shape recognition

- Image of the human body
- Information on its spatial position / pose / morphology

Introduction

What is it? What is it used for?

Applications

- action recognition
- animation
- gaming

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

Pose and Shape: how to define them?

Pose

- Human skeleton
- Joints
- constraints on movement
- Proportions / height

Shape

- General form / appearance of a person
- Weight, clothing, etc...

In Practice

Inputs

2D images

- *in-the-wild* images
- occlusion, challengin viewpoint, crowded scenes, etc...

Depth Images

- Purpose built applications (real time)
- Kinect

In Practice

Inputs

2D images

- *in-the-wild* images
- occlusion, challengin viewpoint, crowded scenes, etc...

Depth Images

- Purpose built applications (real time)
- Kinect

In Practice

Inputs

2D images

- *in-the-wild* images
- occlusion, challengin viewpoint, crowded scenes, etc...

Depth Images

- Purpose built applications (real time)
- Kinect

In Practice

Inputs

2D images

- *in-the-wild* images
- occlusion, challengin viewpoint, crowded scenes, etc...

Depth Images

- Purpose built applications (real time)
- Kinect

In Practice

Outputs

Joints / Keypoints

- Standard approach
- Ideal for MoCap (Motion Capture)

Dense pose mapping

- Projection of the body parts on the image
- difficult to define

In Practice

Outputs

Joints / Keypoints

- Standard approach
- Ideal for MoCap (Motion Capture)

Dense pose mapping

- Projection of the body parts on the image
- difficult to define

In Practice

Outputs

Joints / Keypoints

- Standard approach
- Ideal for MoCap (Motion Capture)

Dense pose mapping

- Projection of the body parts on the image
- difficult to define

In Practice

Outputs

Joints / Keypoints

- Standard approach
- Ideal for MoCap (Motion Capture)

Dense pose mapping

- Projection of the body parts on the image
- difficult to define

In Practice

Outputs

Mesh Representation

- Full 3D model
- SMPL: Skinned Multi-Person Linear mode
- Anatomically correct model

In Practice

Outputs

Mesh Representation

- Full 3D model
- SMPL: Skinned Multi-Person Linear mode
- Anatomically correct model

In Practice

Outputs

Mesh Representation

- Full 3D model
- SMPL: Skinned Multi-Person Linear mode
- Anatomically correct model

Table of contents

Engineered Datasets

MoCap

MoCap

- CMU. Videos associated with 3D keypoints.

Hand-labeled datasets

LSP / MPII / COCO

Mesh Representation

- 2D images
- 2D keypoint annotations
- Hand-labeled (small datasets, LSP: 2K)

Hand-labeled datasets

LSP / MPII / COCO

Mesh Representation

- 2D images
- 2D keypoint annotations
- Hand-labeled (small datasets, LSP: 2K)

Hand-labeled datasets

LSP / MPII / COCO

Mesh Representation

- 2D images
- 2D keypoint annotations
- Hand-labeled (small datasets, LSP: 2K)

Hand-labeled datasets

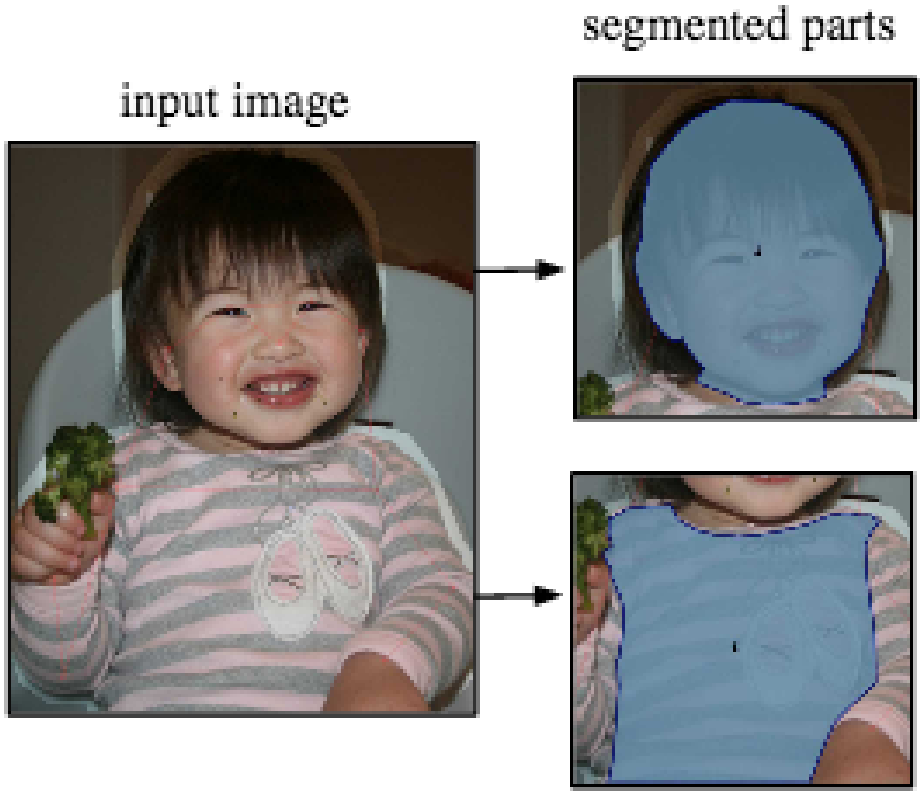
DensePose COCO. by Güler, Neverova, and Kokkinos, 2018

DensePose COCO

- establish ground truth correspondences between the SMPL model and persons appearing in the COCO dataset
- manually-collected (Facebook team)
- 50K humans, 5 million annotated image/surface pairs

