Human Attributes from 3D Pose Tracking

Leonid Sigal[1,3], David J. Fleet[1], Nikolaus F. Troje[2], and Micha Livne[1]

[1] Department of Computer Science, University of Toronto
[2] Department of Psychology and School of Computing, Queen's University
[3] Disney Research, Pittsburgh

Abstract. We show that, from the output of a simple 3D human pose tracker one can infer physical attributes (*e.g.*, gender and weight) and aspects of mental state (*e.g.*, happiness or sadness). This task is useful for man-machine communication, and it provides a natural benchmark for evaluating the performance of 3D pose tracking methods (*vs.* conventional Euclidean joint error metrics). Based on an extensive corpus of motion capture data, with physical and perceptual ground truth, we analyze the inference of subtle biologically-inspired attributes from cyclic gait data. It is shown that inference is also possible with partial observations of the body, and with motions as short as a single gait cycle. Learning models from small amounts of noisy video pose data is, however, prone to over-fitting. To mitigate this we formulate learning in terms of domain adaptation, for which mocap data is uses to regularize models for inference from video-based data.

1 Introduction

The fidelity with which one needs to estimate 3D human pose varies from task to task. One might be able to classify some gestures based on relatively coarse pose estimates, but the communication of many biological and socially relevant attributes, such as gender, age, mental state and personality traits, necessitates the recovery of more subtle cues. It is generally thought that current human pose tracking techniques are insufficient for this task. As a consequence, most previous work on action recognition, gesture analysis, and the extraction of biometrics, has focused on 2D image properties, or holistic spatiotemporal representations. On the contrary, we posit that it is possible to infer subtle human attributes from video-based 3D articulated pose estimates. Further, we advocate the inference of human attributes as a natural, meaningful way to assess the performance of 3D pose tracking techniques.

In this paper, we consider the inference of gender, age, weight and mood from video-based pose estimates. One key problem is the lack of suitable training data comprising labeled image sequences with 3D pose estimates. To deal with this issue, our models are bootstrapped from a substantial corpus of human motion capture data, and then adapted using a simple form of inductive transfer learning. In particular, the adaptation accounts for differences between the distributions of features derived from mocap and the video-based pose tracking data. Ground truth gender, age and weight are provided with the mocap and some video-based pose tracking data. We also consider models trained on *perceived* attributes gathered from human perception experiments over the internet. For various aspects of mental state, like mood (happiness), human perception is, at present, our principal source of (ground truth) training data.

The inference of human attributes has myriad potential uses, ranging from human-computer interaction to surveillance to clinical diagnostics. E.g., biometrics are of interest in security, and retails stores are interested in shopper demographics. The range of potential applications increases further as one considers a wider range of attributes, including, for example, the degree of clinical depression [17], or levels of anxiety.

The goal of this paper is to demonstrate a simple proof-of-concept model for attribute inference. We restrict our attention to walking motions, a generic 3D pose tracker, the extraction of simple motion features, and a very basic set of attributes. Pose tracking from two views is accomplished with an Annealed Particle Filter [8, 29], with a likelihood derived from background subtraction and 2D point tracks. We avoid the use of sophisticated activity-specific prior models (*e.g.*, [18, 30]) that are prone to over-fitting, thereby biasing pose estimates and masking useful information. Following [23, 28, 31, 33] our motion features are derived from a low-dimensional representation of joint trajectories in a body-centric coordinate frame. We then use a regularized form of logistic regression for classification. The experimental results show that one can infer attributes from video pose estimates (at 60–90% accuracy depending on the attribute). We are confident these results can be improved with advances in 3D pose tracking.

2 Background and Related Work

Perception of Biological Motion: Almost 40 years ago, Johansson [12] showed that a simple display with a small number of dots, moving as if attached to major joints of the human body, elicits a compelling percept of a human figure in motion. Not only can we detect people quickly and reliably from such displays, we can also retrieve details about their specific nature. Biological motion cues enable the recognition of familiar people [6, 32], and the inference of attributes such as gender, age, mental state, actions and intentions, even for unfamiliar people [3, 20, 31].

Humans reliably classify gender from point-light walkers with a hit rate (correct classification rate) of 65 to 75%; frontal views are classified best [20, 25, 31]. Studies have focused on cues that mediate gender classification, such as the shoulder-hip ratio [7] or the lateral sway of the upper body that is more pronounced in men [20]. Interestingly, depriving observers of kinematics degrades gender classification rates. When in conflict, information conveyed by dynamic features dominates that of static anthropometrics [20, 31]. Using PCA and linear discriminants Troje [31] modeled such aspects of human perception. Similar models have even been shown to convey information about weight and mood and the degree of depression in clinical populations [17].

Biometrics: Gait analysis is closely related to our task here. There is a growing literature on gait recognition, and on gender discrimination from gait (see [4] for a good overview), and a substantial benchmark datasets exist for gait recognition ([27]). However, such datasets are not well suited for 3D model-based pose tracking as they lack camera calibration and resolution is often poor. Indeed, most approaches to gait recognition rely mainly on background subtraction and properties of 2D silhouettes. Very few approaches exploit articulated models, either in 2D or 3D (although see [33, 35]).

Like gait recognition, gender classification from gait is usually formulated in terms of 2D silhouettes, often from sagittal views where the shape of the upper body, rather than motion, is the primary cue (*e.g.*, [16, 19]). With multiple views some form of voting is often used to merge 2D cues [10]. The use of articulated models for gender discrimination has been limited to 2D partial-body models. Yoo *et al.*, [34] used a set of 19 features, including 2D joint angles, dynamics of hip angles, the correlation between left and right leg angles, and the centre coordinates of the hip-knee cyclogram, with linear and RBF SVMs, and a 3-layer feed-forward neural net for gender classification. Samangooei and Nixon [26] consider video retrieval with physical attributes that include gender, age and weight. But they assume 2D sagittal views and a green screen to simplify the extraction of silhouette-based gait signatures.

