Embedding Context in Vision For Monocular Human Pose Tracking

(Defense Presentation)

Olivier St-Martin Cormier
supervised by Frank Ferrie
McGill University
March 26, 2020

Outline

- **Introduction**
  - Research Context
  - Problem Statement & Thesis
  - Contributions
- **Proposed Approach**
- **Feature Extraction**
  - Silhouettes
  - Metric
- **Pose Estimation**
- **Context Modeling**
- **Experimentation**
  - Small Scale Experiments
  - Human Experiments
- **Conclusion**
  - Key Findings
  - Future Work
  - Contributions

Introduction

Research Context

Collaborative, Human-focused, Assistive Robotics for Manufacturing
Funded by General Motors Canada as part of the NSERC CRD program

Plugfest 3 Video Screenshot

Research Interests:

- Situational Awareness
- Human Posture Tracking

Problem Statement & Thesis

Fundamental Problem:

Infer complex behavior given limited and/or noisy observable data

- **Complex behavior:** Articulated model pose
- **Observable data:** Monocular optical sensor

Thesis:

- We can regularize the incomplete data based on a set of prior models and assumptions based on the observed system's context.
- Low dimensional descriptors such as silhouettes can represent complex deformable three dimensional shapes with high number of degrees of freedom
- A computationally efficient framework to track human postures can be designed

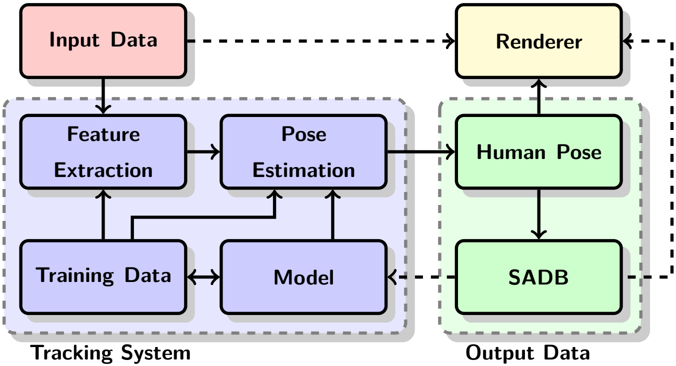
Main Contributions

- **Computational framework for the inference process**
  Hybrid particle filtering and machine learning fusion task
- **Characterization of the relation between low dimensional descriptors and complex structures**
  Survey and evaluation of feature comparion metrics
- **Representations of complex systems**
  Development of the Situational Awareness DataBase (SADB) system
- **Use of deformable 3D mesh models**
  Usage of shape model based on biomechanical data, easily extensible to 3d scans of subject
- **Complete tracking system**
  Design of a flexible tracking system capable of tracking different types of articulated models

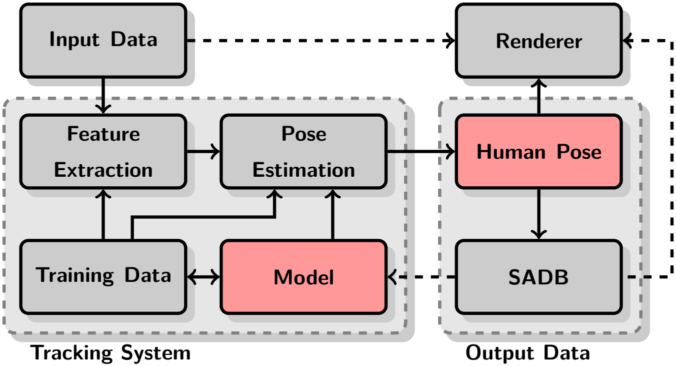
Proposed Approach

Proposed Approach

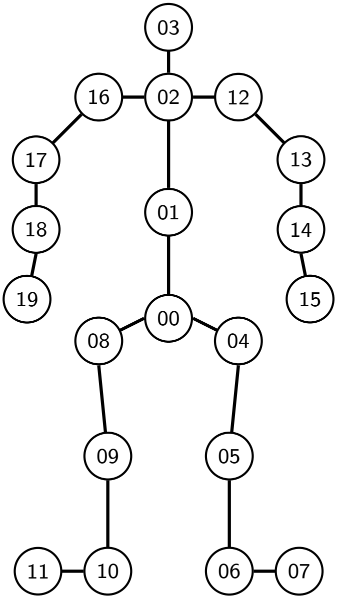
We propose to explore these topics through the development of a complete tracking framework



