Human Attributes from 3D Pose Tracking

MSc. Presentation

October 7, 2011

Micha Livne

Supervisor: David Fleet

University of Toronto, Department of Computer Science

A joint research with:

Leonid Sigal (Disney Research) , Nikolaus Troje (Queen's University)
Perception of Biological Motion

- Detecting animate motion
- Deriving structure from motion
- Recognizing identity from motion
- Recognizing style from motion
- Interpreting action from motion
- Extracting attributes
Perception of Biological Motion

Gunnar Johansson
1911-1998
Biometrics and Gait Analysis

• Use of 3D models is not common.

• Use of articulated models is limited to 2D silhouettes.
Action Recognition

• Holistic space-time features, local interest points.
  - E.g. SURF, HOG, HOF.

• 3D pose estimation is considered too noisy to support activity analysis.

Biomotion Lab (Queen's University)
3D Pose Tracking

- **HUMANeva**: 3D Euclidean distance error as benchmark.
- We suggest **attribute inference** as complementary benchmark.

Biomotion Lab
(Queen's University)
Our Goal

mocap → Feature Extraction → Attribute Inference

Video Tracking → Feature Extraction
Our Goal

Attribute
Inference
Our Goal

Socially Meaningful Attribute Inference:
- Gender
- Weight
- Age
- Mood
Mocap Data: Dmocap

• Mocap - 115 walkers with 4 samples per subject (Biomonition Lab) with known gender, age, weight, etc.

• Perceptual - subjective attributes.
Human Subject Ratings

1. Choose any attribute you are interested in. You then have to indicate names for the two extreme states. For instance, the attribute "gender" can run from "very masculine" to "very feminine", or the attribute "weight" can run from "light" to "heavy". Please indicate also your own age, sex and the country you are from. Only if you have completed the dialog you can start the experiment.

2. Once you click on the Start Button You will be presented with a series of point-light walkers. Look at them and try to rate them according to the attribute you've chosen. You do this by clicking on one of six buttons, which will appear on the right hand side of the screen. A counter keeps track of the number of the ratings.

3. You have to rate at least 20 walkers but it is strongly recommended to rate 50 to 100 or even more to get reasonable results. If you are done, click the FINISH button. The slider that will appear on the top of the screen can be used to change the motion of the walker, based on your ratings. At the same time the data you produced are sent to the BioMotionLab for further evaluation.

Thanks so much for participating.

Name: Micha  Age: 21-30  Sex: m
Country: Canada
Attribute: Intoxication  from: Drunk  to: Sober

Biomotion Lab (Queen's University)
Human Subject Ratings

Biomotion Lab (Queen's University)
Mocap Attribute Data

Observers #  Ratings #

126  8093

694  44657

67  4380
Motion Alignment

• Each motion is a pose trajectory: vector of 15 3D joints position per time step.

• **Direction alignment:** the X-axis coincides with the direction of locomotion.

• **Centring:** "slow" trend is removed in forward and lateral directions.

• **Phase alignment:** Motion phase is aligned.
Fourier Series Representation

• Walking is cyclic: motion is effectively represented by Fourier series.

• Two harmonics are sufficient.

• Each motion is represented by a 226D vector:

\[ 15 \text{ Joints} \times 3D \times 5 \text{ Fourier Coefs} + \text{freq} \]
Fundamental Frequency Estimation

- Fourier model:

\[ m^j(t; \omega, a^j) = r(t) \sum_{h=-2}^{2} e^{-i\omega ht} a^j_h \]

- LS fundamental frequency is given by model fitting:

\[
\arg \min_{\omega} \left( \sum_{j=1}^{45} \min_{a^j} \sum_{t=1}^{T} |f^j(t) - m^j(t; \omega, a^j)|^2 \right)
\]
Fundamental Frequency Estimation

Video Tracking

Model
Subspace Motion Model

• **Dimensionality reduction**: PCA on the Fourier model coefficients \( \{ m_i \}_{i=1}^N \)

\[
c_j = B^T (m_j - \bar{m})
\]

• \( B \equiv [b_1, \ldots, b_K] \) is the subspace basis.