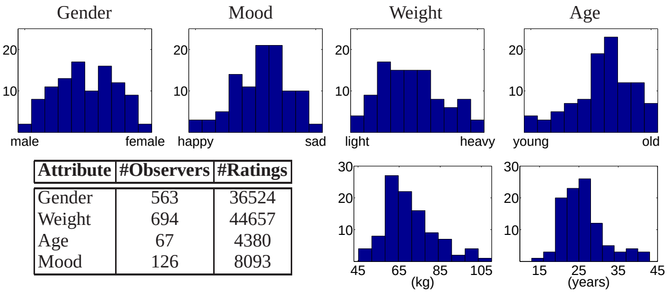
Unlike the gait recognition problem, inferring attributes of unfamiliar people does not presuppose that test subjects exist in the training data. Further, by using 3D articulated tracking we avoid the need for view-based models and constrained domains (*cf.* [10, 26, 34]). The video sequences we use were collected in an indoor environment with different (calibrated) camera locations, most of which did not include a proper sagittal view. Finally, here we infer physical attributes as well as aspects of mental state, like the mood of the subject. To our knowledge this is the first paper that attempts to address recovery of such attributes collectively from video-based 3D pose estimates.

Action Recognition: Like biometrics, most work on action recognition has focused on holistic space-time features, local interest points or space-time shapes (*e.g.*, [9, 14, 21]), in the image domain rather than with 3D pose in a body-centric or world frame. It is widely believed that 3D pose estimation is sufficiently noisy that estimator bias and variance will outweigh the benefits of such compelling representations. Nevertheless, some recent methods have successfully demonstrated that this may not be the case (*e.g.*, [22]). Unlike such work focused on classifying very different motion patterns, we tackle the more subtle problem of inferring meaningful percepts from locomotion.

3D Pose Tracking: The primary benchmark for evaluating techniques for pose tracking, HumanEva [29], uses the 3D Euclidean distance between estimated and ground truth (mocap) joint positions. Errors in joint positions and joint angles are easy to measure, but it is not clear how they relate to task requirements. Will RMSE (root-mean-squared error) of 70mm be sufficient to determine gender or mood, or for gesture recognition? Some trackers with errors of 70mm might preserve the relevant information while others may not. As such, task-specific measures, like attribute inference, complement conventional RMSE measures. In particular, attribute inference is relatively complex as it depends on subtle pose and motion information. Furthermore, unlike many activity recognition tasks, which depend on motion and scene context (*e.g.*, [15]), attribute inference is mainly a function of information intrinsic to the agent or the perception of the agent's motion. Human attributes are of clear social significance, and may be directly relevant to applications. That said, an extensive comparison of different pose trackers based on attribute inference is beyond the scope of this paper.

3 Human Motion and Attribute Data

Models for different attributes are learned from a combination of partially labeled video and motion capture data. Unfortunately, since we had video data from only 20 subjects, models trained on video-based tracking data are prone to over-fitting. On the other hand, models learned from mocap should not be applied blindly to tracking data because many



Attribute	#Observers	#Ratings
Gender	563	36524
Weight	694	44657
Age	67	4380
Mood	126	8093

Fig. 1. Web Attribute Data: The top row shows histograms of average ratings from observers for four attributes. The bottom row histograms show ground truth distributions of weight (kg) and age (yrs). The numbers of observers and walkers rated for each attribute are given in the table.

of the discriminative features in mocap data cannot be reliably estimated during pose tracking. Therefore, as discussed below (Sec. 4), we train from a combination of mocap and tracking data using a simple formulation of transfer learning.

3.1 Motion Capture Data: $\mathcal{D}_{mocap}$

Our source mocap data comprises walking motions from 115 individuals. From 41 physical markers we estimate 15 3D "virtual markers" at major joints of the body, *i.e.*, at shoulder joints, elbows, wrists, hip joints, knees, and ankles, and at the centers of the pelvis, clavicles, and head. Each participant walked for several minutes within the capture volume at their preferred speed, after which we began to record up to 4 trials of walking. The data are also labelled with gender, age and weight (see Fig. 1).

Human Subject Ratings: In addition to physical attributes we also consider perceived attributes, *i.e.*, what people perceive when viewing point-light displays of walking people. With this data one can begin to explore biological cues that convey gender, age and weight. More importantly, this provides us with labels about apparent mental state, such as mood (happiness or sadness).

In a web-based experiment observers were asked to rate walkers using attributes of their choosing. Each observer specified an attribute, and then rated up to 100 walkers (in random order) on a scale of 1 to 6. They were also asked to enter two phrases to indicate what ratings of 1 and 6 represent.[4] From ratings of over 4000 observers, each of whom rated at least 20 walkers, we selected sessions for which the named attribute was one of "gender", "age" or "weight", and the labels for ratings 1 and 6 were meaningful. For "gender" we accepted "male-female" or "masculine-feminine", for "age" they had to contain "young" and "old" (or "elderly"), and for "weight", "light" and "heavy". We accepted any of "mood", "emotion", "happy", or "happiness" for the mood attribute, and ratings 1 and 6 had to include the words "happy" and "sad". The resulting numbers of subjects and trials are given in Fig. 1. For each of the 100 walkers displayed, we computed the average rating, over all observers. Fig. 1 shows the distributions. Although data from experiments like this are noisier than those collected under more controlled conditions, they do reveal consistent perceptual interpretations.