Model & Human Pose



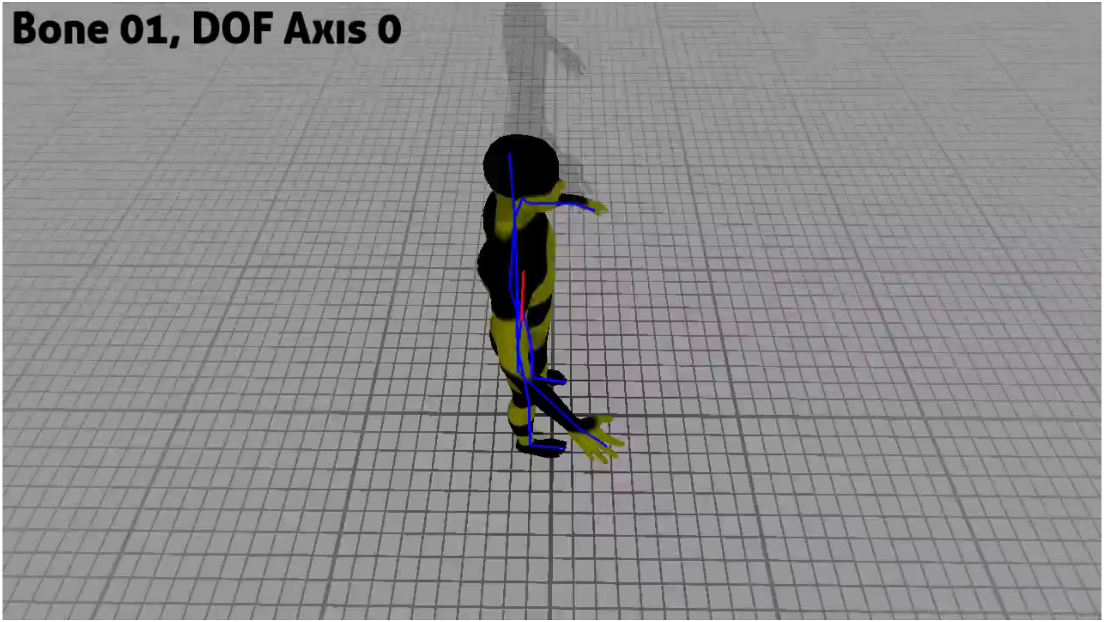
Proposed Skeletal Structure



Node	Parent	Name	DOF
00	–	Root / Pelvis	6
01	00	Abdomen	1
02	01	Thorax	3
03	02	Head	3
04	00	Right Hip	0
05	04	Right Thigh	3
06	05	Right Shin	1
07	06	Right Foot	1
08	00	Left Hip	0
09	08	Left Thigh	3
10	09	Left Shin	1
11	10	Left Foot	1
12	02	Right Shoulder	0
13	12	Right Upper Arm	3
14	13	Right Forearm	2
15	14	Right Hand	2
16	02	Left Shoulder	0
17	16	Left Upper Arm	3
18	17	Left Forearm	2
19	18	Left Hand	2
		Total DOF	37

Proposed Skeletal Structure

Degrees of Freedom Demo



Pose and Silhouette Spaces

We can define a pose vector:

$$\mathbf{P} = [\theta_0, \theta_1, \cdots, \theta_{n-1}, \theta_n], \quad \theta_i \in [0, 1], \quad n = 36$$

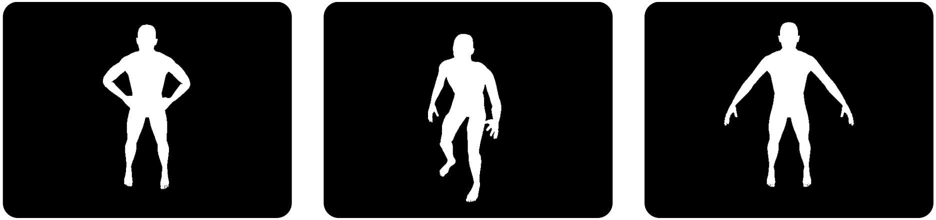
and a pose difference:

$$\mathbf{\Delta_{12}} = \mathbf{P_1} - \mathbf{P_2} = [\delta_0, \delta_1, \cdots, \delta_{n-1}, \delta_n], \quad \delta_i \in [-1, 1], \quad n = 36$$

