Computer Vision and Image Understanding 116 (2012) 648–660

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Human attributes from 3D pose tracking ☆

Micha Livne [a], Leonid Sigal [a,b], Nikolaus F. Troje [c], David J. Fleet [a,*]

[a] Department of Computer Science, University of Toronto, 6 King's College Rd, Toronto, Ontario, Canada M5S 3H5
[b] Disney Research, 4720 Forbes Ave. Pittsburgh, PA 15213, United States
[c] Department of Psychology and School of Computing, Queen's University, Kingston, Ontario K7M3N6, Canada

ARTICLE INFO

Article history:
Received 8 August 2011
Accepted 11 January 2012
Available online 3 February 2012

Keywords:
Human motion
Gait analysis
3D human pose tracking
Transfer learning
Gender recognition
Human attributes

ABSTRACT

It is well known that biological motion conveys a wealth of socially meaningful information. From even a brief exposure, biological motion cues enable the recognition of familiar people, and the inference of attributes such as gender, age, mental state, actions and intentions. In this paper we show that from the output of a video-based 3D human tracking algorithm we can infer physical attributes (*e.g.*, gender and weight) and aspects of mental state (*e.g.*, happiness or sadness). In particular, with 3D articulated tracking we avoid the need for view-based models, specific camera viewpoints, and constrained domains. The 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). We show results on a large corpus of motion capture data and on the output of a simple 3D pose tracker applied to videos of people walking.

© 2012 Elsevier Inc. All rights reserved.

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 spatio-temporal 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 use of tasks like 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 of walking people. 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 (mocap) data, and then adapted using a simple form of transfer learning. In particular, the adaptation accounts for differences between the distributions of features derived from mocap and those obtained from 3D pose sequences obtained from video-based tracking.

In addition to inferring gender, age and weight, we also consider the information conveyed by human motion about properties of mental state. Toward this end we exploit *perceived* attributes data that were gathered from human perception experiments. This allows us to consider differences between ground truth and perceived attributes, and it also allows us to consider attributes like mood, for which we have no ground truth data. For properties of emotional or mental state, at present, human perception is our primary source of training data.

By way of application, the inference of human attributes has myriad potential uses, ranging from human–computer interaction to surveillance to clinical diagnostics. For example, biometrics are of interest in security, and retail 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 [23,28], 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 [11,38] (APF), with a likelihood derived from background subtraction and 2D point tracks. We avoid the use of activity-specific prior models (*e.g.*, [24,39]) that are prone to over-fitting, thereby biasing pose estimates and masking useful information. Following

☆ This paper has been recommended for acceptance by J.K. Aggarwal.

* Corresponding author.
E-mail addresses: mlivne@cs.toronto.edu (M. Livne), lsigal@disneyresearch.com (L. Sigal), troje@queensu.ca (N.F. Troje), fleet@cs.toronto.edu (D.J. Fleet).

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2012.01.003

[32,37,43,47] 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). With a model based on the amplitudes of a Fourier representation, we can achieve a hit rate of 90% in gender classification from video-based tracking data, similar to what can be acheived with mocap data. With improvements in 3D people tracking techniques we should be able to improve inference for a wide range of other attributes.

2. Background and related work

2.1. Perception of biological motion

Almost 40 years ago, Johansson [17] 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 [9,45], and the inference of attributes such as gender, age, mental state, actions and intentions, even for unfamiliar people [5,27,43].

Humans reliably classify gender from point-light walkers with a hit rate (correct classification rate) of 65–75%; frontal views are classified somewhat better than sagittal views [27,34,43]. Studies have focused on cues that mediate gender classification, such as the shoulder–hip ratio [10] or the lateral sway of the upper body that is more pronounced in men [27]. Interestingly, depriving observers of dynamical information degrades gender classification rates. When in conflict, information conveyed by dynamic features dominates that of static anthropometrics [27,43]. By using Principal Component Analysis (PCA) and linear discriminants Troje [43] modelled 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 [23].

2.2. Biometrics

In image-based biometrics there has been a sustained focus of research on recognition and the inference of gender, age and expression from facial images (*e.g.*, [13,29,26]). There is also interesting work on the estimation of gender from hand shape [1]. Human motion complements these sources of information since the face or hands may easily be hidden from view, or poorly resolved in images.

Gait analysis is closely related to our task here. There is a growing literature on gait recognition, and on gender discrimination from gait (see [7] for a good overview), and substantial benchmark data sets exist for gait recognition ([36]). However, such data sets 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 [47,52]).

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.*, [22,25]). When multiple views are available some form of voting is often used to merge 2D cues [15]. The use of articulated models for gender discrimination has been limited to 2D partial-body models. Yoo et al. [51] used a set of 19 features, including 2D joint angles, dynamics of hip angles, the correlation between left and right leg angles, and the center coordinates of the hip–knee cyclogram, with linear and RBF SVMs, and a 3-layer feed-forward neural net for gender classification. We replicated their approach but when applied to our motion capture data we could not achieve the level of performance they report. Samangooei and Nixon [35] consider content-based video retrieval based on infered physical attributes, including gender, age and weight. However, they assume 2D sagittal views and a green screen to simplify the extraction of silhouette-based gait signatures.

With the use of 3D articulated tracking we avoid the need for view-based models, known camera viewpoints, and constrained domains (*cf.* [15,35,51]). The video sequences we use were collected in an indoor environment with different (calibrated) camera locations, most of which did not include proper sagittal or frontal views.

2.3. Action recognition

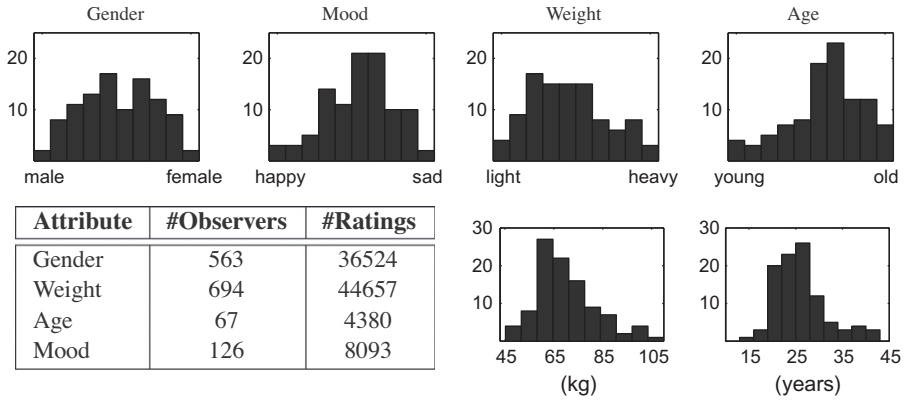
Like biometrics, most work on action recognition has focused on holistic space–time features, local interest points or space–time shapes (*e.g.*, [14,19,30]), in the image domain rather than with 3D pose in a body-centric or world frame of reference.

Holistic approaches focused on global space–time representations, with early methods relying on template-based encoding (*e.g.*, obtained by aggregating differences between subsequent silhouettes [6], computing average silhouette and contour images [48], or derived from person-centered optical flow [12]) and matching. Alternative methods relied on dense spatio-temporal descriptors (*e.g.*, non-negative matrix factorization of HOG descriptors [42]) followed by classifiers. Recently the focus has shifted to local spatio-temporal descriptors. Interest points are often detected as salient regions in the 3D spatio-temporal volumes [14] formed by stacking observed image frames. A number of interest operators have been proposed [21,49]. Descriptors are then utilized to summarize the spatio-temporal volume in the vicinity of the interest point. Examples for such descriptors are patches of normalized derivatives in space and time [30], speeded up robust feature (SURF) [49] and histogram of oriented gradients (HOG) or flow (HOF) [20].

It is widely believed that 3D pose estimation is sufficiently noisy that estimator bias and variance will outweigh the benefits of such compelling representations for action recognition and the analysis of activities. Nevertheless, some recent methods have successfully demonstrated that this may not be the case (*e.g.*, [31,50]). In these papers, 3D pose estimation is introduced as an intermediate latent representation used for action recognition. While these papers focused on classifying grossly different motion patterns, in this paper we tackle the more subtle problem of inferring meaningful attributes from human locomotion.