Hand-labeled datasets

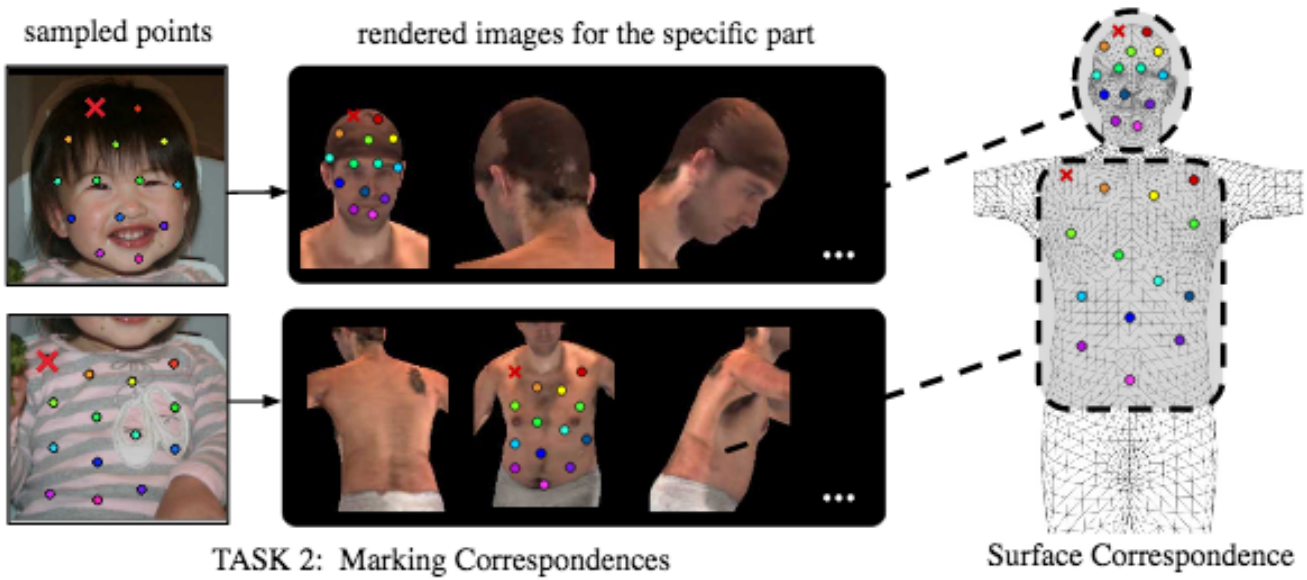
DensePose COCO : annotation system. by Güler, Neverova, and Kokkinos, 2018



TASK 1: Part Segmentation

Hand-labeled datasets

DensePose COCO : annotation system. by Güler, Neverova, and Kokkinos, 2018



Synthetized dataset creation

Varol et al., 2017

SURREAL

- Synthetic hUmans foR REAL tasks
- 6 million frames

Synthetized dataset creation

Varol et al., 2017

SURREAL

- Synthetic hUmans foR REAL tasks
- 6 million frames

Synthetized dataset creation

Varol et al., 2017

SURREAL

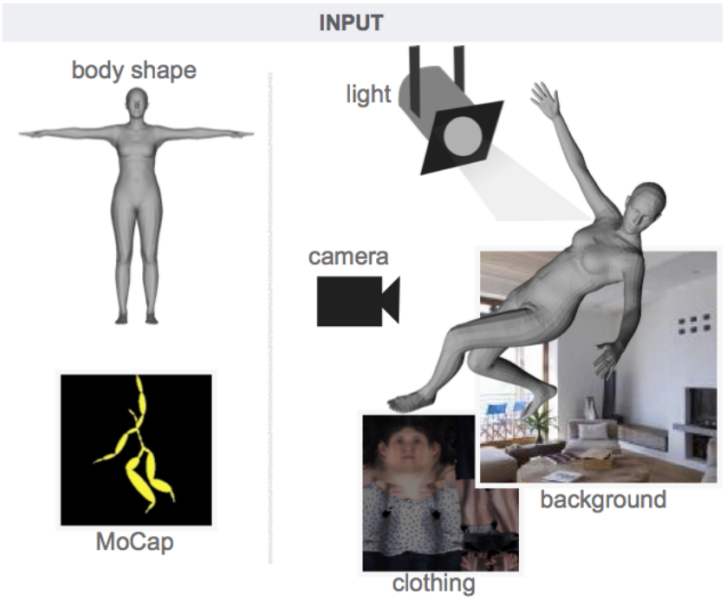
- Synthetic hUmans foR REAL tasks
- 6 million frames

Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

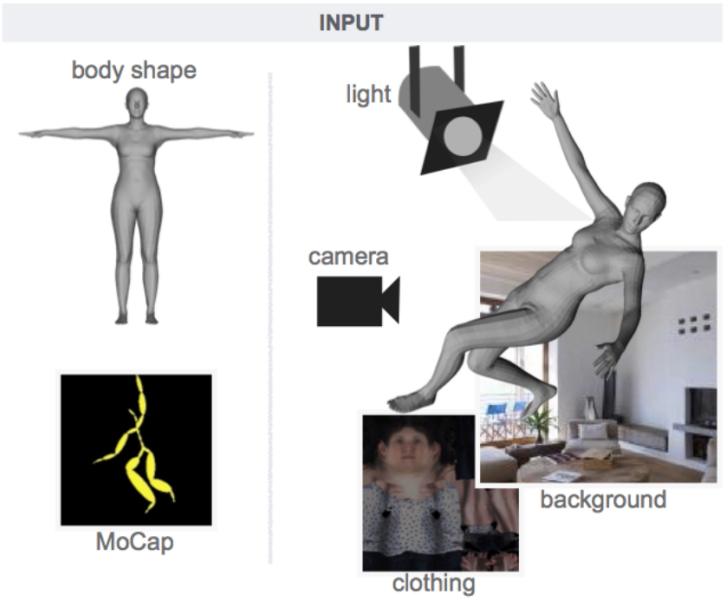


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

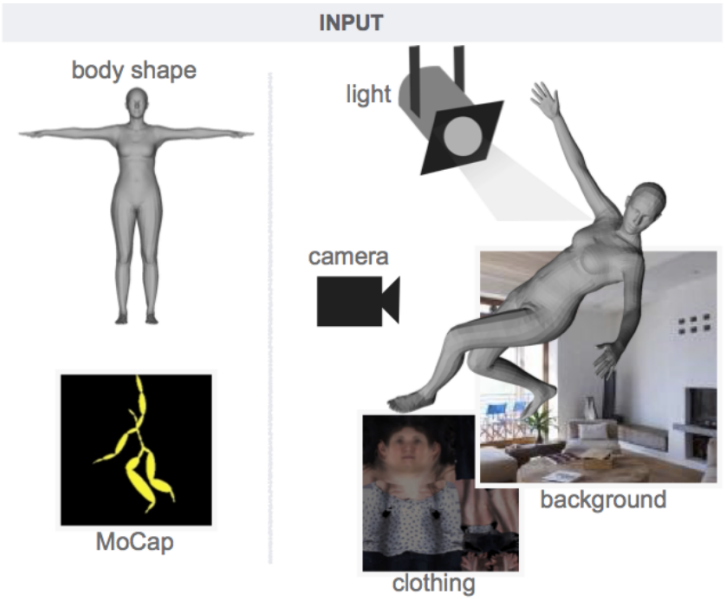


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

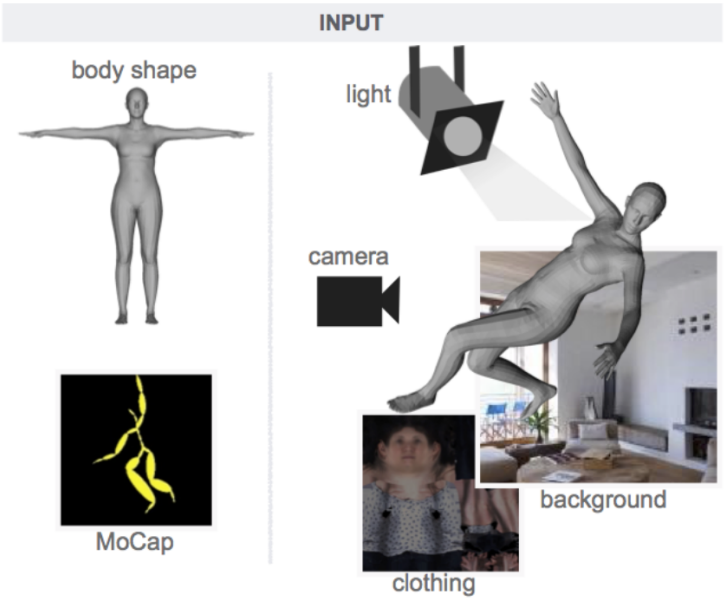


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

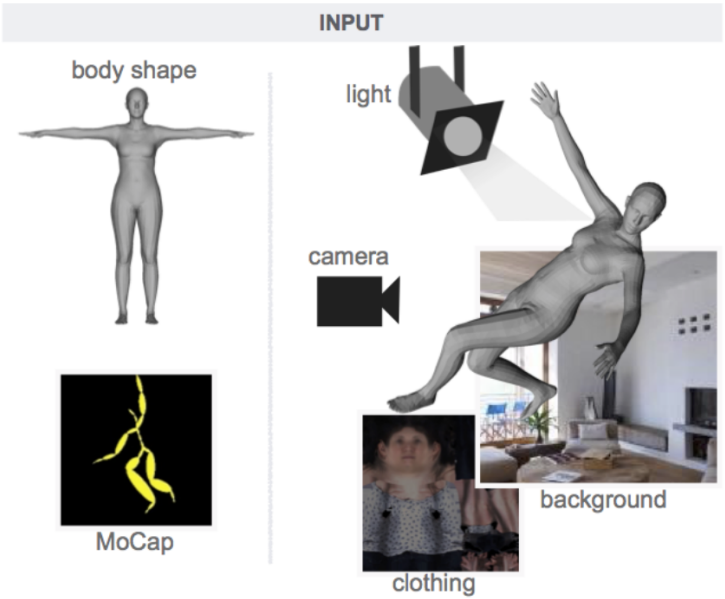


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

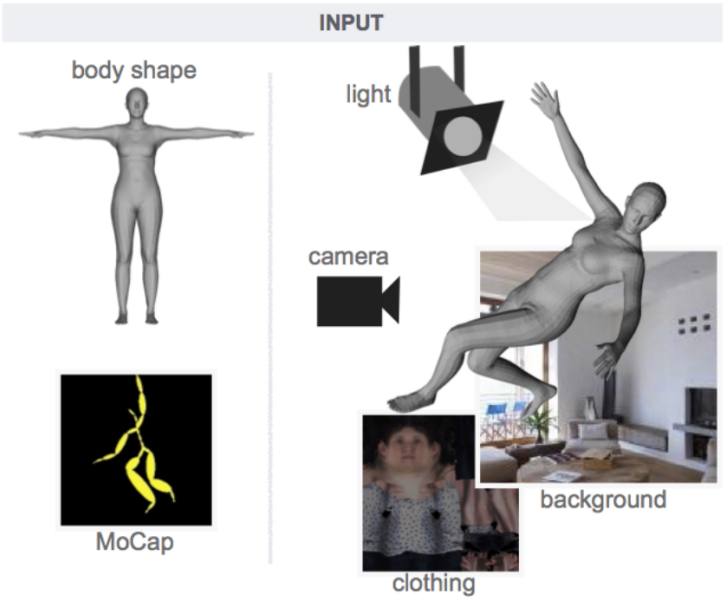


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)