The distance in pose space can be computed as:

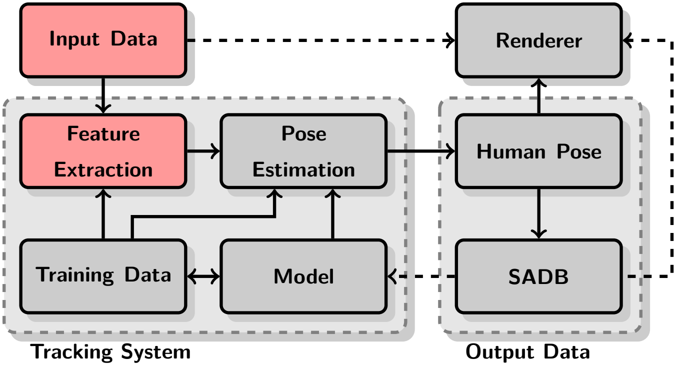
$$D(\mathbf{P_1}, \mathbf{P_2}) = D(\mathbf{P_1}, \mathbf{P_1} + \mathbf{\Delta_{21}}) = \|\mathbf{\Delta_{21}}\|$$

We can render a pose to obtain a silhouette $S = \mathcal{R}(\mathbf{P})$



The distance in silhouette space can be denoted as $D(S_1, S_2)$

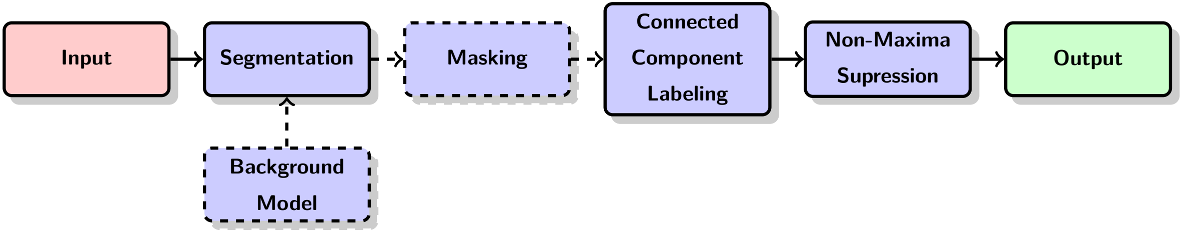
Feature Extraction



- Compress input data into a simpler descriptor
- Silhouettes encode a lot of information about the shape

Silhouette Extraction Approaches

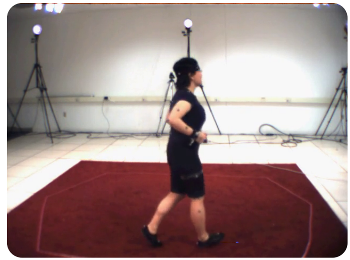
- **Gaussian Mixture Model**
  - Requires learning a background model
  - Susceptible to lighting changes or camera shake
- **Chroma Keying**
  - Requires control of the environment
  - Great segmentation
- **Deep Learning Approaches**
  - Future work



Gaussian Mixture Model - Example

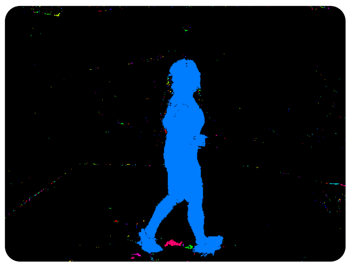
Frame 100 of Jog 2 sequence in HumanEVA

(a) Background



(b) Input

(c) Segmentation

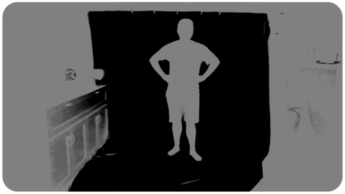


(d) Labeling

(e) Output

Chroma Keying - Example

(a) Input Image



(b) Keyed Image *

(c) Binary Mask **

(d) Masked **

(e) Labels

(f) Non-Maxima Supression

* Chroma Keying done in Natron

** Would not be necessary in a real studio

Silhouette Comparison Metrics

We review 7 metrics commonly used in the literature to find which is better at comparing human silhouettes.