2.4. 3D pose tracking

The primary benchmark for evaluating pose tracking techniques, HUMANEVA [38], uses 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. For instance, it is not clear whether a 70 mm RMSE (Root Mean Squared Error) in pose estimate will or will not be sufficient to determine gender or mood, or for gesture recognition. Some trackers with errors of 70 mm might preserve meaningful information while others may not. As such, task-specific measures, like attribute inference, complement conventional RMSE measures of tracking performance. In particular, attribute inference is relatively complex as it depends on subtle



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 (years). The numbers of observers and walkers rated for each attribute are given in the table.

variations in pose and motion. Furthermore, unlike many activity recognition tasks, which depend on motion and scene context (*e.g.*, [20]), 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

We learn models for different attributes from a combination of partially labeled video and motion capture (mocap) data. The most direct way to learn models for classifying or predicting different attributes is to exploit labeled video-based pose tracking data. However such data are not readily available. In our case we have labeled video data for only 24 subjects, so over-fitting is problematic with such a small training corpus. As an alternative, there exist relatively large mocap corpora, comprising walking motions from hundreds of people. Unfortunately, we found that models learned from such mocap are suboptimal when applied to video-based tracking data because many of the discriminative features in mocap data are not reliably estimated during pose tracking. Our straightforward solution is to train models from a combination of mocap and tracking data, using a simple form of transfer learning.

3.1. Motion capture data: $\mathscr{D}_{mocap}$

Our source mocap corpus comprises walking motions from 115 individuals. From 41 tracked physical markers we estimate 15 3D "virtual markers" located 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. The capture volume was about 5 m long. Each participant walked for several minutes within the capture volume at their preferred speed, after which we began to record up to four trials of walking. The data is labeled with ground truth gender, age and weight [44] (see Fig. 1).

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 as perceived by humans. More importantly, this provides us with labels about apparent mental state, such as mood (happiness or sadness).

In a simple web-based experiment, observers were asked to rate walkers using attributes of their choice.[1] Each observer was asked to enter an attribute description and two phrases to indicate what ratings of 1 and 6 represent, then they proceeded to rate up to 100 walkers (in random order) on a scale of 1–6. 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". As the labels were determined by the observers, an additional pre-processing had to take place in order to group together similar labels. For "gender" we accepted "male–female" or "masculine–feminine", for "age" we accepted "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.

3.2. Video pose tracking data: $\mathscr{D}_{video}$

In a different lab facility and a different subject pool we collected video and mocap data from 24 subjects. Here we obtained synchronized binocular video (30 Hz) and mocap (120 Hz) for each subject. We tracked 2–3 sequences for each of the 24 subjects (12 male, 12 female) walking, with different camera configurations. The camera viewpoints vary from sequence to sequence, but in almost all cases two cameras were within 30 degrees of one frontal and one sagittal view. Each tracking sequence was approximately one to two gait cycles in length (between 40 and 100 view frames).

To obtain 3D pose data from the video data, we used a modified version of an Annealed Particle Filter (APF) for online human pose tracking [11,38]. The likelihood model for the tracker was derived from a probabilistic background model with shadow suppression, and heavy-tailed observation model for 2D point tracks [16] (see Fig. 2 (top)). The point tracking was performed only for body parts that were either always visible throughout the entire sequence, or were occluded only for short periods (*e.g.*, for less than 10 frames). The likelihood for the point tracks was formulated as a truncated Gaussian for robustness, and the same likelihood was used for all subjects. The background model comprised the mean color (RGB) and the mean intensity gradient as functions of image position, To simplify the estimation of the covariance matrix we assume that the 5D covariance matrix was identical for all pixels, and could therefore be estimated from pixel measurements over the entire image. Shadow suppression was also performed to allow a more precise localization of the feet.

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

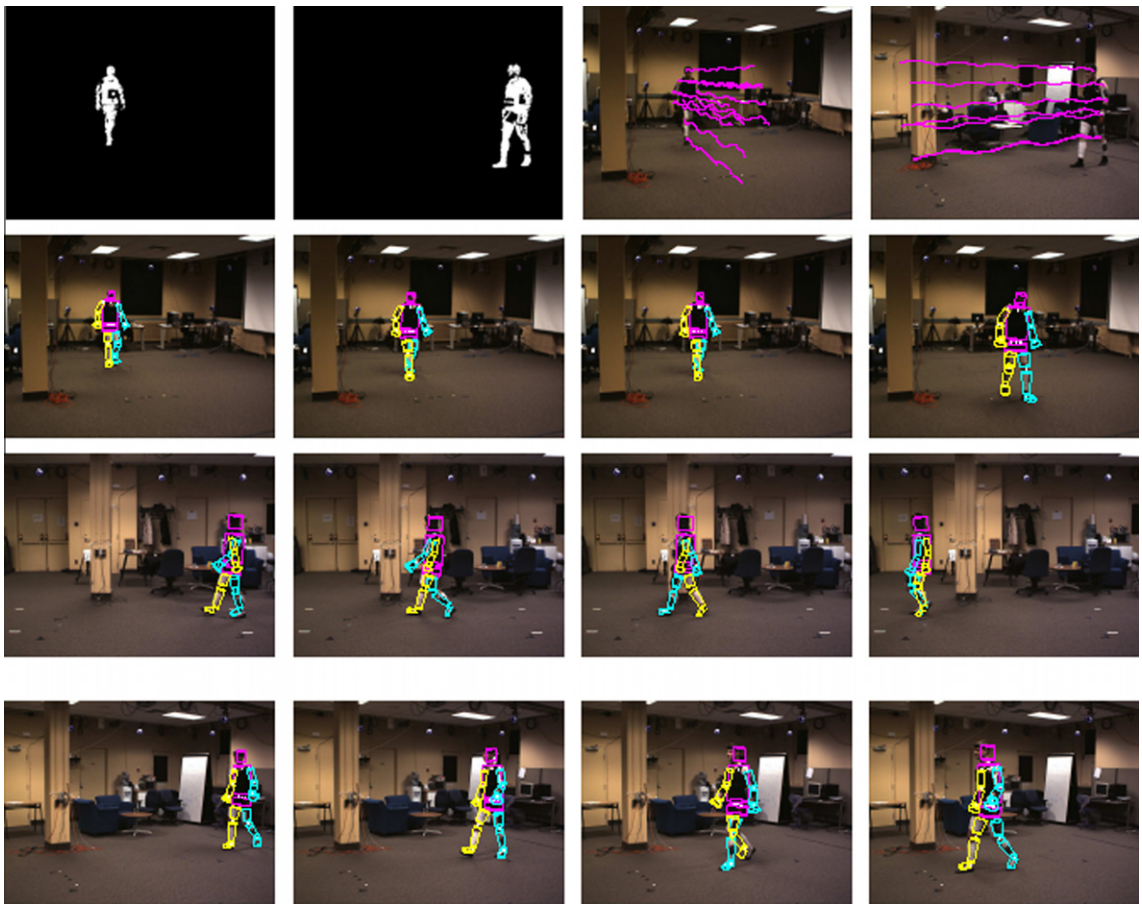


Fig. 2. Video pose tracking: The APF tracker uses a background model and 2D tracked points from two views (top row). 3D motions estimated for three subjects are shown in the bottom three rows, having average error in 3D joint locations of 63.7 (mm), 59.9 (mm), and 82.3 (mm) respectively. Notice the differences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

We used a 15-part body model comprising truncated cylinders, with 34 joint angles plus global pose [38] (40 DOF in total). The prior motion model was a smooth first-order Markov model, with weak joint limits and inter-penetration constraints. The simple prior motion model is weak by comparison to activity-specific models used by most state-of-the-art people tracking methods (*e.g.*, see [46,41,2,40]). This was motivated by our desire to avoid biasing the pose estimates towards a particular population. All experiments used the same APF parameterization (200 particles/layer, 5 layers). This required roughly 2 min/frame in our MATLAB implementation. We use an adopted version of the publicly available APF implementation of [38]. We believe it is possible to estimate partial anthropometrics on-line while tracking [4], but for simplicity we assumed known anthropometrics.

The tracker performed well except when the legs were close to one another. In such rare cases the leg identities were switched. In these cases we did not filter the results in any way; indeed, we report performance on all tracks obtained. We ran the tracker twice on every test sequence. Sample tracking results for three subjects are shown in Fig. 2; in terms of average Euclidean joint errors, the results are comparable to state-of-the-art methods (*e.g.*, see [38]). The average Euclidean error in 3D joint locations over all of the runs had a mean of 73 mm and a standard deviation of 19 mm.