• For \( K = 16 \) we captures well over 90% of the model variance.
Attribute Inference From Motion Capture

- Classification Rate vs. Number of PCs
- MSE Error (kg) vs. Number of PCs
- Classification Rate vs. Sequence Length (gaits)
Subspace Coefficients: \textit{Dmocap}

Gender

Weight
<table>
<thead>
<tr>
<th>Sagittal 2D Pose</th>
<th>Frontal 2D Pose</th>
<th>Lower Body</th>
<th>Upper Body</th>
<th>All Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.86</td>
<td>0.84</td>
<td>0.75</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>0.83</td>
<td>0.76</td>
<td>0.77</td>
<td>0.79</td>
<td>0.83</td>
</tr>
<tr>
<td>0.75</td>
<td>0.70</td>
<td>0.76</td>
<td>0.77</td>
<td>0.82</td>
</tr>
</tbody>
</table>

3D Model
- Normalized
- Motion Only

Data Sub-sets
- Different
- Models
- Dmocap
- Results
## Different Models: Dmocap Results

<table>
<thead>
<tr>
<th>Weight (RMSE)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Markers</strong></td>
<td>5.44</td>
<td>9.78</td>
<td>10.86</td>
</tr>
<tr>
<td><strong>Upper Body</strong></td>
<td>5.91</td>
<td>10.19</td>
<td>11.14</td>
</tr>
<tr>
<td><strong>Lower Body</strong></td>
<td>6.63</td>
<td>9.52</td>
<td>12.49</td>
</tr>
<tr>
<td><strong>Frontal 2D Pose</strong></td>
<td>5.59</td>
<td>9.79</td>
<td>10.82</td>
</tr>
<tr>
<td><strong>Sagittal 2D Pose</strong></td>
<td>10.07</td>
<td>11.41</td>
<td>12.26</td>
</tr>
</tbody>
</table>

**Weight:** 45 - 110 kg

**standard deviation:** 12.5 kg
# Different Models: Dmocap Results

<table>
<thead>
<tr>
<th>Age (RMSE)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Markers</td>
<td>6.89</td>
<td>6.84</td>
<td>6.63</td>
</tr>
<tr>
<td>Upper Body</td>
<td>6.92</td>
<td>6.74</td>
<td>6.46</td>
</tr>
<tr>
<td>Lower Body</td>
<td>7.44</td>
<td>7.46</td>
<td>7.67</td>
</tr>
<tr>
<td>Frontal 2D Pose</td>
<td>7.08</td>
<td>7.12</td>
<td>6.87</td>
</tr>
<tr>
<td>Sagittal 2D Pose</td>
<td>7.02</td>
<td>6.94</td>
<td>6.91</td>
</tr>
</tbody>
</table>

**Age:** 13 - 43 years

**Standard deviation:** 6.6 years
## Predicting Human Ratings

a.k.a: Perceived Attributes

In all cases, our classifiers are consistently better in predicting human ratings.

<table>
<thead>
<tr>
<th>Mocap (%)</th>
<th>Gender</th>
<th>Weight</th>
<th>Age</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full 3D</strong></td>
<td>93</td>
<td>94</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td><strong>Height Normalized</strong></td>
<td>92</td>
<td>94</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td><strong>Motion Only</strong></td>
<td>93</td>
<td>94</td>
<td>87</td>
<td>94</td>
</tr>
</tbody>
</table>
Attribute Inference Results: Motion Capture

Gender Features
Attribute Inference Results: Motion Capture

Weight Features

X = RED
Y = GREEN
Z = WHITE

X = RED
Y = GREEN
Z = BLUE
But What About Tracking?

- We can use mocap for attribute inference.
- Mocap is almost noiseless.
- Can we do the same with tracking?
Video Pose Tracking Data: Dvideo

- **Dvideo**: 24 subjects with binocular video (30Hz) and mocap (120Hz) synchronized and one to two gait cycles in length.

- The 3D pose tracker is a modified version of an Annealed Particle Filter (APF) [deutscher 2005, Sigal 2010].