[4] http://www.biomotionlab.ca/Demos/BMLrating.html

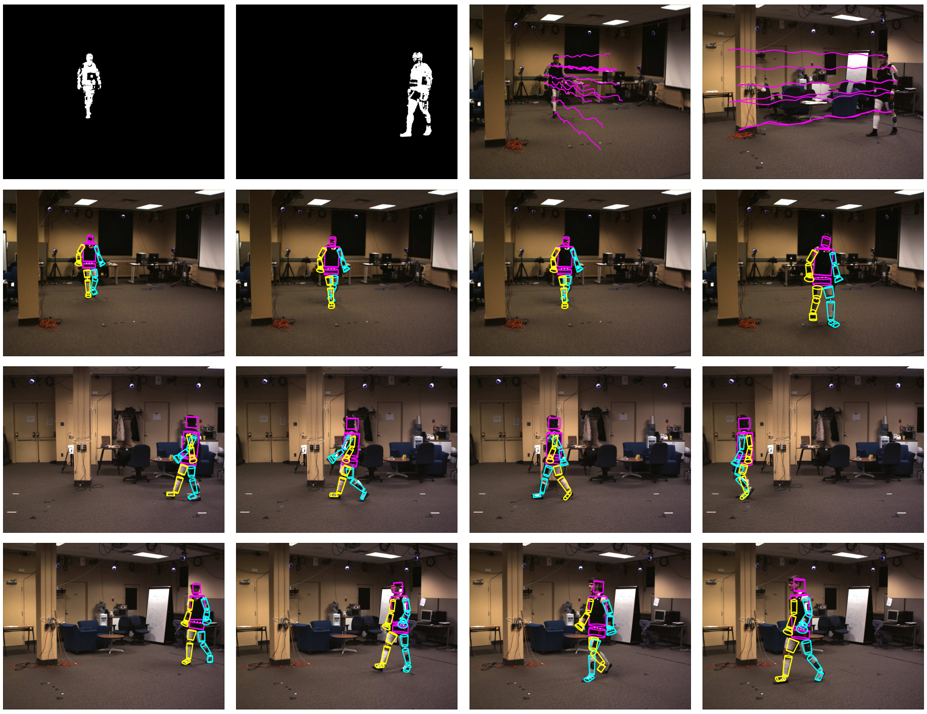


Fig. 2. Video Pose Tracking: The APF tracker uses a background model and 2D tracked points from two views (top row). Tracking output for three subjects are shown in the bottom three rows, with average error in 3D joint locations of 63.7 (mm), 59.9 (mm), and 82.3 (mm) respectively. Notice the differences in camera orientations and the background.

3.2 Video Pose Tracking Data: $\mathcal{D}_{video}$

In addition to the mocap above, we also have synchronized binocular video (30Hz) and mocap (120hz). We captured 2-3 sequences for each of 20 subjects (10 male, 10 female) walking, with different camera configurations, but usually with views that were within $30°$ of frontal and sagittal. Each sequence was approximately two gait cycles in length.

The 3D pose tracker was a modified version of an Annealed Particle Filter (APF) [8, 29]. The likelihood used a combination of a probabilistic background model with shadow suppression, and 2D point tracks [11] (see Fig. 2 (top)). Point tracks were only used for body parts that remain visible, the likelihood for which was formulated as a truncated Gaussian (for robustness). The same likelihood was used for all subjects. We used a 15-part body model comprising truncated cylinders, with 34 joint angles plus global pose [29] (40 DOFs in total). The prior motion model was a smooth first-order Markov model, with weak joint limits and inter-penetration constraints. The lack of an activity-specific prior motion model was motivated by the desire to avoid biasing the pose estimates towards a particular population. All experiments used the same APF setup (200 particles/layer, 5 layers), requiring roughly 2 minutes/frame (Matlab). We believe it is possible to estimate partial anthropometrics online while tracking [2], but for simplicity we assumed known anthropometrics.

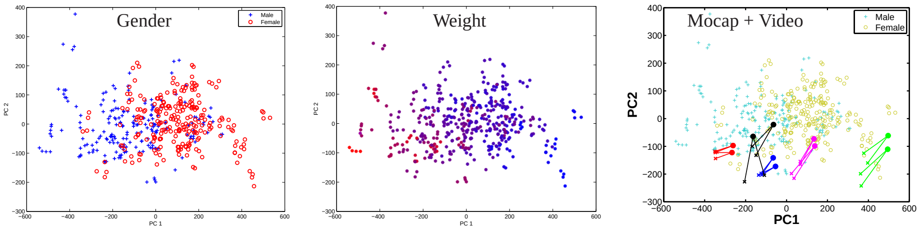


Fig. 3. Subspace Visualization: The distribution of motions in $\mathcal{D}_{mocap}$ in the first 2 principal dimensions is shown. (Left) Males (blue +) and females (red o). (Middle) Weight is depicted with blended colors: Heavy (red) and light (blue). (Right) Video pose tracks and mocap from 5 subjects in $\mathcal{D}_{video}$ are shown in 2 subspace dimensions: (color coded); circles indicate two video trials, crosses corresponding tracks; (cyan – $\mathcal{D}_{mocap}$ males, yellow – $\mathcal{D}_{mocap}$ females).

The tracker performed well except when the legs were close; in rare cases the leg identities were switched. In these cases we did not filter the results in any way. In fact we report performance on all tracks obtained. We ran the tracker twice on every test sequence (yielding 80 pose trajectories). Sample tracking results for three subjects are shown in Fig. 2; in terms of the average Euclidean joint errors, the results are comparable to state-of-the-art [29]. The average Euclidean error in 3D joint locations over the 80 runs had a mean of 73mm and a standard deviation of 19mm.

Finally, note that pose data in $\mathcal{D}_{video}$ and $\mathcal{D}_{mocap}$ have structual differences. To facilitate video tracking the body model in $\mathcal{D}_{video}$ had fewer degrees of freedom. Also the mocap protocol used to estimate joint positions differed in $\mathcal{D}_{video}$ and $\mathcal{D}_{mocap}$.

3.3 Motion Representation

Following [28, 31] we represent each motion as a pose trajectory, *i.e.*, a vector comprising the 15 3D joint positions at each time step.[5] We exploit the periodic nature of locomotion, expressing each motion as a Fourier series [23, 31]; two harmonics are sufficient for walking [31]. To represent each pose trajectory, we encode the mean (DC) pose, along with the Fourier coefficients at the fundamental frequency and its second harmonic. This yields a 225-D vector for each motion (*i.e.*, 5 real-valued Fourier coefficients for each of 15, 3D markers). This encoding is somewhat robust to the noise in the 3D poses within a trajectory, allowing us to better deal with the poor SNR of the video-based pose data.