3.3. Motion representation

Following [37,43] we represent each motion as a pose trajectory, *i.e.*, a vector comprising the 15 3D joint positions at each time step. We begin by aligning all motions to walk in the same direction and with the same gait phase as a function of time. Then we extract the model parameters that will be used for attribute inference.

3.3.1. Motion alignment

Each walking motion is first aligned with the world coordinate frame such that the *X*-axis coincides with the direction of locomotion (perpendicular to the coronal plane), the *Z*-axis is aligned with gravity, and the *Y*-axis is perpendicular to the sagittal plane. Next, we remove slow trends in the forward and lateral directions (*i.e.*, in *XY*-plane), based on the motion of the "center-of-mass", *i.e.*, the average of all 15 joint markers. This alignment centers the moving figure at the origin as though walking in place in the world "X" direction. With all such motions aligned in this way we then represent each joint tracjectory in terms of a sinusoidal Fourier basis. This Fourier-based representation is also then phase-shifted so that the different motions are aligned with respect to gait phase.

3.3.2. Fourier series

We exploit the periodic nature of locomotion by expressing each motion as a Fourier series [32,43,44]. Two harmonics are thought to be sufficient for walking [43]. In fact, when modeling noisy tracking data, higher harmonics are too noisy.[2] Thus, restricting ourselves to two harmonics acts as an efficient noise filtering. To represent each pose trajectory, we encode the mean (DC) pose, along

[2] In our experiments, using the 3rd and 4th harmonics did not change our results significantly. In some cases, performance actually decreased due to noise (*e.g.*, age inference based on tracking data).

with the Fourier coefficients at the fundamental frequency and the second harmonic. This yields a 226-D features vector for each motion (*i.e.*, five real-valued Fourier coefficients for each of 15, 3D markers, and the fundamental frequency). This encoding is somewhat robust to noise in the 3D pose trajectories as it presupposes that the motion is periodic with only two harmonics. This is especially useful when dealing with noisy video-based pose data.

The estimation of the fundamental frequency, $\omega \in \mathbb{R}^+$, must be done with some care because the motions are often less than two periods in length, and the video-based pose data are noisy. In particular, the simple approach of choosing the frequency that maximizes power in a discrete-time Fourier transform is not reliable. Instead, we take the finite temporal support into account in a simple generative model.

For each of 15 joints we have motion along the three coordinate axes, *X*, *Y*, and *Z*, producing 45 time series. For $j \in [1, 45]$, let $f^j(t)$ denote the jth motion for $t \in [1, \ldots, T]$. We approximate each of these motions using a DC term and two harmonics that are measured within a temporal window $r(t)$ that is 1 for $1 \leqslant t \leqslant T$, and zero otherwise; *i.e.*,

$$m^j(t; \omega, \mathbf{a}^j) = r(t) \sum_{h=-2}^{2} e^{-i\omega ht} a_h^j.$$

Here, the parameters of the model include ω, the fundamental frequency, and a vector of complex-valued Fourier coefficients $\mathbf{a}^j = (a_{-2}^j, \ldots a_2^j)$.[3] The least-squares fundamental frequency ω^* is then given by

$$\omega^* = \arg\min_{\omega} \left(\sum_{j=1}^{45} \min_{\mathbf{a}^j} \sum_{t=1}^{T} |f^j(t) - m^j(t; \omega, \mathbf{a}^j)|^2 \right). \quad (1)$$

Fundamental frequencies estimated using (1) were more accurate than simply choosing the frequency that maximized the discrete-time power spectrum.

3.3.3. Subspace motion model

Let the Fourier-based representation of the N motions in our data-set be $\{\mathbf{m}_i\}_{i=1}^N$, where $\mathbf{m}_i \in \mathbb{R}^{226}$. To further reduce the dimension of this motion feature vector, we applied PCA. Empirical results showed that more than 90% of the data variance is captured in 16 dimensions. We find that, in practice, using more than 16 dimensions does not improve the accuracy of attribute prediction in an appreciable way. We also considered a robust PCA variant, but the motion capture data is sufficiently well behaved that this also had no significant impact.

Let $\mathbf{B} \equiv [\mathbf{b}_1, \ldots, \mathbf{b}_K]$ denote the subspace basis, where K is typically 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 the sample mean of the motion data $\{\mathbf{m}_j\}$. Fig. 3 depicts the distribution of gender and weight in the first two principal directions. Even though only two of the 16 dimensions of the latent representation are depicted, one can already see some degree of attribute separability in the distribution of the data.

Of course there are other possible motion features. For example, Yoo et al. [51] use features of an articulated model extracted from a sagittal view of walking people, from which they achieve 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 corpus, $\mathscr{D}_{mocap}$, compared to hit rates of 80–90% obtained here (*cf.* Fig. 8).

[3] Since the motions f^j are real-valued, there are only 5 degrees of freedom in the Fourier coefficients a_h^j.

4. Learning

As discussed above, learning is based on partially labeled video and motion capture (mocap) data, combined with a simple form of transfer learning. $\mathscr{D}_{mocap}$ provides a significant corpus of labeled mocap, but the subspace motion features from $\mathscr{D}_{mocap}$ and $\mathscr{D}_{video}$ have different distributions (*e.g.*, see Fig. 4). First, the pose data in $\mathscr{D}_{video}$ is based on a different joint parametrization. There are fewer joint degrees of freedom in the model, at knees and elbows, for example, to simplify parameter estimation and tracking. More importantly, the 3D pose data from video tracking has a much lower signal to noise ratio. Tracking errors are the major noise source in $\mathscr{D}_{video}$, especially when parts of the body are occluded, or confused with one another (*e.g.*, the feet). Indeed, some features that are highly discriminative in $\mathscr{D}_{mocap}$ will be uninformative in $\mathscr{D}_{video}$. Other features are indeed discriminative in $\mathscr{D}_{video}$, but are distributed differently than in $\mathscr{D}_{mocap}$.

For example, Fig. 4 depicts the subspace representation of $\mathscr{D}_{video}$ data in the $\mathscr{D}_{mocap}$ subspace. This depictions clearly indicates that the subspace features of both data-sets are capable of supporting gender classification, but it is also clear that the two feature distributions require different decision boundaries. As a consequence, we cannot simply learn models from $\mathscr{D}_{mocap}$ and then blindly apply them to data from a video-based pose tracker. Conversely, learning models directly from our relatively small corpus of noisy video data in $\mathscr{D}_{video}$ is prone to over-fitting.

To mitigate these problems we formulate the learning problem as a form of transfer learning called *domain adaptation* (*e.g.*, see [33]). 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.*, [3,8]). The resulting models generalize better than those learned from the video-based pose data directly, and they produce better results than the direct application of models learned from $\mathscr{D}_{mocap}$.

It is important to note that, prior to learning the initial models from the source motion capture data, $\mathscr{D}_{mocap}$, the pose data in $\mathscr{D}_{mocap}$ were converted into the parameterization used by the tracker in $\mathscr{D}_{video}$. We make this conversion by simply dropping the degrees of freedom from certain joints (*e.g.*, for a knee we only take rotation angle in a sagittal plane and drop the remaining two degrees of freedom). Second, note that the PCA basis used for the latent representation of the (converted) mocap data in $\mathscr{D}_{mocap}$ is also used as the latent basis for the data in $\mathscr{D}_{video}$.