We can use the following properties to compare metrics

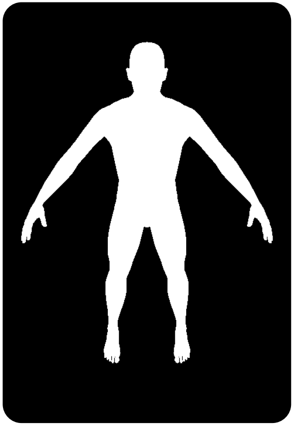
- Metric Property (Triangle inequality): $D(S_1, S_2) + D(S_2, S_3) \geq D(S_1, S_3)$
- Increases monotonically with distance
- Robustness to noise and shadows
- Correlation between $D(S_1, S_2)$ and $D(\mathbf{P_1}, \mathbf{P_2})$, given $S_n = \mathcal{R}(\mathbf{P_n})$

Key Finding:
Pixel count and Shape Contexts perform best, but Shape Contexts are more computationaly expensive

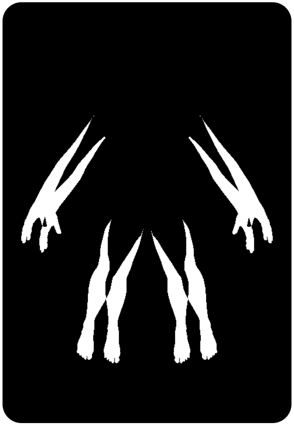
Silhouette Comparison Metrics - Pixel Count

Count the number of pixels that differ between 2 silhouettes

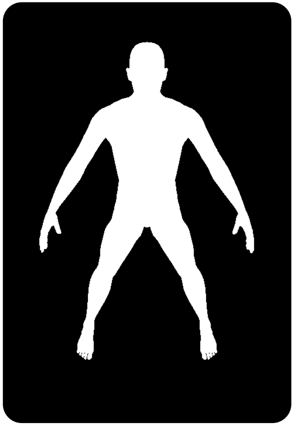
$$D(S_1, S_2) = \sum_{x=0}^{n} \sum_{y=0}^{m} S_1[x, y] \oplus S_2[x, y]$$



S_1

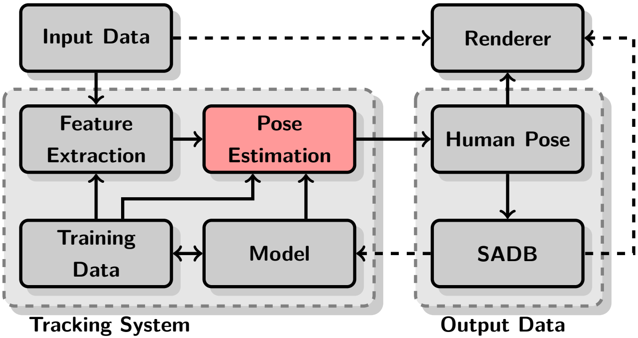


$S_1 \oplus S_2$



S_2

Particle Filtering



Mathematical Formulation

$$\underbrace{p(x_t|y_{1:t})}_{\text{Posterior at } t} \propto \underbrace{p(y_t|x_t)}_{\text{Likelihood}} \int \underbrace{\overbrace{p(x_t|x_{t-1})}^{\text{Temporal Prior}} \overbrace{p(x_{t-1}|y_{1:t-1})}^{\text{Posterior at } t-1} dx_{t-1}}_{\text{Predictive Density}}$$

- State vector: x_t
- Sequence of measurements: $y_{1:t}$

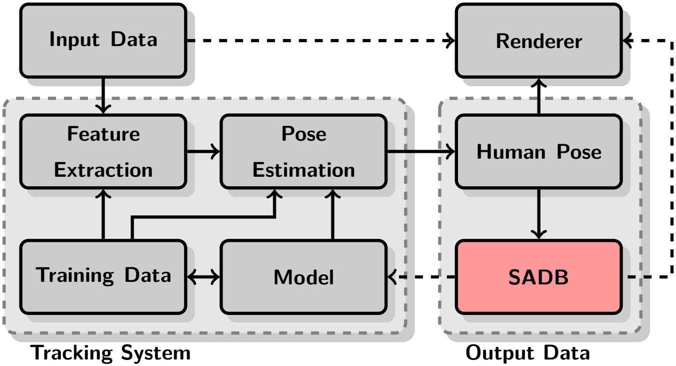
Practical Formulation

- **Initialization**
  Multiple options (Gradient Descent, CNN, Random Sampling, etc)
- **Particle Weighting**
  Comparing silhouettes using a similarity metric
- **Result Computation**