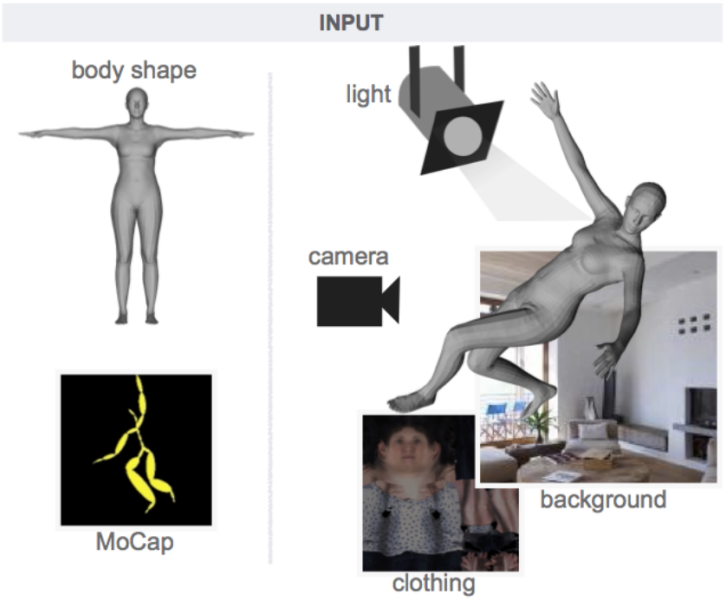


Synthetized dataset creation

Varol et al., 2017

Pose / shape recognition

- Model (Mosh + SMPL)
- Shape (CAESAR)
- Pose (MoCap)
- Texture / clothes
- Light
- Camera
- Background (LSUN)



Synthesized dataset creation

Varol et al., 2017

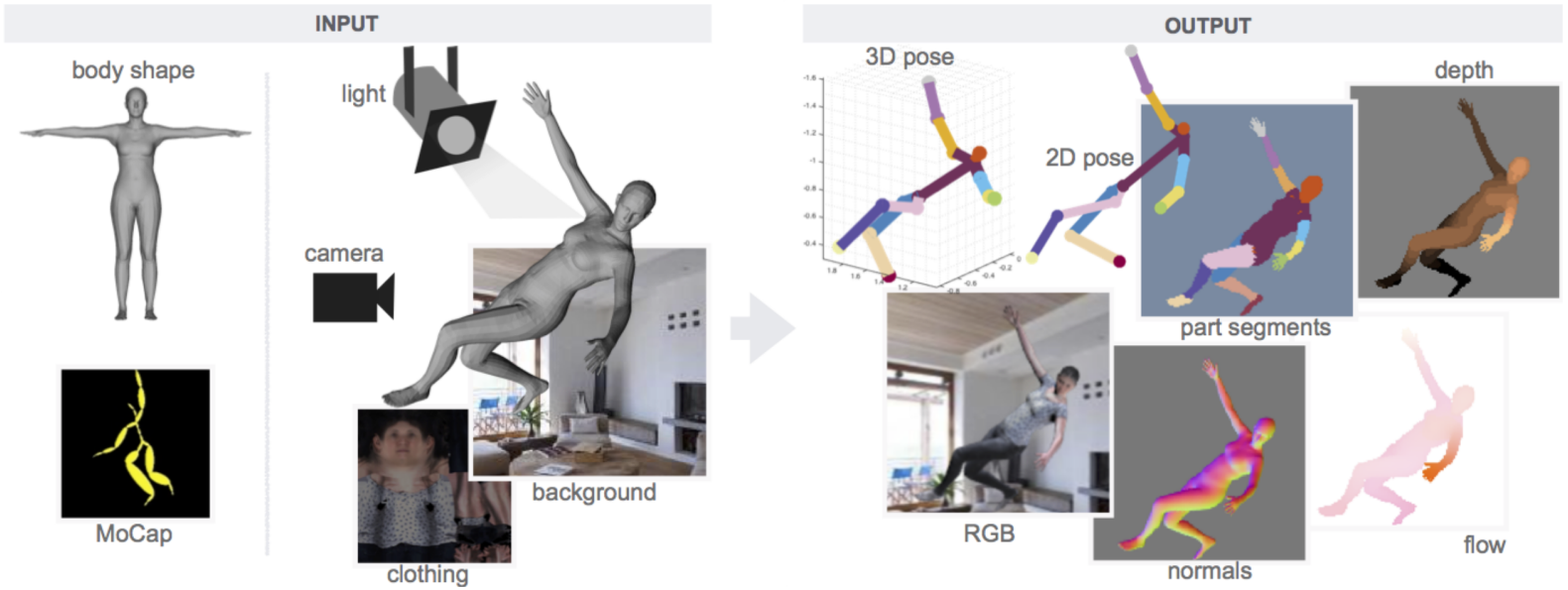


Table of contents

Iterative Error Feedback

Carreira et al., 2016

- a generic framework for modeling rich structure in both input and output space
- standard ConvNet with a simple feedback connection
- focus on the correction the network should make on the estimation and not the estimation itself

Iterative Error Feedback

Carreira et al., 2016

- a generic framework for modeling rich structure in both input and output space
- standard ConvNet with a simple feedback connection
- focus on the correction the network should make on the estimation and not the estimation itself