4.1. Logistic classifier with transfer learning

In more detail, we use logistic regression for the inference of binary attributes. 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.*,

$$p(g = 1|\mathbf{c}, \theta) = \frac{1}{1 + \exp(-\mathbf{c}^T\mathbf{w} - b)} \equiv \sigma(\mathbf{c};\ \mathbf{w},\ b). \quad (2)$$

where $\mathbf{c}$ is the vector of model subspace coefficients. The weights that define the decision hyperplane are found by maximum likelihood. For example, given IID *source* mocap data of subspace coefficients and gender g, $\mathscr{D} = \left\{\mathbf{c}_j^s,\ g_j^s\right\}_{j=1}^{N_s}$, the data likelihood is

$$p(\mathscr{D}) = \prod_{j=1}^{N_s} p(g = g_j|\mathbf{c}_j, \theta). \quad (3)$$

With some manipulations the negative log likelihood of the source data becomes

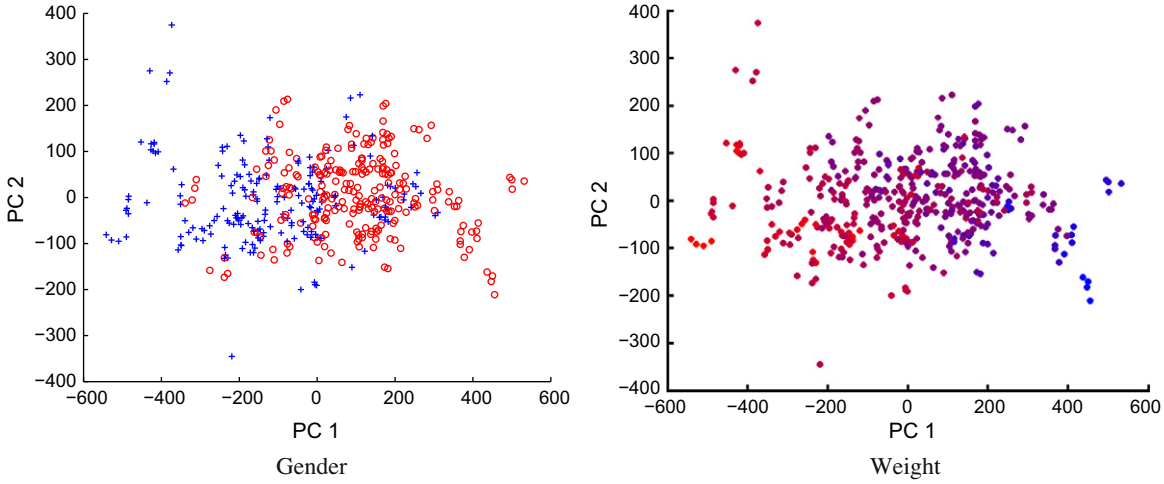


Fig. 3. Subspace visualization: The distribution of motions in $\mathscr{D}_{mocap}$ in the first 2 principal dimensions (out of the total of 16) is shown. (Left) Males (+) and females (o). (Right) Weight is depicted with blended colors: Heavy (red) and light (blue). Notice that, although not completely seperable in just the first two dimensions, there is a clear relation between the measured attribute and the subspace coefficients.

$$\mathscr{L}_s(\mathbf{w},b) = -\log\prod_{j=1}^{N_s}\sigma\left(\mathbf{c}_j^s;\mathbf{w},b\right)^{g_j^s}\left(1-\sigma\left(\mathbf{c}_j^s;\mathbf{w},b\right)\right)^{1-g_j^s}. \tag{4}$$

The parameters are found by minimizing the negative log likelihood, *i.e.*, $\theta^s = (\mathbf{w}^s, b^s) = \arg\min \mathscr{L}_s$.

While such source models perform well on other test mocap data, they do not produce good predictions when applied to 3D pose data from video tracking. To adapt the model learned from $\mathscr{D}_{mocap}$ to the target data $\mathscr{D}_{video}$, following [8], 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 comprises the negative log likelihood of the video training data, $\left\{\mathbf{c}_j^t, g_j^t\right\}_{j=1}^{N_t}$, and a quadratic regularizer:

$$\mathscr{L}_t(\mathbf{w},b) = -\log\prod_{j=1}^{N_t}\sigma\left(\mathbf{c}_j^t;\mathbf{w},b\right)^{g_j^t}\left(1-\sigma\left(\mathbf{c}_j^t;\mathbf{w},b\right)\right)^{1-g_j^t} + \lambda\|\mathbf{w}-\mathbf{w}^s\|^2. \tag{5}$$

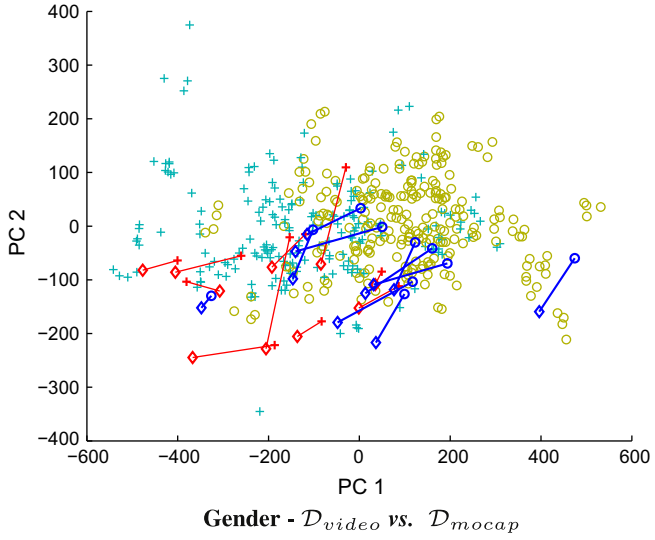


Fig. 4. Video and mocap consistency – gender: video pose tracks and mocap from 10 subjects in $\mathscr{D}_{video}$, along with $\mathscr{D}_{mocap}$ are shown in two subspace dimensions. $\mathscr{D}_{mocap}$: +(male), o (female). $\mathscr{D}_{video}$: +(male, mocap), ◇ (male, tracking), o (female, mocap), ◇ (female, tracking). Notice the bias with the $\mathscr{D}_{video}$ subspace representation in the $\mathscr{D}_{mocap}$ subspace. This motivates the need for transfer learning.

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 anisotropic prior might be set according to the ratio of variances in the subspace projections of $\mathscr{D}_{mocap}$ and $\mathscr{D}_{video}$ respectively. Nevertheless the experiments reported below are based on an isotropic prior.

Minimization of $\mathscr{L}_t$ is accomplished with Newton iterations to solve for critical points, *i.e.*,

$$\frac{\partial\mathscr{L}_t}{\partial\mathbf{w},b} = \sum_{j=1}^{N_t}\left(\sigma\left(\mathbf{c}_j^t;\mathbf{w},b\right)-g_j^t\right)\begin{pmatrix}\mathbf{c}_j^t\\1\end{pmatrix}+\lambda\begin{pmatrix}\mathbf{w}-\mathbf{w}^s\\0\end{pmatrix}=\mathbf{0}. \tag{6}$$

Leave-one-out 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.

One can generalize the approach to model the ratings data by replacing the ground truth g in (5) with the average rating (scaled to $(0,1)$). Treating the average rating as the expected value of g over different observers, (5) can be interpreted as the expected likelihood of the data. While the approach formulated here presupposes labeled 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.*, see [3]).

4.2. LS regressors with transfer learning

The same form of domain adaptation can also be applied to help learn models for predicting real-valued attributes, such as age or weight. For example, let $\mathbf{w}_{LS}^s$ be the least-squares optimal weight vector for a linear regressor that predicts an attribute, a, from the subspace representation of the mocap data in $\mathscr{D}_{mocap}$. This provides the source model.

Using domain adaptation we can then formulate the target model in terms of a least-squares predictor for weight from the video-based pose tracking data in $\mathscr{D}_{video}$. This is just another least-squares optimization, but with a regularizer that biases the weight vector toward the source model parameters, $\mathbf{w}_{LS}^s$. That is, the adapted LS predictor for real-valued attribute a minimizes

$$\mathscr{L}_c(\mathbf{w},b) = \sum_{j=1}^{N_t}\left(\mathbf{w}^T\mathbf{c}_j^t+b-a_j^t\right)^2+\lambda\|\mathbf{w}-\mathbf{w}_{LS}^s\|^2. \tag{7}$$

5. Attribute inference from $\mathcal{D}_{mocap}$

We begin with the models learned from the labeled source mocap data $\mathcal{D}_{mocap}$. This includes classifiers for gender, and regressors for predicting age and weight, and models for predicting perceived human ratings. For gender classification, with a 16 dimensional subspace representation, we obtain a correct classification rate of 90% (based on leave-one-our validation testing, see Fig. 8). For weight regression, with a 16 dimensional subspace representation, we obtain a RMSE of 5.4 kg (see Fig. 9).

Fig. 7 (left) shows how gender classification depends on the subspace dimension of the motion representation. With fewer than 16 dimensions important information is lost. Classification performance with more than 20 dimensions yields marginal gains; with a 16D subspace the correct classification rate for gender is over 90%. Fig. 7 (middle) shows the behavior of a LS predictor for weight. The weights of our 115 walking subjects ranged from 50 to 100 kg, while the RMSE of predictions (leave-one-out testing with 16D features) is 5.4 kg. Fig. 7 (right) shows that gender can be classified with as little as one gait cycle (consistent with human perception [18]).