$$\mathbf{P_t} = \sum_{k=0}^{N-1} w_t^k x_t^k$$

- **Particle Propagation**
  Predict changes in the system
  - Gaussian Diffusion
  - First Order Motion Model
  - Second Order Motion Model
  - Learned Model
  - Physics Simulation
  - Hybrid Models
- **Particle Resampling**
  Keep particles focused around simulation

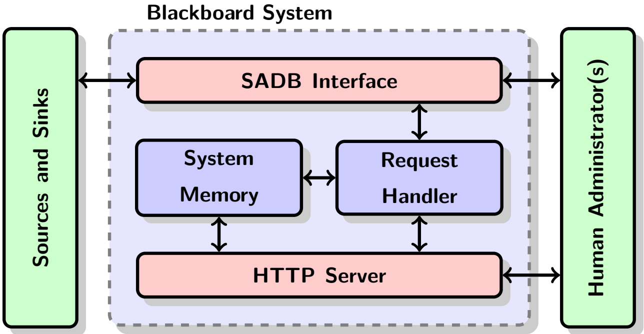
Situational Awareness Database



Knowledge Representation and Organization

Proposed system block diagram

Situational Awareness DataBase (SADB)

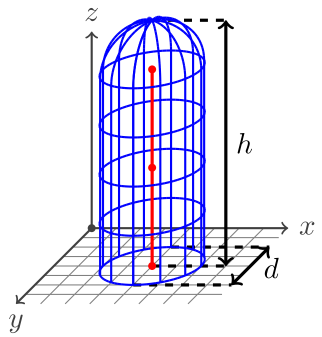


Key Features

- NoSQL API interface to data
- Data is organized as objects
- Each object consists of a time-coded series of values
- Values available on an any-time basis via interpolation
- Objects are classified in categories
- System queries in the form of set operations

Small Scale Experiments

Small Scale Model



Synthetic data allows us to record:

- **Image Data**: input and output
- **Pose Error**: distance in pose space
- **Visual Acuity**: silhouette similarity metric
- **End Effector Position**: position of the end of the model
- **End Effector Error**: distance between ground truth and tracker's output
- **Particle Poses**: pose vector of each particle
- **Particle Statistics**: dispersion, weight distribution, etc

Experiment Areas

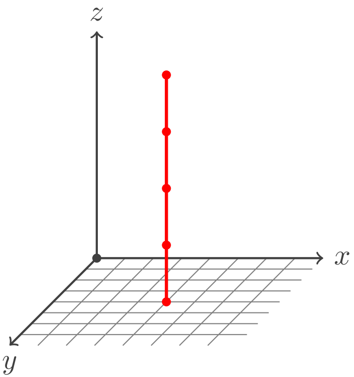
The goal is to find the best set of parameters before applying the tracker to human sequences

- Particle Propagation Models
- Resampling Method
- Resampling Amounts
- Initialization Strategies
- Metric Behavior Over Time
- Effect of Occlusion
- Effect of Ambiguity
- Effect of Model Complexity