  - The likelihood used a combination of a probabilistic background model with shadow suppression and 2D point tracks.
Video Pose Tracking Data: Dvideo

Average error: \((x, y, z) = (63.7, 59.9, 82.3) \text{ [mm]}\)
Video and mocap consistency

Good News:

dVideo has trends in dMocap subspace representation.
Video and mocap consistency

Bad News:

Dvideo and Dmocap have different distributions.
Learning: Problem

**Problem:** Dmocap and Dvideo have different distributions:

- Pose data in Dvideo is based on a different joint parametrization.
- There are fewer joint degrees of freedom in the model.
- The 3D pose data from video tracking has a much lower signal to noise ratio.
Learning: Solution

Solution: transfer learning - domain adaptation.

• Learn source models from mocap training data.

• Adapt to the video-feature domain.
Learning: Resulting Models

The resulting models:

• Generalize better than those learned from the video-based pose data directly.

• Produce better results than the direct application of models learned from mocap.
Binary Attributes: Logistic Classifier

**A logistic model**: the posterior probability of an attribute is a Sigmoidal function of distance from a planar decision boundary:

\[
p(g = 1 \mid \mathbf{c}, \theta) = \frac{1}{1 + \exp(-\mathbf{c}^T \theta)} \equiv \sigma(\mathbf{c}^T \theta)
\]

Given IID source data \( \{\mathbf{c}^s_j, g^s_j\}_{j=1}^{N_s} \) the combined likelihood is:

\[
p(\mathcal{D}) = \prod_{j=1}^{N_s} p(g = g_j \mid \mathbf{c}_j, \theta)
\]
Binary Attributes: Logistic Classifier

The negative log likelihood of the source data becomes:

$$\mathcal{L}_s(\theta) = -\log \prod_{j=1}^{N_s} \sigma(c_j^s; \theta)^{g_j^s} (1 - \sigma(c_j^s; \theta))^{1 - g_j^s}$$

with model parameters being estimated using maximum likelihood:

$$\theta^s = \arg \min \mathcal{L}_s$$
We learn a logistic model on the target training data with a Gaussian prior centred at the source model:

\[
\mathcal{L}_t(\theta) = -\log \prod_{j=1}^{N_t} \sigma(c_j^t; \theta)^{g_j^t} (1 - \sigma(c_j^t; \theta))^{1 - g_j^t} + \lambda \|\theta - \theta^s\|^2
\]
Continuous Attributes: LS Regressors with Transfer Learning

• Similar technique for predicting real-valued attributes, such as age or weight.

• We formulate the target model in terms of a least-squares predictor:

\[
\mathcal{L}_c(\theta) = \sum_{j=1}^{N_t} (\theta^T c_j^t - a_j^t)^2 + \lambda \cdot ||\theta - \theta_{LS}^s||^2
\]
Gender Inference From Dvideo

![Graphs of gender hit rate vs. prior strength for Mocap and Tracking. The graphs show the performance of different methods: Full 3D, Height Norm, and Motion.](image)
### Pose Tracking Results: Gender

<table>
<thead>
<tr>
<th></th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>mocap (%)</strong></td>
<td>82</td>
<td>64</td>
<td>69</td>
</tr>
<tr>
<td><strong>mocap tracking</strong></td>
<td>62</td>
<td>62</td>
<td>64</td>
</tr>
<tr>
<td><strong>transfer learning</strong></td>
<td>87</td>
<td>71</td>
<td>71</td>
</tr>
<tr>
<td><strong>Pmin</strong></td>
<td>79</td>
<td>62</td>
<td>72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>mocap (%)</strong></td>
<td>51</td>
<td>56</td>
<td>55</td>
</tr>
<tr>
<td><strong>mocap tracking</strong></td>
<td>41</td>
<td>55</td>
<td>51</td>
</tr>
<tr>
<td><strong>transfer learning</strong></td>
<td>68</td>
<td>70</td>
<td>59</td>
</tr>
<tr>
<td><strong>Pmin</strong></td>
<td>58</td>
<td>60</td>
<td>49</td>
</tr>
</tbody>
</table>

- Using mocap classifier
- No transfer learning
- Best results with transfer learning
Pose Tracking Results: Gender