As described above, gender classification is based on logistic regression with a planar decision boundary:

$$\mathbf{w}^T\mathbf{c} + b = \mathbf{w}^T\mathbf{B}\mathbf{m} + b = 0, \tag{8}$$

where $\mathbf{w}$ and b are the weight vector and bias in (2), $\mathbf{c}$ are subspace coefficients of the motion model, $\mathbf{B}$ is the subspace basis, and $\mathbf{m}$ is the 226-vector in the joint-based Fourier series motion representation. Fig. 5 (left) depicts the weights $\mathbf{w}^T\mathbf{B}$ in the Fourier series representation. The size and color of disk at each joint depicts the relative magnitudes of the weights in the X, Y, and Z directions. From the figure one can see that gender classification relies heavily on body shape, while the motion coefficients (1st, 2nd harmonics) play a somewhat lesser role. Fig. 5 (right) shows how the weights decrease from strongest to weakest; the most dominant features are those corresponding to lateral shape and motion (perpendicular to the sagittal plane), consistent with studies of human perception [43].

Fig. 6 depicts the feature weights used in least-squares weight prediction. In contrast to gender classification, weight prediction relies more significantly on motion than the mean pose. This is understandable since a person's weight may not affect their skeletal structure, but may affect the way they walk since soft tissue may restrict certain movements. It is also interesting to note that most of the motion features have weights of similar magnitude (see *cf.* Fig. 5 (right)). Similar results appear with the inference of age.

5.1. Normalized models

When inferring attributes from video motion estimates, we may not have access to full 3D pose. For example, with monocular tracking one might be able to 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.*, [4]). In extreme cases a pose tracker may have no anthropometric knowledge whatsoever. To explore these cases we computed two further subspace representations of the mocap corpus, $\mathcal{D}_{mocap}$. In one model, all walkers were normalized to be the same height. The fact that males are on average taller than females is therefore lost. In the other model, the individual anthropometrics are removed so that motion is the only remaining cue. The removal of the anthropometrics was accomplished by converting joint positions into joint angles, and then using the mean anthropometrics from the subject pool to convert back into joint positions to reconstruct the motions. This removes both height differences as well as other information that might discriminate gender, such as the distance between shoulders or hips.

The first row of results in Fig. 8 reports the correct classification rate for gender, determined using leave-one-out validation. Compared to a 90% hit rate for the 3D model, the height-normalized model has a hit rate of 83%, while the performance of the anthropometrically-normalized model drops to 82%.

The first row of Fig. 9 gives the RMSE of linear least-squares weight and age predictors, again based on leave-one-out testing. Within the training mocap corpus, $\mathcal{D}_{mocap}$, weights vary between 45 and 110 kg, with a standard deviation of 12.5 kg. RMSE for the 3D model is 5.4 kg, increasing to 10.9 kg for the anthropometrically-normalized model. When comparing weight to age inference, age varied between 13 and 43 years with a standard deviation of 6.6 years. Age is predicted with a RMSE of 6.9 years, larger than the population standard deviation. Thus, while weight is predicted with reasonable accuracy, it is clear that age is poorly predicted.

5.2. Incomplete and unreliable data

To infer attributes from video-based pose estimates, one might wish to be able to cope with missing data, since parts of the body may be occluded. It is therefore of interest to ask how performance degrades when partial data is available. To that end, let $\mathbf{m} \in \mathbb{R}^{226}$ 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 $\mathbf{P}$ comprises only those rows of the identity matrix that correspond to the observed joints. It then follows from the gener-

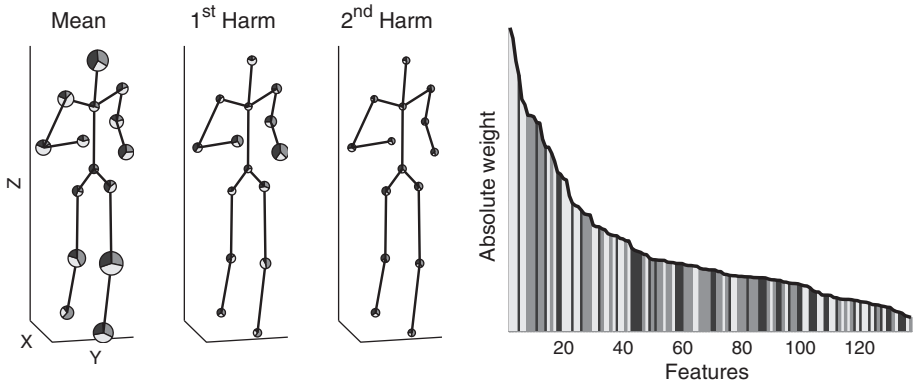


Fig. 5. Feature weights in gender classification: $X/Y/Z$ dimensions are mapped respectively to Dark gray/Light gray/Gray. (Left) Our model is depicted with joints size radius proportional to the weight of that joint (per different harmonics). Within each joint, a pie diagram represents which dimension ($X/Y/Z$) was the most important. (Right) The corresponding bar chart lists all features sorted according to importance, where Dark gray represents motions in the X direction (normal to the coronal plane), and Light gray represents motion in the Y direction (normal to the sagittal plane).

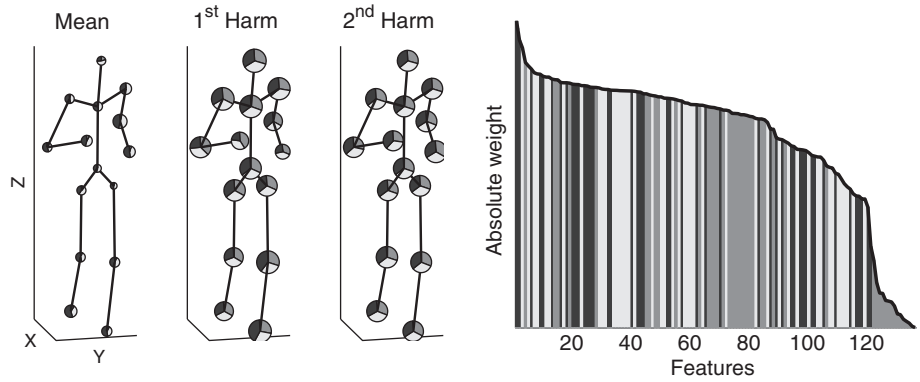


Fig. 6. Feature weights for weight regression: *X*/*Y*/*Z* dimensions are mapped respectively to Dark gray/Light gray/Gray. (Left) Our model is depicted with joints size radius proportional to the weight of that joint (per different harmonics). Within each joint, a pie diagram represents which dimension (*X*/*Y*/*Z*) was the most important. (Right) The corresponding bar chart lists all features sorted according to importance, where Dark gray represents motions in the *X* direction (normal to the coronal plane), and Light gray represents motion in the *Y* direction (normal to the sagittal plane).

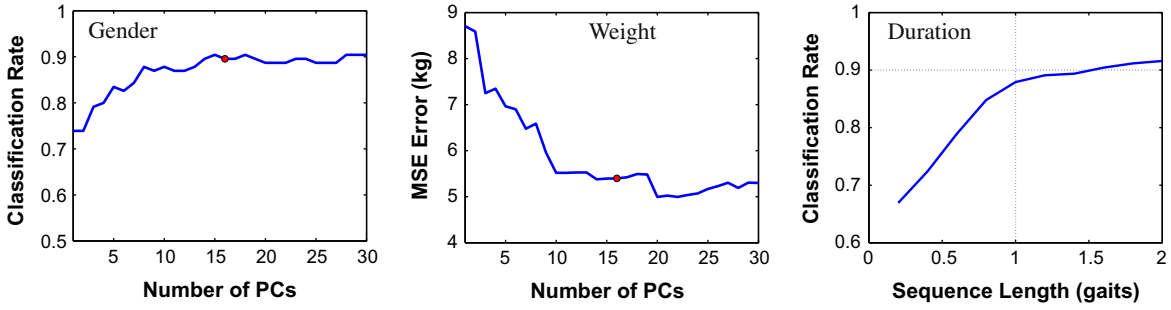


Fig. 7. Effect of subspace dimension and sequence length: Leave-one-out validation is used to asses the effect of subspace dimension on the classification performance for gender classification (left) and the RMSE of the least-squares weight regressor (middle). The right plot shows the dependence of gender classification on the duration (in gait cycles) of mocap data (based again on leave-one-out testing).

ative 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.*,

$$\mathbf{c}_0 = (\mathbf{B}^T\mathbf{P}^T\mathbf{P}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}^T(\mathbf{m}_0 - \mathbf{P}\bar{\mathbf{m}}), \tag{9}$$

where **B** is the basis matrix of principal directions, and **c** is the corresponding coefficient matrix.