Let the Fourier-based representation of these N motions be $\{\mathbf{m}_j\}_{j=1}^{N}$, where $\mathbf{m}_j \in \mathbb{R}^{225}$. Not surprisingly we find that the dimension of the representation can be reduced significantly with PCA. Since the SNR of the mocap data is much higher than the tracking data, we compute the subspace basis from the mocap data (from the 115 subjects described above in Sec. 3.1). Well more than 90% of the data variance is captured in 16 dimensions; in practice, using more than 16 dimensions does not improve the accuracy of attribute prediction appreciably.

[5] Initially all the walkers are aligned. The world frame is oriented so subjects are walking along the X-axis. We remove slow trends in the forward and lateral directions, based on the motion of the COM (*i.e.*, the average of all 15 joint markers) the XY plane.

Let $\mathbf{B} \equiv [\mathbf{b}_1, ..., \mathbf{b}_K]$ denote the subspace basis, where K is usually 16 or below. Further, let $\mathbf{c}_j$ denote the subspace coefficients for $\mathbf{m}_j$; *i.e.*, $\mathbf{c}_j = \mathbf{B}^T(\mathbf{m}_j - \bar{\mathbf{m}})$ where $\bar{\mathbf{m}}$ is sample mean of the motion data $\{\mathbf{m}_j\}$. Fig. 3 depicts the distribution of gender and weight in the first two principal directions. While not linearly separable, the attribute structure is clearly evident.

Of course there are other possible motion features. For example, Yoo *et al.* [34] use features of an articulated model extracted from a sagittal view of walking people, from which they acheive good gender classification with SVMs. Based on their paper, our implementation of their features with several different classifiers produces no better than 75% correct gender classification on our mocap dataset $\mathcal{D}_{mocap}$, compared to hit rates of 80%-90% obtained here (cf. Fig. 5).

4 Learning

$\mathcal{D}_{mocap}$ provides a significant corpus of labeled mocap, *but* the subspace motion features from $\mathcal{D}_{mocap}$ and $\mathcal{D}_{video}$ have different distributions. First, the pose data in $\mathcal{D}_{video}$ is based on a different joint parameterization (more suitable for video-based pose tracking). More importantly, the video tracking data has a lower SNR and is often biased because certain parts of the body (*e.g.*, the feet) are not tracked well. Indeed, some features that are highly discriminative in $\mathcal{D}_{mocap}$ will be uninformative in $\mathcal{D}_{video}$. Conversely, learning models from the small corpus of noisy video data in $\mathcal{D}_{video}$ is prone to over-fitting.

To mitigate these problems we formulate the learning problem as a form of transfer learning, called *domain adaptation*. It is applicable when the source ($\mathcal{D}_{mocap}$) and target ($\mathcal{D}_{video}$) domains share the same features, but have significantly different feature distributions (*e.g.*, see [24]). Intuitively, we learn source models from the mocap training data. The source models are then adapted to the video-feature domain through the minimization of a loss function on the target data that is biased toward the source model (*e.g.*, [1, 5]). The resulting models generalize much better than those learned from the video-based pose data directly, and they produce much better results than the direct application of models learned from $\mathcal{D}_{mocap}$.

In more detail, we use logistic classifiers for the inference of binary attributes and for predicting human ratings. A logistic model expresses the posterior probability of an attribute, $g \in \{0, 1\}$, as a sigmoidal function $\sigma(\cdot)$ of distance from a planar decision boundary, defined by parameters $\theta \equiv (\mathbf{w}, b)$; *i.e.*,

$$p(g = 1 \,|\, \mathbf{c}, \theta) \;=\; \frac{1}{1 + \exp(-\mathbf{c}^T\mathbf{w} - b)} \;\equiv\; \sigma(\mathbf{c}^T\mathbf{w} + b)\ . \tag{1}$$

The weights that define the decision hyperplane are found by ML optimization. That is, given source mocap data, $\{\mathbf{c}_j^s, g_j^s\}_{j=1}^{N_s}$, the optimized parameters are found by minimizing the negative log likelihood of the data with respect to the weight vector $\mathbf{w}$ and the bias offset b, *i.e.*, $\theta^s = (\mathbf{w}^s, b^s) = \arg\min \mathcal{L}_s$, where

$$\mathcal{L}_s(\mathbf{w}, b) \;=\; -\log \prod_{j=1}^{N_s} \sigma(\mathbf{c}_j^s; \mathbf{w}, b)^{g_j^s}\ (1 - \sigma(\mathbf{c}_j^s; \mathbf{w}, b))^{1-g_j^s}\ . \tag{2}$$

To adapt the model learned from $\mathcal{D}_{mocap}$ to the target data $\mathcal{D}_{video}$, following [5], we learn a logistic model on the target training data with a Gaussian prior centered at the source model. That is, we minimize a loss function that is a combination of the negative log likelihood of the video training data, $\{\mathbf{c}_j^t, g_j^t\}_{j=1}^{N_t}$, $N_t \ll N_s$, and a quadratic regularizer:

$$\mathcal{L}_t(\mathbf{w}, b) \; = \; -\log \prod_{j=1}^{N_t} \sigma(\mathbf{c}_j^t; \mathbf{w}, b)^{g_j^t} \, (1-\sigma(\mathbf{c}_j^t; \mathbf{w}, b))^{1-g_j^t} \; + \; \lambda ||\mathbf{w} - \mathbf{w}^s||^2 \, . \quad (3)$$

While this formulation assumes an isotropic prior, with variance $1/\lambda$, the loss function is easily generalized to an anisotropic prior that allows some weights to drift more than others. The covariance for an anisotopic prior might be set according to the ratio of variances in the subspace projections of $\mathcal{D}_{mocap}$ and $\mathcal{D}_{video}$ respectively. Nevertheless the experiments reported below are based on an isotropic prior.