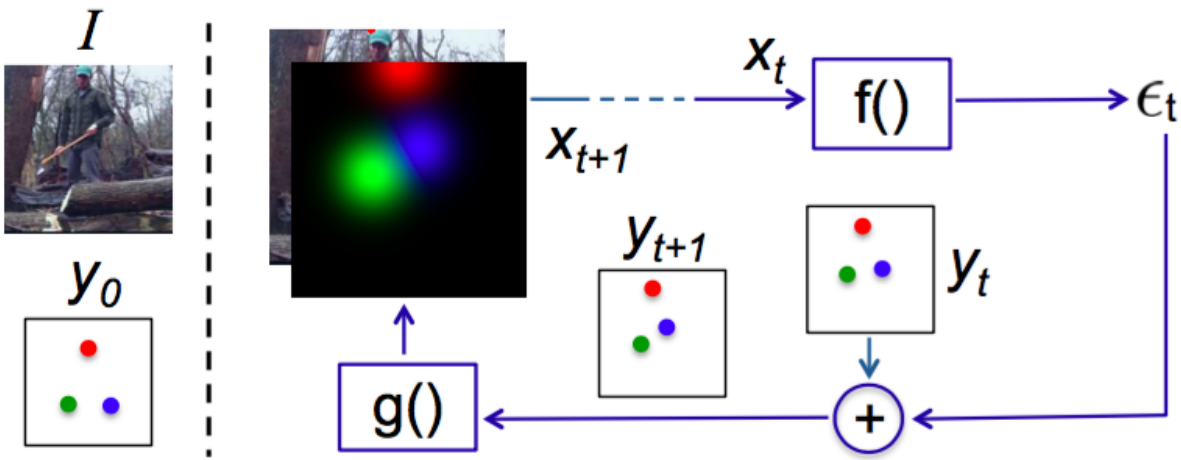
Iterative Error Feedback

Carreira et al., 2016

- a generic framework for modeling rich structure in both input and output space
- standard ConvNet with a simple feedback connection
- focus on the correction the network should make on the estimation and not the estimation itself

Iterative Error Feedback

Architecture



- $\epsilon_t = f(x_t)$ predicted error
- $y_{t+1} = y_t + \epsilon_t$ corrected estimation
- $x_{t+1} = I \oplus g(y_{t+1})$

Human Mesh Recovery

Kanazawa et al., 2018

Task

reconstruct a full 3D mesh of a human body from a single RGB image

Problems

- *in-the-wild* images with 2D annotations but not 3D
- depth ambiguity : same 2D projection for different 3D configurations
- regression rotation matrices is a challenge

Human Mesh Recovery

Kanazawa et al., 2018

Task

reconstruct a full 3D mesh of a human body from a single RGB image

Problems

- *in-the-wild* images with 2D annotations but not 3D
- depth ambiguity : same 2D projection for different 3D configurations
- regression rotation matrices is a challenge

Human Mesh Recovery

Kanazawa et al., 2018

Task

reconstruct a full 3D mesh of a human body from a single RGB image

Problems

- *in-the-wild* images with 2D annotations but not 3D
- depth ambiguity : same 2D projection for different 3D configurations
- regression rotation matrices is a challenge

Human Mesh Recovery

Kanazawa et al., 2018

Task

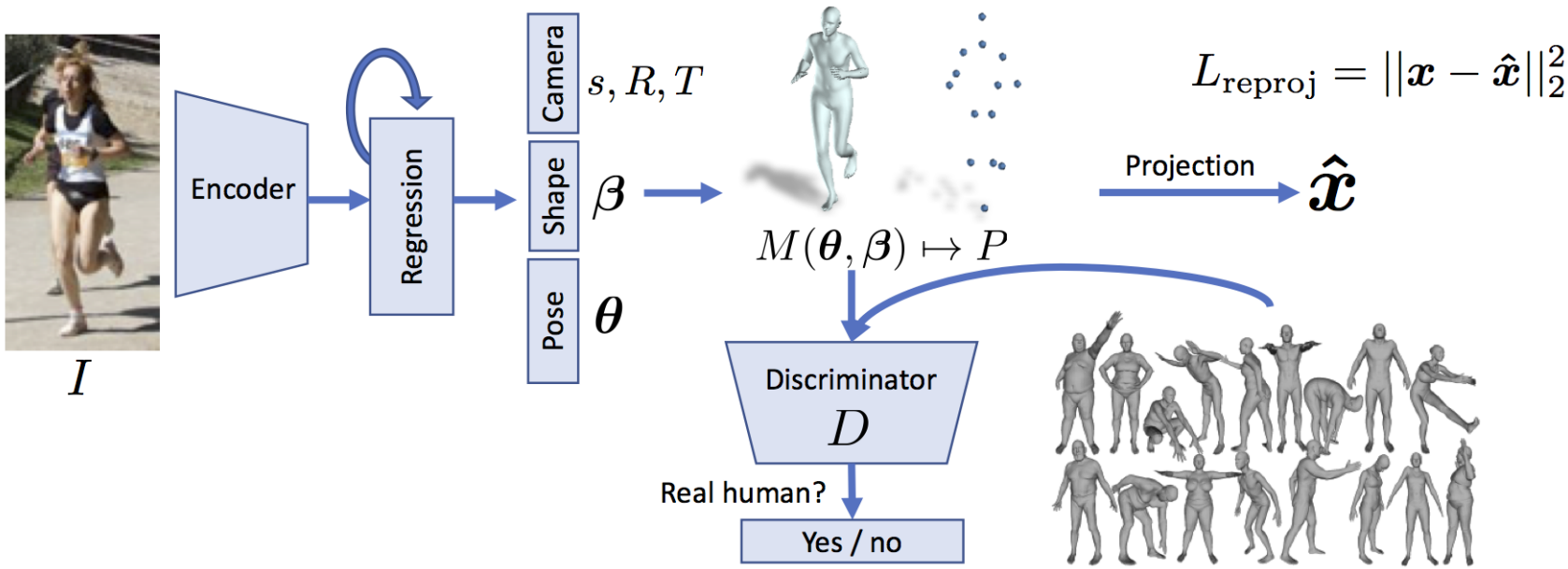
reconstruct a full 3D mesh of a human body from a single RGB image