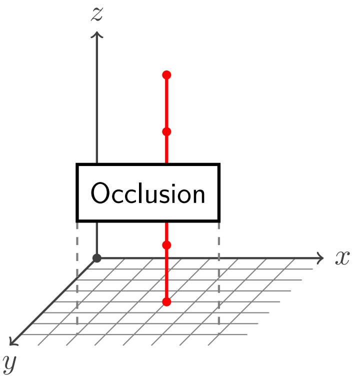
Experiment Configurations

We can experiment with parameters in different conditions

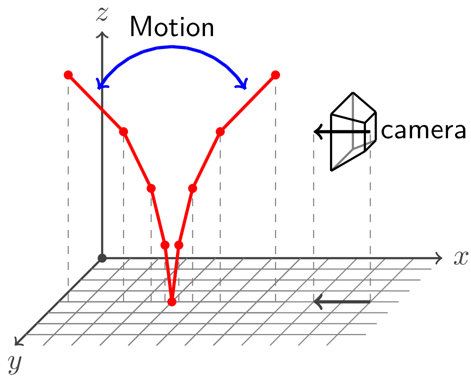
Basic Configuration



Occlusion Simulation



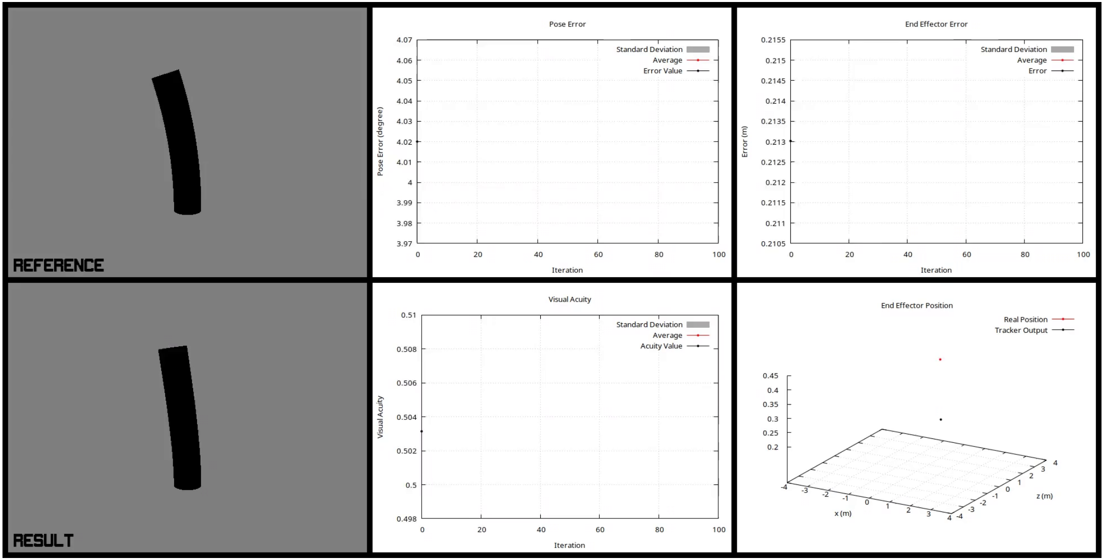
Ambiguous Motion



Small Scale Experiment Example

Randomly selected video

Simple motion, Gaussian Prior, 100 Particles, 36 DOF



flash-

Human Experiments

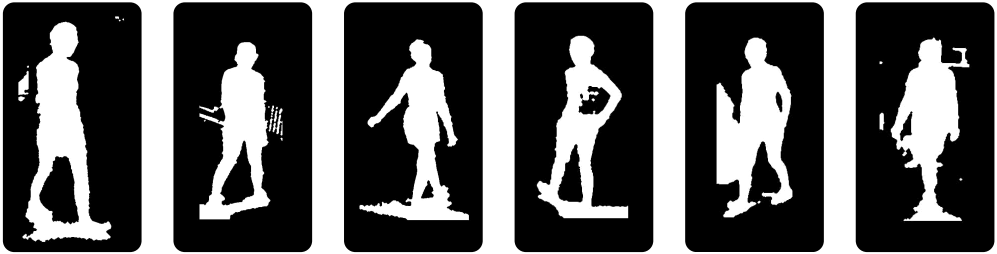
Experiment Cases

The goal is to determine if the tracking approach generalizes to full human pose tracking.

Testing Datasets:

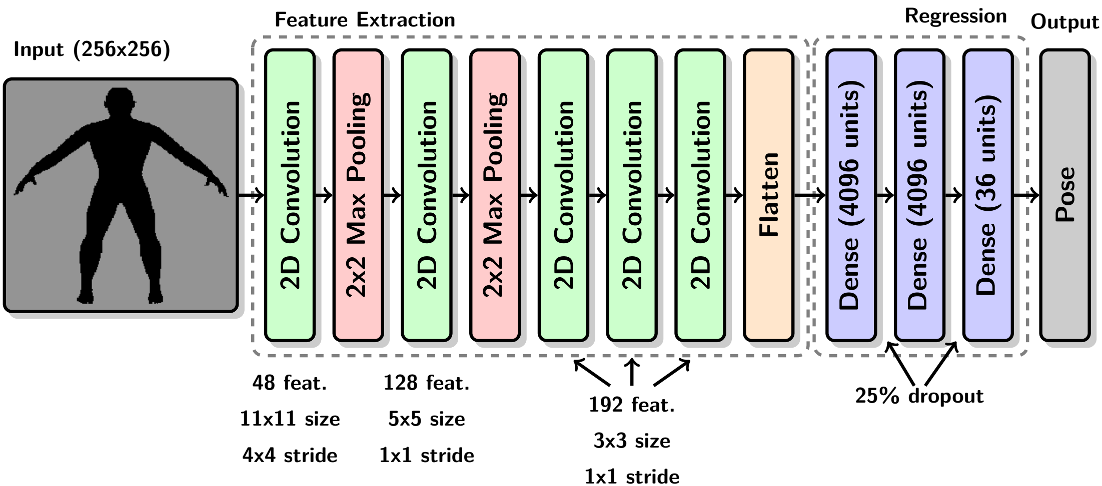
- Synthetic Data
- HumanEVA Dataset
- Human3.6M Dataset
- Chroma-Keyed Sequence

Problems with Human3.6M:



Network Architecture

Architecture based on “Deeppose” (CVPR 2014)



* all activation functions are rectified linear units (ReLU)

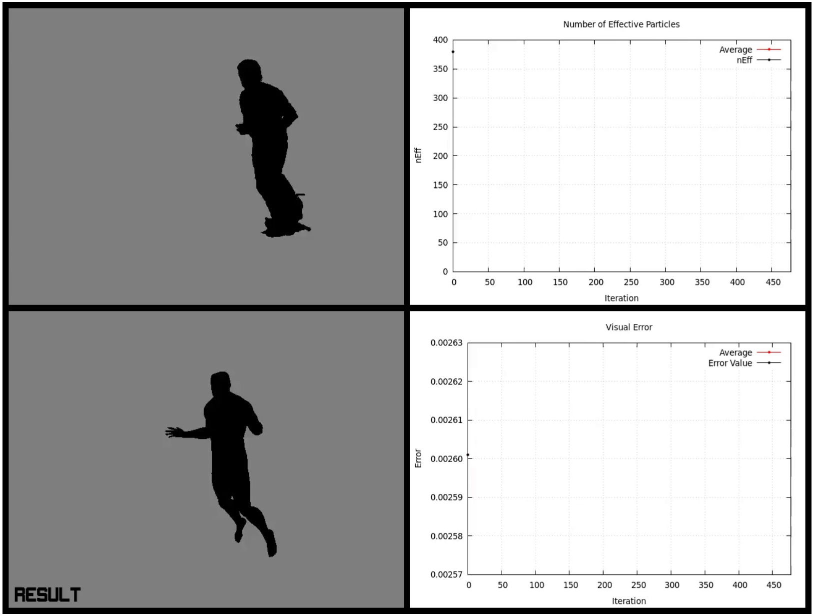
Training Process

- Use our generative model to generate training data
- Use multiple mesh model to generalize to different shape models