Cross-validation is used to determine λ. Also, note that we do not regularize the bias offset since it is often convenient to allow b to vary freely to account for any bias in the tracking data. Minimization of $\mathcal{L}_t$ is accomplished with Newton iterations to solve for critical points, *i.e.*,

$$\frac{\partial \mathcal{L}_t}{\partial \mathbf{w}, b} \; = \; \sum_{j=1}^{N_t} (\sigma(\mathbf{c}_j^t; \mathbf{w}, b) - g_j^t) \begin{pmatrix} \mathbf{c}_j^t \\ 1 \end{pmatrix} \; + \; \lambda \begin{pmatrix} \mathbf{w} - \mathbf{w}^s \\ 0 \end{pmatrix} \; = \; \mathbf{0} \, . \quad (4)$$

One can generalize the approach to model the ratings data by replacing the ground truth g in (3) with the average rating (scaled to $(0, 1)$). Treating the average rating as the expected value of g over different observers, (3) can be interpreted as the expected likelihood. Also, while the approach formulated here presupposes labelled target data, it is also possible to extend the technique to the semi-supervised case where the target video data is not labeled (*e.g.*, [1]).

In addition to simple classifiers for binary attributes, we also consider domain-adapted least-squares (LS) regressors for real-valued attributes, such as age and weight. For example, the adapted LS predictor for real-valued attribute a minimizes

$$\mathcal{L}_c(\mathbf{w}, b) \; = \; \sum_{j=1}^{N_t} \left[(\mathbf{w}^T \mathbf{c}_j^t + b) - a_j^t \right]^2 \; + \; \lambda ||\mathbf{w} - \mathbf{w}_{LS}^s||^2 \, . \quad (5)$$

where $\mathbf{w}_{LS}^s$ is the LS optimal weight vector learned from the mocap data in $\mathcal{D}_{mocap}$.

5 Models and Analysis of Source Data: $\mathcal{D}_{mocap}$

We first learn models for the inference of different attributes using the labelled mocap corpus, $\mathcal{D}_{mocap}$. We tried learning with several different loss functions, including Gaussian class-conditional models and linear/RBF SVMs, but none generalized significantly better than logistic or linear LS regression. In all cases we characterize the expected performance of the classifier/regressor using leave-one-out cross-validation.

Figure 4 (left) shows how gender classification depends on the subspace dimension of the motion representation. With fewer than 16 dimensions important information is

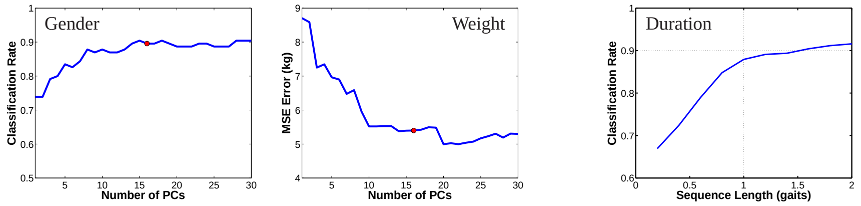


Fig. 4. Effect of Subspace Dimension and Sequence Length: Leave-one-out cross validation is used to asses the effect of subspace dimension on the correct-classification rate for the ground truth gender classification (left) and the RMSE of the real-valued weight regressor (middle). The right plot shows the dependence of gender classification on the duration (in gait cycles) of mocap sequences (based again on leave-one-out cross-validation).

lost. Classification performance with more than 20 dimensions yields marginal gains; with a 16D subspace the correct classification rate for gender is 90%. Fig. 4 (middle) shows the behaviour of a LS predictor for weight. The weights of our 115 walking subjects ranged from 50 to 100 kg, while the RMSE of predictions (16D features and leave-one-out cross-validation) is 5.4 kg. Fig. 4 (right) shows that gender can be classified with as little as one gait cycle (consistent with human perception [13]).

Normalized Models: To infer attributes from video pose estimates, we may not have access to full 3D pose. For example, with monocular tracking one might be able estimate 3D pose only up to the overall scale of the subject. Many 3D pose trackers simply assume the subject is average height (*e.g.*, [2]). In extreme cases a pose tracker may have no anthropometric knowledge whatsoever. To explore these cases we computed two further subspace representations of the data in $\mathcal{D}_{mocap}$. First all walkers were normalized to be the same height, and second, all anthropometrics are removed (by computing joint angles and then using the mean anthropometrics to reconstruct the motions).

The first row of results in Fig. 5 gives the gender hit rate (*i.e.*, correct classification rate) and the RMSE of linear LS predictors for weight and age, all based on leave-one-out (LOO) testing. One can see that the two normalized models are less informative than using the full 3D data. Predictions from the height-normalized models are somewhat better than the anthropometric-normalized models as expected. Also note that while predictions of gender and weight are quite good, age is poorly predicted. The walking subjects in this dataset ranged in age from roughly 18 to 35 years, while the RMSE for age prediction is 6.9 years.

Incomplete Data: To infer attributes from video-based pose estimates, we must be able to cope with missing data, since parts of the body may be partially or entirely occluded. Let $\mathbf{m} \in \mathbb{R}^{225}$ be a *complete* measurement vector (*i.e.*, the Fourier coefficients for each joint). Let the observed measurements be $\mathbf{m}_0 = P\mathbf{m}$, where the matrix P comprises only those rows of the identity matrix that correspond to the observed joints. It then follows from the generative subspace model, *i.e.*, $\mathbf{m} = \mathbf{B}\mathbf{c} + \bar{\mathbf{m}}$, that a LS pseudo-inverse can be used to estimate the subspace coefficients $\mathbf{c}_0$ from $\mathbf{m}_0$, *i.e.*,

$$\mathbf{c}_0 \;=\; (\mathbf{B}^T P^T P \mathbf{B})^{-1} \mathbf{B}^T P^T (\mathbf{m}_0 - P\bar{\mathbf{m}}) \; . \qquad (6)$$