<table>
<thead>
<tr>
<th></th>
<th>mocap (%)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>mocap</td>
<td>82</td>
<td>64</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>tracking</td>
<td>62</td>
<td>62</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>transfer learning</td>
<td>87</td>
<td>71</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td><strong>Pmin</strong></td>
<td>79</td>
<td>62</td>
<td>72</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>tracking (%)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>mocap</td>
<td>51</td>
<td>56</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>tracking</td>
<td>41</td>
<td>55</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>transfer learning</td>
<td>68</td>
<td>70</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td><strong>Pmin</strong></td>
<td>58</td>
<td>60</td>
<td>49</td>
<td></td>
</tr>
</tbody>
</table>

- **Pmin**: statistical significance
- In 9 out of 10 runs we should get same or better results than **Pmin**.

\[
\int_{p_{min}}^{1} f(p \mid k, n) dp = 0.9
\]
Weight Inference From Video Tracking
Additional Model: Matching Mocap Results With Video

- Tracking still performs significantly worse than mocap.
- Phase noise is a significant source of tracking noise.

**Solution**: use amplitude only.
Amplitude-Based Model

Results: Gender

- Full 3D
- Height Norm
- Motion

Graphs showing gender hit rate vs. prior strength for mocap and tracking.
### Amplitude-Based Model Results: Gender

<table>
<thead>
<tr>
<th>mocup (%)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>mocup</td>
<td>88</td>
<td>70</td>
<td>67</td>
</tr>
<tr>
<td>tracking</td>
<td>77</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>transfer learning</td>
<td>93</td>
<td>70</td>
<td>67</td>
</tr>
<tr>
<td>Pmin</td>
<td>86</td>
<td>60</td>
<td>58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>tracking (%)</th>
<th>3D Model</th>
<th>Height Normalized</th>
<th>Motion Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>mocup</td>
<td>88</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>tracking</td>
<td>91</td>
<td>69</td>
<td>51</td>
</tr>
<tr>
<td>transfer learning</td>
<td>91</td>
<td>75</td>
<td>67</td>
</tr>
<tr>
<td>Pmin</td>
<td>84</td>
<td>67</td>
<td>58</td>
</tr>
</tbody>
</table>
Inference of Perceived Attributes: Video Tracking

<table>
<thead>
<tr>
<th>Tracking (%)</th>
<th>Gender</th>
<th>Weight</th>
<th>Age</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D Model</td>
<td>80</td>
<td>80</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>Height Normalized</td>
<td>85</td>
<td>83</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>Motion Only</td>
<td>93</td>
<td>76</td>
<td>98</td>
<td>93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mocap (%)</th>
<th>Gender</th>
<th>Weight</th>
<th>Age</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full 3D</td>
<td>93</td>
<td>94</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>Height Normalized</td>
<td>92</td>
<td>94</td>
<td>89</td>
<td>85</td>
</tr>
<tr>
<td>Motion Only</td>
<td>93</td>
<td>94</td>
<td>87</td>
<td>94</td>
</tr>
</tbody>
</table>

| Height Normalized | 80 | 80 | 87 | 87 |
| Motion Only      | 85 | 83 | 93 | 91 |
Demo

Biomotion Lab

Nikolaus Troje

http://www.biomotionlab.ca/Demos/BMLwalker.html
Discussion

• Inferring significant physical attributes (e.g. gender, weight) and aspects of mental state (e.g. happiness) is feasible.

• Current state-of-the-art 3D tracking methods are capable of matching mocap data in predicting some certain attributes (e.g. gender).
Discussion

• Perceived attributes are easier to predict.

• Relying on 3D RMSE is unreliable and illusive indicator for many real-life tasks of a 3D pose tracker.

• We suggest using attribute inference as a complementary quality indicator.
Future Work

• Collect larger data set and explore stronger tracking prior models.

• Test the inference of attributes with monocular pose tracking methods.

• Possible tracking method enhancement: estimate attributes as part of the tracking.
Acknowledgements

• This work was financially supported in part by NSERC Canada and the Canadian Institute for Advanced Research.

• This research is a Joint research with:
  Nikolaus Troje - Queen's University.
  Leonids Sigal - Disney Research.
Questions?