Problems

- *in-the-wild* images with 2D annotations but not 3D
- depth ambiguity : same 2D projection for different 3D configurations
- regression rotation matrices is a challenge

Human Mesh Recovery

Kanazawa et al., 2018



Human Mesh Recovery

Kanazawa et al., 2018

- real time
- no prior assumption (on angle joint limit for example)
- adapted to *in-the-wild* images with only 2D annotations

Pose Recognition from Single Depth Images

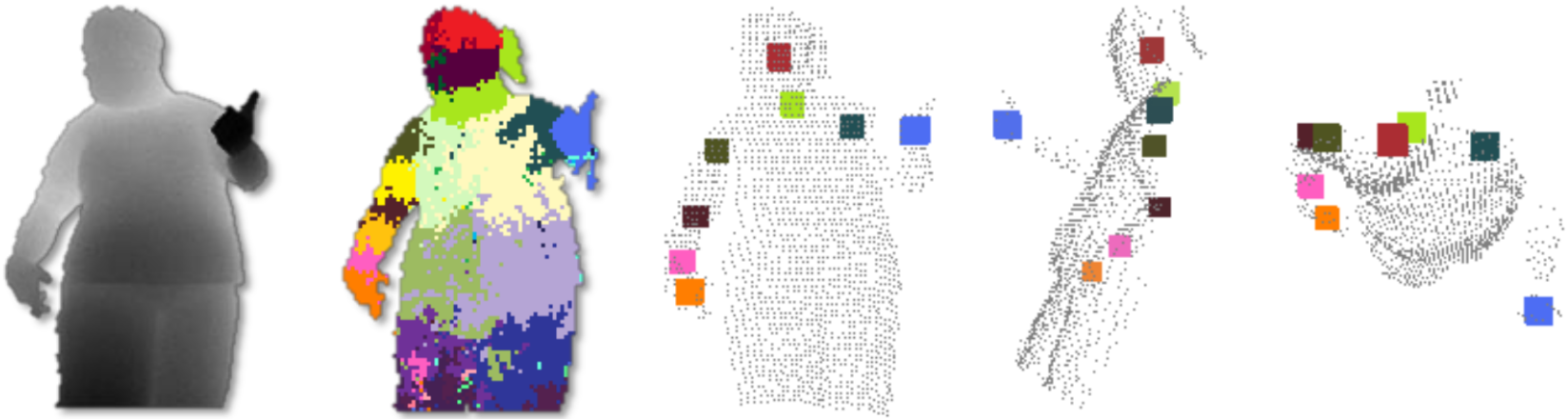
Shotton et al., 2013

Pose Recognition from Single Depth Images

Shotton et al., 2013

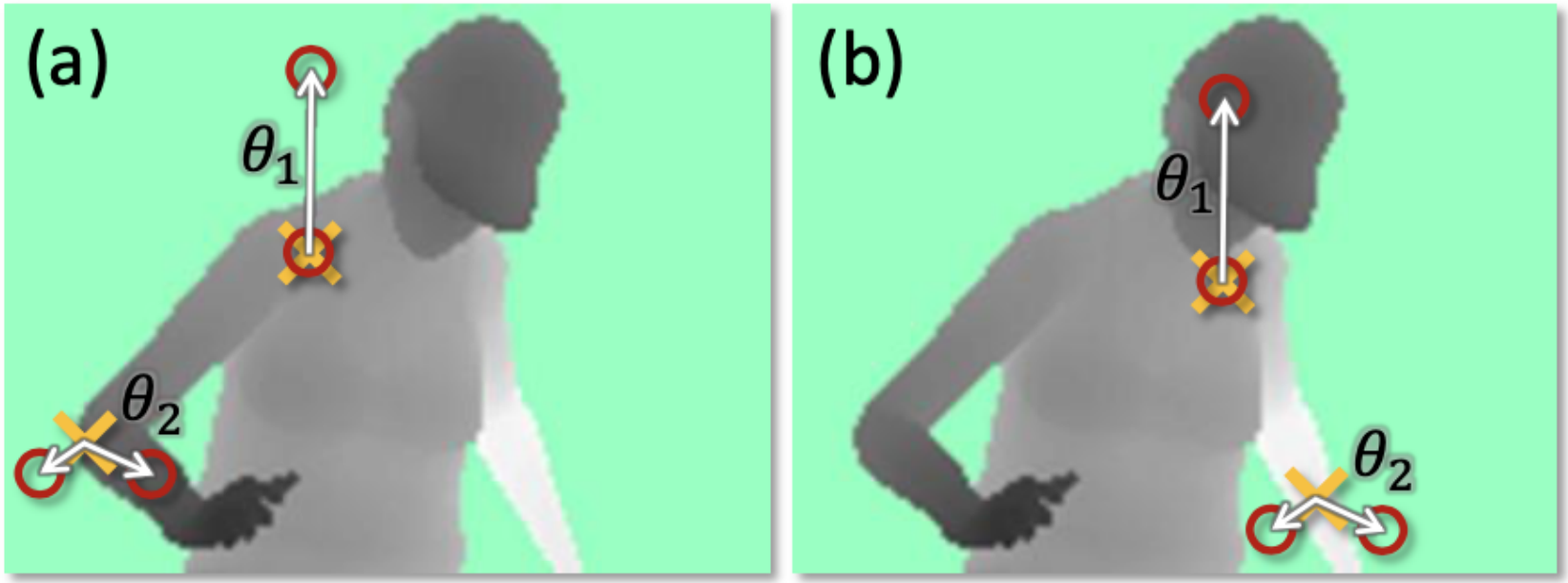
Pose Recognition from Single Depth Images