The columns in Fig. 5 report model performance when data from model joints of the upper body, or from the lower body, are used. Also reported are results when one

	Gender (% correct)			**Weight** (RMSE kg)			**Age** (RMSE yrs)		
	Full 3D	**Height Norm.**	**Motion Only**	**Full 3D**	**Height Norm.**	**Motion Only**	**Full 3D**	**Height Norm.**	**Motion Only**
Full 3D Pose	89.6	86.1	81.7	5.4	9.7	10.9	6.9	6.9	6.4
Upper 3D Body	87.8	86.1	80.9	5.9	9.9	11.0	7.0	7.1	6.3
Lower 3D Body	84.4	80.0	73.9	6.2	9.4	12.2	7.2	7.2	7.3
Frontal 2D Pose	87.0	80.0	76.5	5.5	9.6	10.8	7.0	7.1	6.9
Sagittal 2D Pose	80.9	83.5	79.1	9.9	11.5	12.2	7.1	7.0	6.7

Fig. 5. Inference with $\mathcal{D}_{mocap}$ Models: To assess performance, with and without missing data, we build 3 models: **Full 3D** uses known anthropometrics and kinematics; **Height Normalized** is learned from mocap that is height normalized; and **Motion Only** uses only kinematic information (all walkers have the same limb lengths). The lack of anthropometrics degrades performance, but the inference of gender and weight are above chance in all models. We also report how performance varies with different subsets of markers (*e.g.*, upper/lower body) or 2D projections. Again, despite degradation in performance, the models continue to predict attributes well.

	Gender	**Weight**	**Age**	**Mood**
Full 3D	94	93	88	94
Height Normalized	93	93	86	93
Motion Only	93	94	86	93

Fig. 6. Inference of Perceived Attributes: We report the accuracy of predictions of human ratings for gender, weight, age and mood, all from the source mocap dataset $\mathcal{D}_{mocap}$. Perceived attributes are quantized to one bit based on the average rating for each subject, and the output of the logistic regressor is thresholded at 0.5. The table shows the fraction of subjects for which the classifier matches the quantized rating. Notice that perceived attributes are generally better predicted by the learned models than are ground truth attributes (*cf.* age in Fig. 5).

uses 2D data under orthographic projection from frontal or sagittal views. Interestingly, the observation that frontal views are more informative than sagittal views is consistent with studies of human perception [31].

Predicting Human Ratings: It is also interesting to consider how well one can predict *perceived attributes*. This is a scientific curiosity for physical attributes like gender, age and weight. For mood, however, we have no physical ground truth. Rather, the perceived mood is our only labelled data source. For all attributes, because our perceptual rating data are noisy, we quantize ratings of each attribute to one bit; *i.e.*, each walker is (perceived to be) (1) male or female, (2) heavy or light, (3) young or old, and (4) happy or sad. Then, the average attribute rating for a given training subject (scaled to $(0, 1)$) is taken to be the corresponding probability of being male, heavy, old, and happy, respectively. We use logistic regression to predict these probabilities, with leave-one-out measures of performance given in Fig. 6.

It is striking that, in all cases, we can do a better job predicting human ratings than ground truth. Human observers are, purportly using the available visual cues in a consistent manner, even if it is inconsistent with the ground truth. In particular, while true age is very hard to predict, perceived age is predicted well; *it's not how old you are, it's how old you look*. While interesting, this also shows clearly that percevied attributes may be biased, and therefore require qualification.

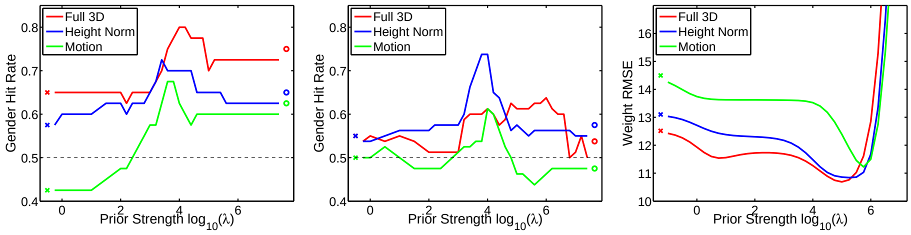


Fig. 7. Domain Adaptation: (a) Gender classification from the *mocap* in $\mathcal{D}_{video}$ for 20 test subjects (from leave-one-out performance), as a function of the strength of the prior λ, for each of 3 models (full 3D, height normalized, motion only). (b) Gender classification from the video-based *pose tracking data* for 20 test subjects (leave-one-out performance). (c) RMSE of weight estimates from *pose tracking data*, for 20 test subjects, as a function of the strength of the prior.

6 Attribute Inference from $\mathcal{D}_{video}$

Given the source models learned from $\mathcal{D}_{mocap}$, we use domain adaptation to learn models for the test pose data in $\mathcal{D}_{video}$. Not only is this useful in generating models for the video pose tracking data, it is also useful in building a classifier from the test mocap in $\mathcal{D}_{video}$. The reason is that the pose data in $\mathcal{D}_{video}$ is noisier and is parameterized differently from that in $\mathcal{D}_{mocap}$. The mocap in $\mathcal{D}_{mocap}$ allows for variable joint locations, while the parameterization of the tracker used in $\mathcal{D}_{video}$ has fixed joints. The tracker also has a fewer DOFs. Hence there are structural differences even between the mocap in $\mathcal{D}_{mocap}$ and that in $\mathcal{D}_{video}$.