HumanEVA Dataset Experiment Example

Randomly selected video for HumanEVA results

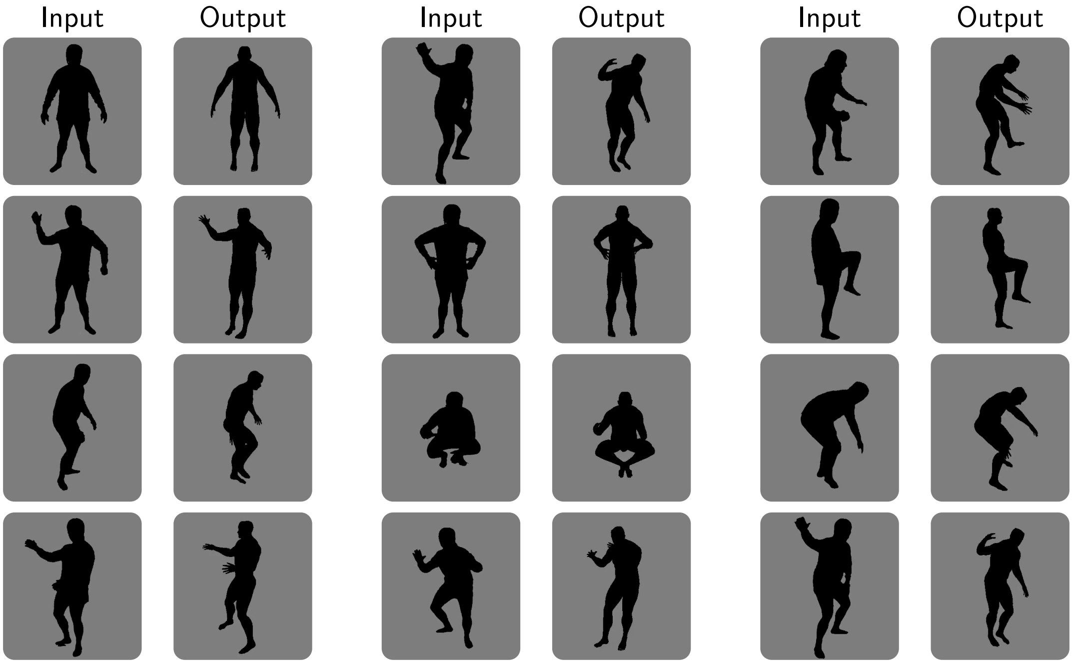
Subject 2, Jog 2 Sequence, Camera 3, 36 DOF



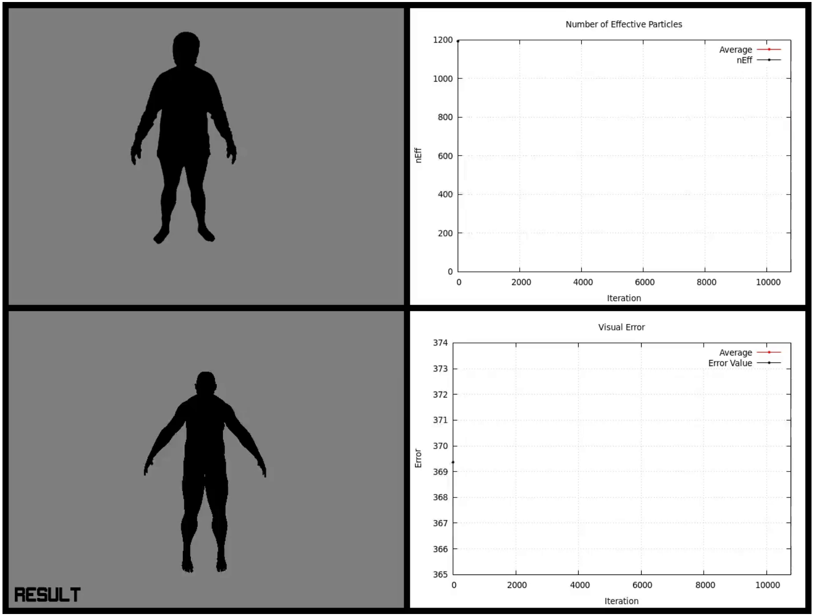
flashvars=source=Videos/HumanDemo-

Human Experiment

Sample frames from the chroma-keyed sequence



Chroma-Keyed Human Experiment



Conclusion

Key Findings

- Low dimensionality descriptors correlate to high dimensional changes in pose
- The use of a synthetic generative model can be used as a
  - basis for controlled experimentation
  - prior model for a particle filter
  - source of training data for a CNN
- We can bootstrap the development of a CNN by providing it with a descriptor instead of raw images
- The limiting factor of the overall is the accuracy of the silhouette extraction algorithm

Future Work

- Evaluate better silhouette segmentation algorithms
- Apply data augmentation techniques to make CNN more robust to errors in silhouettes
- Extend approach to handle multiple cameras
- Track additional parameters (height, age, sex, etc)
- Replace virtual silhouette renderer with a CNN

Generated silhouette

Final Thoughts - Research Impacts

- **To the best of our knowledge, the research presented here is the first time 3d human pose has been tracked with such a high number of DOFs from a monocular camera configuration.**
- **Situational awareness modeling is beneficial for both humans and machines.**
  - Humans: provides transparency and increases trust in the system by exposing inner workings
  - Machine: provides cues to interpret recorded data in a more meaningful way
- **Combining classical vision and deep learning approaches addresses shortcomings of both.**

Thank You