The columns in Figs. 8 and 9 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 uses only 2D data from frontal and sagittal views under orthographic projection. It is clear that performance deteriorates when fewer measurements are available.

	Gender (success rate)		
Data \ Model	3D Model	Height Norm.	Motion Only
All Markers	0.90	0.83	0.82
Upper Body	0.88	0.79	0.77
Lower Body	0.75	0.77	0.76
Frontal 2D Pose	0.84	0.76	0.70
Sagittal 2D Pose	0.86	0.83	0.75

Fig. 8. Gender inference with $\mathscr{D}_{mocap}$ 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), as explained in Section 5.1. The lack of anthropometrics degrades performance, but the inference of gender is 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.

5.3. Predicting human ratings

It is also interesting to consider how well one can predict aspects of mental (or emotional) state. To this end we consider the prediction of *perceived attributes*. While physical attributes like gender, age and weight have ground truth values, mood for instance, has no physical ground truth per se. Rather, in our case, the perceived mood is our only source of labeled data. Because our perceptual rating data are noisy, we first quantize the human 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 empirical 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. 10.

It is striking that, in all cases, our classifiers are remarkably consistent in predicting human ratings. In most cases they do as well or better than classifiers that predict ground truth attributes (*e.g.*, gender). Human observers are purportedly using the available visual cues in a consistent manner, even when it might be inconsistent with the ground truth. And this overt information is captured in our subspace representation. In particular, while true age is 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 perceived attributes may be biased, and thus should be interpreted with care.

Data \ Model	Weight (RMSE kg)			Age (RMSE yrs)		
	3D Model	Height Norm.	Motion Only	3D Model	Height Norm.	Motion Only
All Markers	5.44	9.78	10.86	6.89	6.84	6.63
Upper Body	5.91	10.19	11.14	6.92	6.74	6.46
Lower Body	6.37	9.52	12.49	7.44	7.46	7.57
Frontal 2D Pose	5.59	9.79	10.82	7.08	7.12	6.87
Sagittal 2D Pose	10.07	11.41	12.26	7.02	6.95	6.91

Fig. 9. Weight and age inference with $\mathscr{D}_{mocap}$ models: To assess performance, with and without missing data, we build three models (as explained in Fig. 8). Note that the full 3D model is always the best or close to the best result. We also report how performance varies with different subsets of markers (*e.g.*, upper/lower body) or 2D projections. Weight predictions are reasonably good, especially for the full 3D model, but age is predicted poorly in all cases when compared the population standard deviation of 6.6 years (Section 5.1).

	Gender	Weight	Age	Mood
Full 3D	93	94	89	94
Height Normalized	92	94	89	95
Motion Only	93	94	87	94

Fig. 10. 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 $\mathscr{D}_{mocap}$. 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. It is also interesting to note that the people from whom ratings were obtained had not absolute 3D data available to them since they viewed only 2D projections (on a monitor) of the true 3D motion. So one might not expect 3D data to provide much additional information over height normalized data.

6. Attribute inference from $\mathcal{D}_{video}$

Given the source models learned from $\mathscr{D}_{mocap}$, we use domain adaptation to learn models for the video-based motion data in $\mathscr{D}_{video}$, as explained in Section 4. 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 $\mathscr{D}_{video}$. The reason is that the mocap and video-based pose data in $\mathscr{D}_{video}$ are parametrized differently from that in $\mathscr{D}_{mocap}$. The source pose data $\mathscr{D}_{mocap}$ allows for variable joint locations, and three rotational DOFs for all joints. To simplify state estimation during video tracking, the body parameterization used in video tracking, as well as the mocap in $\mathscr{D}_{video}$, has fixed joint locations, the knees and elbows have only one roational DOF, and the shoulders have fewer degrees of movement. Hence, there are structural differences even between the mocap in $\mathscr{D}_{mocap}$ and that in $\mathscr{D}_{video}$. Finally, as discussed above, it is also clear that the 3D pose data based on video pose tracking is also much noisier than the source pose data in $\mathscr{D}_{mocap}$.

Fig. 11 (left) show the leave-one-out performance for gender classification based on the mocap in $\mathscr{D}_{video}$, with domain adaptation from $\mathscr{D}_{mocap}$. The curves show how performance depends on adaptation from the source model, as a function of λ (see (5) in Section 4). The highest hit rates occur with λ between 10^3 and 10^4. For comparison, the crosses (x) depict hit rates when there is no domain adaptation (*i.e.*, with $\mathbf{w}^s = \mathbf{0}$ in Eq. (5)). The circles (o) depict hit rates when the classifiers are trained solely on the source data $\mathscr{D}_{mocap}$ (with no domain adaptation) and then tested on the mocap in $\mathscr{D}_{video}$. Remember that, due to different parameterizations of body, the mocap features in $\mathscr{D}_{mocap}$ and $\mathscr{D}_{video}$ are distributed differently.

6.1. Target pose tracking data

Fig. 11 (right) shows leave-one-out performance for gender classification from the video-based 3D pose tracking data (two trials of the APF, for each of 2 walking sequences for each of 24 subjects). As above, the curves show the dependence on the strength of the prior from the source model. The crosses (x) depict the performance rates when trained only on the pose track data, with no domain adaptation. The circles (o) depict hit rates from classifiers trained solely on the source mocap data $\mathscr{D}_{mocap}$.

Fig. 12 reports numerical results for gender classification, from both the mocap and the video-based data in $\mathscr{D}_{video}$ (*cf.* Fig. 11). As above, we show results from three models: C_{mocap} is learned solely from the source mocap $\mathscr{D}_{mocap}$; C_{track} is learned solely from test

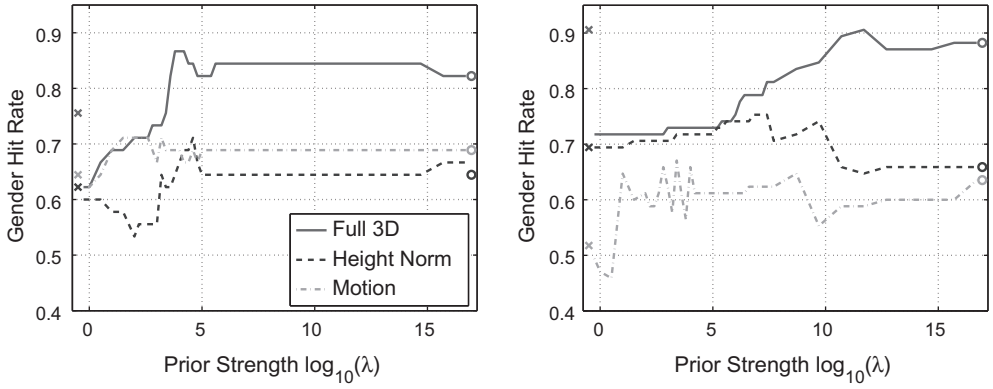


Fig. 11. Domain adaptation (gender): (left) Gender classification from the *mocap* data in $\mathscr{D}_{video}$ for 24 test subjects, as a function of the strength of the prior λ, for each of three models (full 3D, height normalized, motion only). (right) Gender classification from the video-based *pose tracking data*. All results were generated using hit rates computed using leave-one-out validation. The crosses (x) depict hit rates when trained on the video-based pose tracking data, without domain adaptation. The circles (o) depict hit rates when the classifiers are trained solely on the source data $\mathscr{D}_{mocap}$ (with no domain adaptation).