Domain Adaptation: Figure 7 (left) show the leave-one-out hit rates for gender classifiers learned from $\mathcal{D}_{video}$ with domain adaptation from $\mathcal{D}_{mocap}$. The curves show how performance depends on adaptation from the source model, as a function of λ (see (3) in Sec. 4). The highest hit rates occur with λ between 10^3 and 10^4. For comparison, the crosses (x) depict the hit rate when there is no domain adaptation (i.e., with $\mathbf{w}^s = \mathbf{0}$ in (3)). The circles (o) depict the hit rate when the classifiers are trained solely on the source data $\mathcal{D}_{mocap}$ (with no domain adaptation) and then tested on the mocap in $\mathcal{D}_{video}$. Remember that the body model in $\mathcal{D}_{video}$ has fewer degrees of freedom and was estimated using a different mocap protocol from that in the original mocap in $\mathcal{D}_{mocap}$. Hence even the mocap motion features in $\mathcal{D}_{mocap}$ and $\mathcal{D}_{video}$ are distributed differently, and hence the value of domain adaptation.

Pose Tracking Data: Figure 7 (middle) shows leave-one-out hit rates for gender from video-based 3D pose tracking data (two trials of the APF, for each of 2 walking sequences for each of 20 subjects). As above, the curves show the dependence on the strength of the prior from the source model. The crosses (x) depict hit rates with no domain adaptation (from pose tracking data alone), and the circles (o) depict the hit rates from classifiers trained solely on the source mocap data $\mathcal{D}_{mocap}$. It is not clear why the full 3D model with pose tracking data is much worse than that with mocap input.

Figure 7 (right) shows how predictions of weight from video-based 3D pose data depends on domain adaptation. As above, the crosses (x) and the circles (o) show that predictions are poor when based solely on the data in $\mathcal{D}_{mocap}$ or in $\mathcal{D}_{video}$. With domain adaptation the results improve significantly. The standard deviation of the weight among

	Gender - mocap			**Gender - tracking**			**Weight - mocap**			**Weight - tracking**		
	(% correct, $\lambda = 10^4$)			(% correct, $\lambda = 10^4$)			(RMSE kg, $\lambda = 10^{1.5}$)			(RMSE kg, $\lambda = 10^5$)		
	Full 3D	**Height Norm.**	**Motion Only**	**Full 3D**	**Height Norm.**	**Motion Only**	**Full 3D**	**Height Norm.**	**Motion Only**	**Full 3D**	**Height Norm.**	**Motion Only**
C_{mocap}	75.0	65.0	62.5	53.8	57.5	47.5	5.7	10.9	6.6	51.4	42.1	42.7
C_{track}	65.0	57.5	42.5	55.0	55.0	50.0	4.0	7.3	6.9	12.5	13.1	14.5
$C_{trackTL}$	**77.5**	**70.0**	**67.5**	**61.3**	**73.8**	**61.3**	**3.6**	**7.6**	**6.0**	**10.6**	**10.9**	**12.4**

Fig. 8. Attributes from Mocap and Pose Tracking Data: The tables reports leave-one-out performance on gender classification and weight prediction from test mocap and pose tracking data in the target dataset $\mathcal{D}_{video}$ of 20 subjects. There are 40 mocap sequences (2 walks/subject), and 80 pose trajectories from video tracking (2 tracking trials per sequence). Results from 3 models are reported: C_{mocap} is learned from the source mocap $\mathcal{D}_{mocap}$; C_{track} is learned solely from test data $\mathcal{D}_{video}$; $C_{trackTL}$ is learned with $\mathcal{D}_{video}$ and domain adaptation from $\mathcal{D}_{mocap}$.

the test subjects is approximately 12kg. With domain adaptation, with $\lambda = 10^5$, the RMSE decreases to approximately 10.6. These results with tracking data are worse than those based on training mocap data in Fig. 5, but we find them encouraging nonetheless.

Figure 8 gives numerical results for gender classification and weight prediction, from both test mocap and test pose tracking data (like the plots in Fig. 7). As above, we show results from three models: C_{mocap} is learned solely from the source mocap $\mathcal{D}_{mocap}$; C_{track} is learned solely from test data $\mathcal{D}_{video}$; $C_{trackTL}$ is learned with $\mathcal{D}_{video}$ and domain adaptation from $\mathcal{D}_{mocap}$. Not surprisingly, the predictions of gender and weight from on video tracking data are not as reliable as those from the mocap. They are, however, encouraging. While not shown in the figure, we also note that errors in gender classification are reasonably consistent between the test mocap and the test tracking data. Approximately 85% of the motions classified from the pose tracking data are concistent with classification from the corresponding mocap. Thus, while some of the errors in Fig. 8 are due to noise in the pose tracking data, some are due to the fact that indeed some females consistently walk like males and vice versa.

Inference of Perceived Attributes: Figure 9 reports leave-one-out hit rates in the prediction of the *perceived* attributes. Like the above experiment in Fig. 6 we quantize perceptual ratings to one bit and use logistic regression for classification (*e.g.*, happy vs. sad). For the purposes of this experiment we also consider the perceptual data as the *ground truth* (indeed for perceived mental state, *e.g.*, mood, that is our only source of data label) and look at the consistency of predictions between the leave-one-out model trained with mocap and with video tracking results from $\mathcal{D}_{video}$.

The consistency between the mocap and pose tracking is very good, with consistent classification rates between 74% to 93%. It is interesting to note that we can recover the mental state – mood (happiness), with 85% to 86% accuracy. Like the results reported in Fig. 6 the perceived age is predicted well when compared to our models for predicting true age.

7 Discussion

This paper demonstrates that one can, from the output of a video-based, 3D human pose tracker, infer physical attributes (*e.g.*, gender and weight) and aspects of mental

	Gender	Weight	Age	Mood
$C_{trackTL}$ (Full 3D)	83	79	93	86
$C_{trackTL}$ (Height Normalized)	74	79	90	85

Fig. 9. Classification of Perceived Attributes with Respect to MoCap: The table reports consistency of leave-one-out performance on *perceived* gender, weight, age and mood (happiness) between test mocap and pose tracking data in the target dataset $\mathcal{D}_{video}$ of 20 test subjects. We use predicted attribute values for test mocap as targets to train $C_{trackTL}$ binary classifiers (learned with $\mathcal{D}_{video}$ and domain adaptation from $\mathcal{D}_{mocap}$, all with $\lambda = 10^4$).

state (*e.g.*. happiness). The models are used to infer binary attributes (gender) and real-valued attributes (weight). We also consider the prediction of perceived attributes based on human perceptual experiments. This is useful for infering attributes such as mood where human judgements are our source of ground truth. Learning is accomplished using datasets comprising labelled mocap and video-based 3D pose estimates. These sources of training data are combined with a simple for of domain adaptation.