Shotton et al., 2013



Pose Recognition from Single Depth Images

Shotton et al., 2013



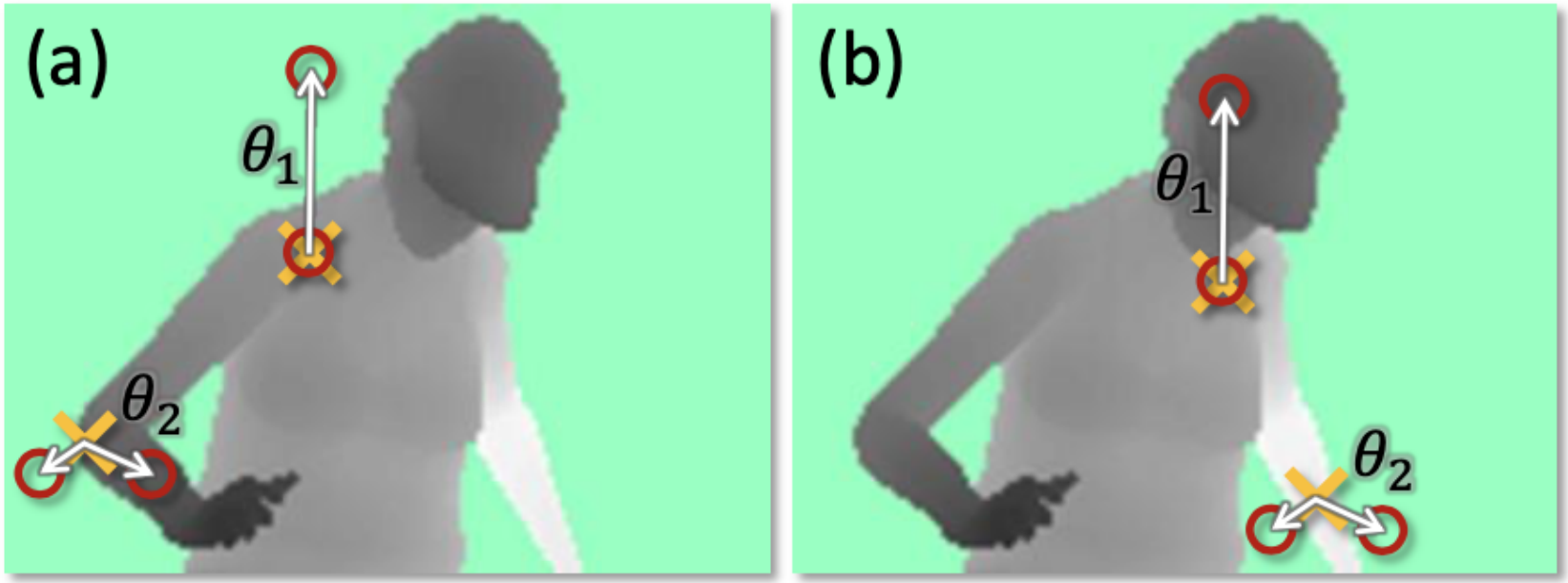
Pose Recognition from Single Depth Images

Shotton et al., 2013

$$f_\theta(I, \mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right) , \quad (1)$$

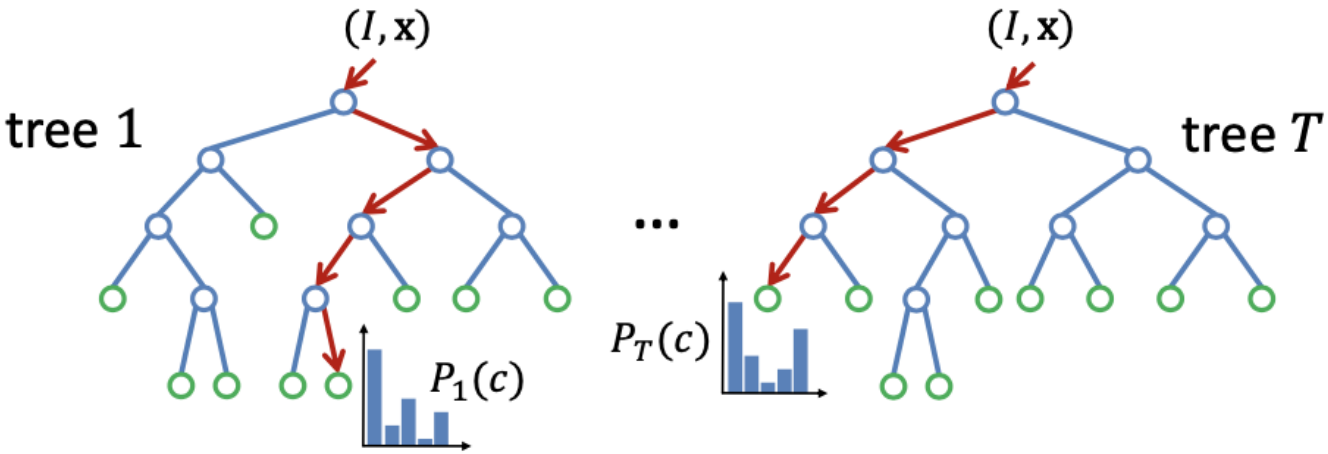
Pose Recognition from Single Depth Images