		Gender (%correct)		
		3D Model	Height Norm.	Motion Only
mocap	C_{mocap}	0.82	0.64	0.69
	C_{track}	0.62	0.62	0.64
	$C_{trackTL}$	0.87	0.71	0.71
	$\mathbf{p_{min}}$	0.79	0.62	0.62
tracking	C_{mocap}	0.51	0.56	0.55
	C_{track}	0.41	0.55	0.51
	$C_{trackTL}$	0.68	0.70	0.59
	$\mathbf{p_{min}}$	0.58	0.6	0.49

Fig. 12. Gender inference from Mocap and pose tracking data: The table gives leave-one-out performance for gender classification from mocap and pose tracking data in $\mathscr{D}_{video}$. There are 46 mocap sequences (~2 walks/subject), and 86 pose trajectories from video tracking (~2 tracking trials per sequence). Results from three models are reported: C_{mocap} is learned from the source mocap in $\mathscr{D}_{mocap}$; C_{track} is learned solely from $\mathscr{D}_{video}$ data; $C_{trackTL}$ is learned with $\mathscr{D}_{video}$ and domain adaptation from $\mathscr{D}_{mocap}$. p_{min} is the minimum success rate of a classifier that would generate results as good or better than ours 9 out of 10 times.

(mocap and video) data $\mathscr{D}_{video}$; $C_{trackTL}$ is learned from $\mathscr{D}_{video}$ with domain adaptation from $\mathscr{D}_{mocap}$. It is clear that transfer learning $C_{trackTL}$ yields either the best result, or close to best results in each case. For the target mocap data, we find correct classification rates for gender at 87% for the full 3D model, very close to results on the source data. Results for the height-normalized model and the anthropometric-normalized model are not quite as good, both with hits rates of 68%.

To provide a measure of statistical significance, Fig. 12 also includes the value of p_{min} for each model. Let k be the number of correct classifications out of n trials, and p be the true probability of correct classification. We assume that k conditioned on p, n is Binomial, and that we have uniform prior over p. We then define p_{min} to satisfy

$$\int_{p_{min}}^{1} f(p|k,n)\,dp = 0.9. \tag{10}$$

That is, p_{min} represents the lowest probability of correct classification such that in 9 out of 10 runs of our experiment we should get results at least as good as those we obtained in the leave-one-out scores. When p_{min} is bigger than 0.5 it confirms that our classifier is almost certainly not performing at chance on this data.

Fig. 13 (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 $\mathscr{D}_{mocap}$ or in $\mathscr{D}_{video}$. With domain adaptation, with $\lambda = 10^6$, the RMSE decreases to approximately 12 kg. In comparison to Fig. 13 (left),the mocap based prediction with best RMSE of approximately 6 kg, it is evident that the 3D tracker fails to capture some essential information that is needed for useful weight prediction. Fig. 14 provides numerical results for weight prediction. (*cf.* Fig. 13). As above, we show results from three models: C_{mocap} is learned solely from the source mocap $\mathscr{D}_{mocap}$; C_{track} is learned solely from test (mocap and video) data $\mathscr{D}_{video}$; $C_{trackTL}$ is learned from $\mathscr{D}_{video}$ with domain adaptation from $\mathscr{D}_{mocap}$.

6.2. Amplitude-based model

Clearly the results in Figs. 12 and 11(right) reveal that the results on the video-tracking pose data are not as good as those for the mocap data in $\mathscr{D}_{mocap}$. The major problem stems from noise in the pose tracking, much of which was manifested in the relative phases of some of the joint motion. This rendered the phase-based alignment of walking motions unreliable.

A simple but effective way of removing noise in the relative phases of different joint trajectories is to use only the amplitudes of the Fourier coefficients rather than their real and imaginary parts. That is, rather than a 226 dimensional Fourier description, with amplitude alone, the Fourier description has 136 dimensions (45 for the mean pose, 90 for the amplitudes of the first two harmonics, and the fundamental frequency). We then apply dimensionality reduction and learn the classifiers as above. The net result is a significant improvement in gender classification from the video-based pose data. Fig. 15 shows result with domain adaptation applied to a subspace representation of the amplitude coefficients. The corresponding numerical results are reported in Fig. 16. Note that the classification rates for both the video data and the mocap data are improved. Indeed, the performance on the video-based approaches the performance on the mocap data. Interestingly, prediction of ground truth weight and age did not improve appreciably with an amplitude-based motion represetation.

6.3. Inference of perceived attributes

Next we turn to consider how well we can infer perceived attributes. To this end, Fig. 17 reports leave-one-out hit rates in

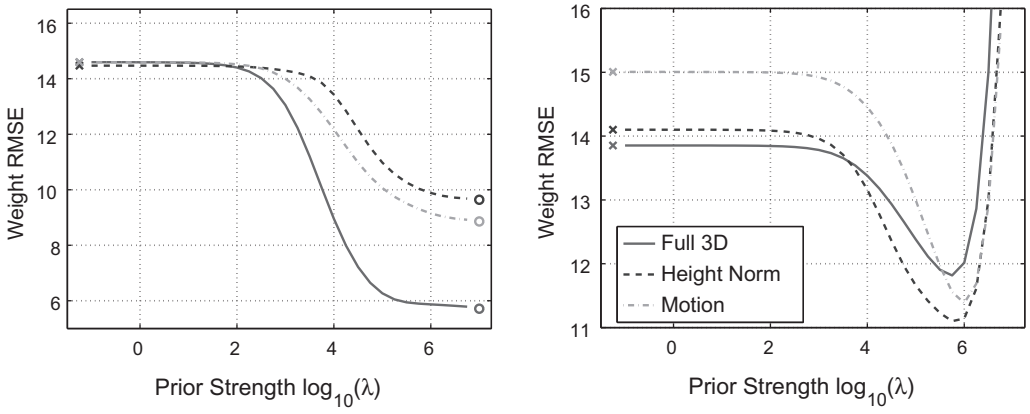


Fig. 13. Domain adaptation (weight): (right) RMSE of weight estimates from *mocap*, for 24 test subjects, as a function of the strength of the prior for each of 3 models (full 3D, height normalized, motion only). (left) RMSE of weight estimates from video-based *pose tracking data*. All results were generated using leave-one-out testing. The crosses (x) depict hit rates when there is no domain adaptation. The circles (o) depict hit rates when the classifiers are trained solely on the source data $\mathscr{D}_{mocap}$ (with no domain adaptation).

		Weight (RMSE yrs)		
		3D Model	Height Norm.	Motion Only
mocap	C_{mocap}	5.84	9.78	9.04
	C_{track}	14.31	14.34	14.51
	$C_{trackTL}$	5.91	9.81	9.09
tracking	C_{mocap}	62.42	51.53	50.69
	C_{track}	13.73	14.10	15.03
	$C_{trackTL}$	11.71	11.02	11.28

Fig. 14. Weight inference from mocap and pose tracking data: The tables reports leave-one-out cross-validation performance on weight prediction from mocap and pose tracking data in the $\mathscr{D}_{video}$ data-set of 24 subjects. There are 46 mocap sequences (∼2 walks/subject), and 86 pose trajectories from video tracking (∼2 tracking trials per sequence). Results from three models are reported: C_{mocap} is learned from the source mocap corpus $\mathscr{D}_{mocap}$; C_{track} is learned solely from $\mathscr{D}_{video}$ data; $C_{trackTL}$ is learned with $\mathscr{D}_{video}$ and domain adaptation from $\mathscr{D}_{mocap}$.

		Gender (%correct)		
		3D Model	Height Norm.	Motion Only
mocap	C_{mocap}	0.88	0.70	0.67
	C_{track}	0.77	0.42	0.42
	$C_{trackTL}$	**0.93**	**0.70**	**0.67**
	p_{min}	0.86	0.6	0.58
tracking	C_{mocap}	0.88	0.66	0.64
	C_{track}	0.91	0.69	0.51
	$C_{trackTL}$	**0.91**	**0.75**	**0.67**
	p_{min}	0.84	0.67	0.58

Fig. 16. Gender inference from mocap and pose tracking data (amplitude representation): The table reports leave-one-out hit rates for gender classification, as in Fig. 12. However, instead of representing motions by real and imaginary Fourier coefficients, an amplitude-only model was used. Again, as in Fig. 12, p_{min} is the minimum success rate of a classifier that will generate our results in 9 out of 10 runs.

the prediction of four *perceived* attributes, namely gender, weight, age and mood. Like the above experiment in Fig. 10 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 $\mathscr{D}_{video}$.

The consistency between the mocap and pose tracking is indeed good, with consistent classification rates between 76% and 98%. It is interesting to note that we can recover the mental state – mood (happiness), with up to 93% accuracy, and age with up to 98%. The high rates of consistency demonstrates how perceived attributes are predicted better than ground truth.