To our knowledge, this is the first paper in the literature that attempted to infer such perceptually and biologically meaningful attributes from 3D video-based pose estimates. In the future we hope to collect large datasets and explore stronger tracking prior models trained from large collections of mocap data. We also hope to be able to test the inference of attributes with monocular pose tracking methods. While the results reported here are interesting in their own right, we also suggest that tasks like this provide a natural way to assess the fidelity with which people trackers estimate 3D pose.

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

References

1. Arnold, A., Nallapati, R., Cohen, W.: A comparative study of methods for transductive transfer learning. In: ICDM Workshop on Mining and Management of Biological Data (2007)
2. Balan, A., Sigal, L., Black, M., Davis, J., Haussecker, H.: Detailed human shape and pose from images. In: Proc. IEEE CVPR (2007)
3. Blakemore, S., Decety, J.: From the perception of action to the understanding of intention. Nature Reviews Neuroscience 2(8), 561–567 (2001)
4. Boyd, J., Little, J.: Biometric gait recognition. Advanced Studies in Biometrics: Summer School on Biometrics (2003)
5. Chelba, C., Acero, A.: Adaptation of maximum entropy capitalizer: Little data can help a lot. In: Conf. on Empirical Methods in Natural Language Processing (2004)
6. Cutting, J.E., Kozlowski, L.T.: Recognizing friends by their walk: Gait perception without familiarity cues. Bulletin of the Psychonomic Society 9(5), 353–356 (1977)
7. Cutting, J.E., Proffitt, D.R., Kozlowski, L.T.: A biomechanical invariant of gait perception. J. Exp. Psych.: Human Perception and Performance 4, 357–372 (1978)
8. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. IJCV 61(2), 185–205 (2005)
9. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. PAMI 29(12), 2247–2253 (2007)

10. Huang, G., Wang, Y.: Gender classification based on fusion of multi-view gait sequences. Proc. ACCV pp. 462–471 (2007)
11. Jepson, A., Fleet, D., El-MAraghi., T.: Robust online appearance models for vision tracking. IEEE Trans. PAMI 25(10), 1296–1311 (2003)
12. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception & Psychophysics 14(2), 201–211 (1973)
13. Johansson, G.: Spatio-temporal differentiation and integration in visual motion perception. Psychological Research 38, 379–393 (1976)
14. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV (2007)
15. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. IEEE Conf. CVPR (2008)
16. Lee, L., Grimson, E.: Gait analysis for recognition and classification. In: Proc. IEEE Int. Conf. Auto. Face and Gesture Recog. (2002)
17. Lemke, M., Wendorff, T., Mieth, B., Buhl, K., Linnemann, M.: Spatiotemporal gait patterns during over ground locomotion in major depression compared with never depressed controls. J Psychiatr Res 34, 277–283 (2000)
18. Li, R., Tian, T.P., Sclaroff, S.: Simultaneous learning of non-linear manifold and dynamical models for high-dimensional time series. In: IEEE ICCV (2007)
19. Li, X., Maybank, S., Yan, S., Tao, D., Xu, S.: Gait components and their applications to gender recognition. IEEE Trans. SMC, Part C 38(2) (2008)
20. Mather, G., Murdoch, L.: Gender discrimination in biological motion displays based on dynamic cues. Proceedings of the Royal Society of London Series B 258, 273–279 (1994)
21. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV 79(3), 299–318 (2008)
22. Ning, H., Xu, W., Gong, Y., Huang, T.: Latent pose estimator for continuous action recognition. In: Proc. ECCV. pp. 419–433 (2008)
23. Ormoneit, D., Sidenbladh, H., Black, M., Hastie, T., Fleet, D.: Learning and tracking human motion using functional analysis. In: Work. Human Modeling, Anal. & Syn. (2000)
24. Pan, S., Yang, Q.: A survey on transfer learning. IEEE Trans. KDE 12 (2009)
25. Pollick, F., Kay, J., Heim, K., Stringer, R.: Gender recognition from point-light walkers. J. Exp. Psych.: Human Perception and Performance 31(6), 1247–1265 (2005)
26. Samangooei, S., Nixon, M.: Performing content-based retrieval of humans using gait biometrics. Multimedia Tools and Applications (Oct 2009)
27. Sarkar, S., Phillips, J., Liu, Z., Robledo, I., Grother, P., Bowyer, K.: The human id gait challenge problem: Data sets, performance, and analysis. IEEE TPAMI 27, 162–177 (2005)
28. Sidenbladh, H., Black, M., Fleet, D.: Stochastic tracking of 3d human figures using 2d image motion. Proc. ECCV 2, 702–718 (2000)
29. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV (2010)
30. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. In: Int. Conf. Machine Learning. pp. 759–766 (2004)
31. Troje, N.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. J. Vision 2(5), 371–387 (2002)
32. Troje, N., Westhoff, C., Lavrov, M.: Person identification from biological motion: effects of structural and kinematic cues. Percept Psychophys 67(4), 667–675 (2005)
33. Urtasun, R., Fleet, D., Fua, P.: Motion models for 3D people tracking. CVIU 104(2-3), 157–177 (2006)
34. Yoo, J.H., Hwang, D., Nixon, M.: Gender classification in human gait using support vector machine. Advances Concepts for Intelligent Vision Systems (2006)
35. Zhang, R., Vogler, C., Metaxas, D.: Human gait recognition. IEEE Workshop on Articulated and Nonrigid Motion (2004)