Shotton et al., 2013



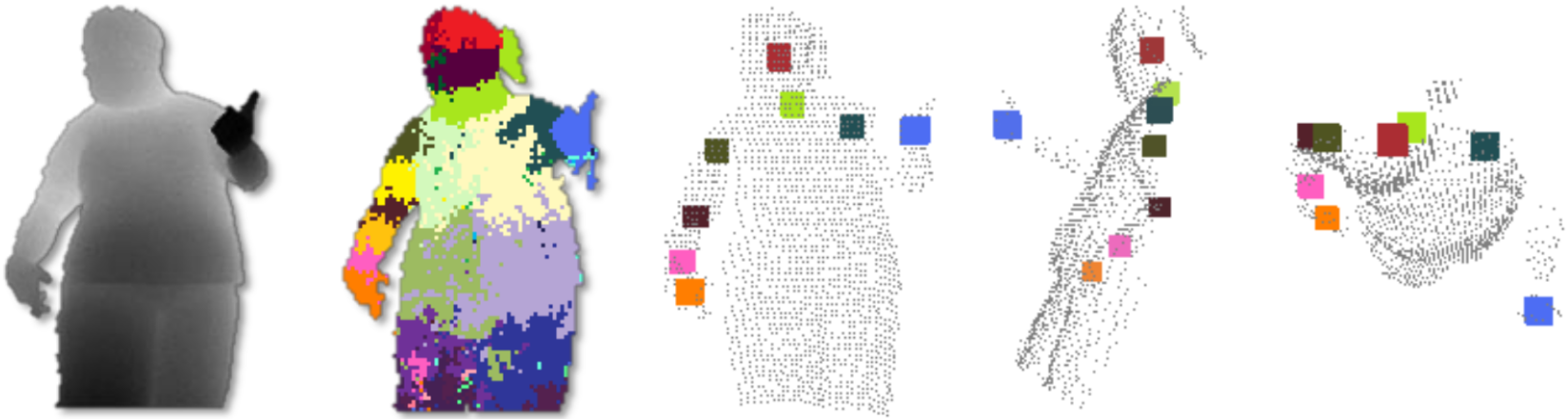
Pose Recognition from Single Depth Images

Shotton et al., 2013



Pose Recognition from Single Depth Images

Shotton et al., 2013



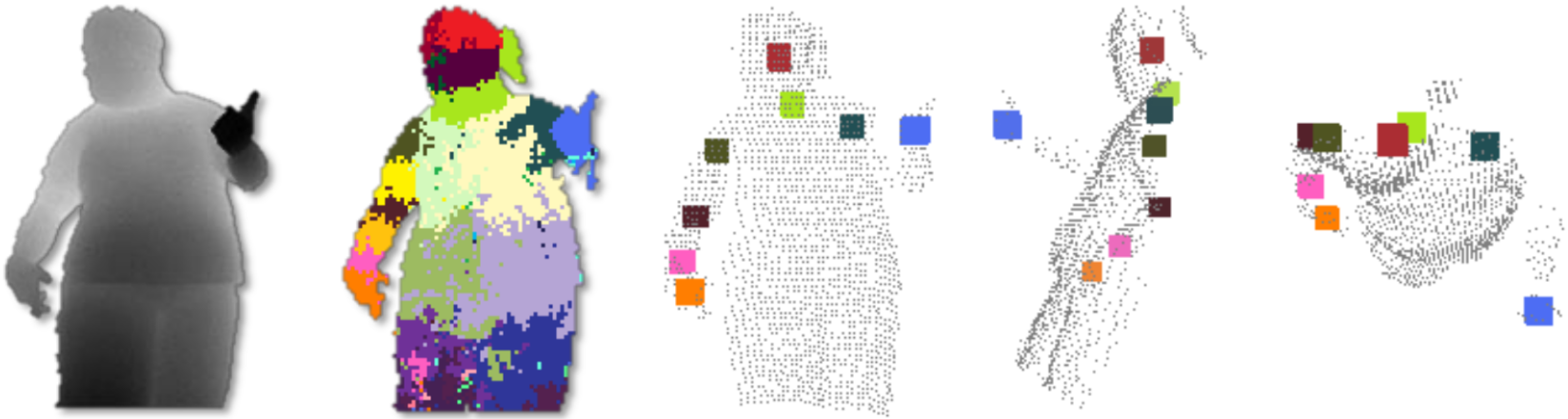
Pose Recognition from Single Depth Images

Shotton et al., 2013

$$f_c(\hat{\mathbf{x}}) \propto \sum_{i=1}^{N} w_{ic} \exp\left(-\left\|\frac{\hat{\mathbf{x}} - \hat{\mathbf{x}}_i}{b_c}\right\|^2\right), \qquad (7)$$

Pose Recognition from Single Depth Images

Shotton et al., 2013



References

Carreira, Joao et al. (2016). "Human Pose Estimation with Iterative Error Feedback". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr.2016.512. URL: http://dx.doi.org/10.1109/CVPR.2016.512.

Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos (2018). "Densepose: Dense human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306.

Kanazawa, Angjoo et al. (2018). "End-to-End Recovery of Human Shape and Pose". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/cvpr.2018.00744. URL: http://dx.doi.org/10.1109/CVPR.2018.00744.

Shotton, Jamie et al. (2013). "Real-time human pose recognition in parts from single depth images". In: *Communications of the ACM* 56.1, pp. 116–124.

Varol, Gul et al. (2017). "Learning from Synthetic Humans". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr.2017.492. URL: http://dx.doi.org/10.1109/CVPR.2017.492.