	$C_{trackTL}$ (% correct)			
	Gender	**Weight**	**Mood**	**Age**
3D Model	80	80	87	87
Height Normalized	85	83	91	93
Motion Only	93	76	93	98

Fig. 17. Classification of perceived attributes with respect to mocap: The table reports consistency of leave-one-out cross-validation performance on *perceived* gender, weight, attractiveness, mood (happiness) and age between mocap and pose tracking data in the target data-set $\mathscr{D}_{video}$ of 24 test subjects. We predict $\mathscr{D}_{video}$ mocap attribute values by using $\mathscr{D}_{mocap}$ models. We then use the predicted attribute values as ground-truth targets to train $C_{trackTL}$ binary classifiers (learned with $\mathscr{D}_{video}$ and domain adaptation from $\mathscr{D}_{mocap}$).

7. Discussion

This paper demonstrates that one can infer significant physical attributes (*e.g.*, gender and weight) and aspects of mental state (*e.g.*, happiness) from the output of a video-based, 3D human pose tracker. The models are used to infer binary attributes (gender) and real-valued attributes (*e.g.*, weight). We also consider the prediction of perceived attributes based on human perceptual experiments. This is useful when inferring attributes such as mood where human perception is our source of ground truth. Learning is accomplished using data sets comprising labeled mocap and video-based 3D pose estimates. These sources of training data are combined with a simple forms of domain adaptation. In addition, we demonstrate that current state-of-the-art tracking methods are capable of matching accurate mocap data in predicting at least

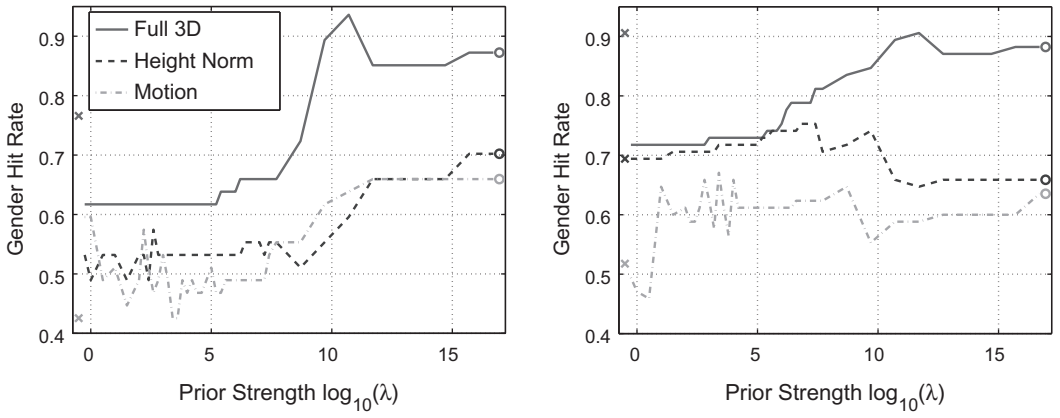


Fig. 15. Amplitude-based model with domain adaptation (gender): (left) Gender classification from *mocap* of $\mathscr{D}_{video}$ for 24 test subjects, as a function of the strength of the prior λ, for each of 3 models (full 3D, height normalized, motion only). (left) Gender classification from the video-based *pose tracking data*. All results were generated using leave-one-out hit rates. Notice how by using the amplitude model, classification rate of video-based pose tracking data matches that of the mocap based data. The crosses (x) depict hit rates when there is no domain adaptation. The circles (o) depict hit rates when the classifiers are trained solely on the source data $\mathscr{D}_{mocap}$ (with no domain adaptation).

certain attributes (*e.g.*, gender) by using an appropriate model. However, a more reliable video tracking is needed in order to capture a more subtle attributes such as weight and age, or at least a one that captures the required information with high enough SNR ratio.

We also demonstrate how predicting perceived attributes is more accurate than predicting ground truth in gender, weight and age. In other words, it does not matter how old you really are, it is more important whether you move as an old or a young person. In addition, we show that relying on Euclidean RMSE alone is illusive and unreliable indicator for many real-life tasks of a 3D pose tracker (Section 6.2). Instead we suggest using attribute inference as an additional quality indicator which can indicate whether essential motion information is preserved.

In the future we hope to collect large data sets 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. One possible pose tracking method is to estimate attributes as part of the tracking process, in order to create a better generative model. By doing so, both Euclidean tracking accuracy and attribute inference might improve. While the results reported here are interesting in their own right, this is one of the first papers to suggest that tasks like as attribute inference provide a natural way to assess the fidelity with which people trackers estimate 3D pose.

Acknowledgments

This work was financially supported in part by NSERC Canada, the Canadian Institute for Advanced Research, the GRAND Network Centre of Excellence, and the University of Toronto.

References

[1] G. Amayeh, G. Bebis, M. Nicolescu, Gender classification from hand shape, in: Proceedings of IEEE Workshop on Biometrics, 2008.
[2] M. Andriluka, S. Roth, B. Schiele, Monocular 3d pose estimation and tracking by detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[3] A. Arnold, R. Nallapati, W. Cohen, A comparative study of methods for transductive transfer learning, in: Proceedings of ICDM Workshop on Mining and Management of Biological Data, 2007.
[4] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker, Detailed human shape and pose from images, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[5] S. Blakemore, J. Decety, From the perception of action to the understanding of intention, Nature Reviews Neuroscience 2 (8) (2001) 561–567.
[6] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.
[7] J. Boyd, J. Little, Biometric gait recognition, Adv. Stud. Biomet.: Summer School Biomet. (2003).
[8] C. Chelba, A. Acero, Adaptation of maximum entropy capitalizer: little data can help a lot, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2004.
[9] J.E. Cutting, L.T. Kozlowski, Recognizing friends by their walk: gait perception without familiarity cues, Bull. Psychonom. Soc. 9 (5) (1977) 353–356.
[10] J.E. Cutting, D.R. Proffitt, L.T. Kozlowski, A biomechanical invariant of gait perception, J. Experim. Psychol.: Human Percep. Perf. 4 (1978) 357–372.
[11] J. Deutscher, I. Reid, Articulated body motion capture by stochastic search, Int. J. Computer Vision 61 (2) (2005) 185–205.
[12] Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, Recognizing action at a distance, in: Proceedings of IEEE International Conference on Computer Vision, 2003.
[13] X. Geng, Z.H. Zhou, K. Smith Miles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234–2240.
[14] L. Gorelick, M. Blank, E. Shechtman, M. Irani, Ronen Basri, Actions as space–time shapes, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2247–2253.
[15] G. Huang, Y. Wang, Gender classification based on fusion of multi-view gait sequences, in: Proceedings of Asian Conference on Computer Vision, 2007, pp. 462–471.
[16] A. Jepson, D.J. Fleet, T. El-MAraghi, Robust online appearance models for vision tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1296–1311.
[17] G. Johansson, Visual perception of biological motion and a model for its analysis, Percep. Psychophys. 14 (2) (1973) 201–211.
[18] G. Johansson, Spatio-temporal differentiation and integration in visual motion perception, Psychol. Res. 38 (1976) 379–393.
[19] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: Proceedings of IEEE International Conference on Computer Vision, 2007.
[20] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.
[21] Ivan Laptev, Tony Lindeberg, Space–time interest points, in: Proceedings of IEEE International Conference on Computer Vision, 2003, pp. 432–439.
[22] L. Lee, E. Grimson, Gait analysis for recognition and classification, in: Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, 2002.
[23] M. Lemke, T. Wendorff, B. Mieth, K. Buhl, M. Linnemann, Spatiotemporal gait patterns during over ground locomotion in major depression compared with never depressed controls, J. Psychiat. Res. 34 (2000) 277–283.
[24] Rui Li, Tai-Peng Tian, Stan Sclaroff, Simultaneous learning of non-linear manifold and dynamical models for high-dimensional time series, in: Proceedings of IEEE International Conference on Computer Vision, 2007.
[25] X. Li, S. Maybank, S. Yan, D. Tao, S. Xu, Gait components and their applications to gender recognition, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 38 (2) (2008).
[26] M.M. Toews, T. Arbel, Detection, localization, and sex classification of faces from arbitrary viewpoints and under occlusion, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1567–1581.
[27] G. Mather, L. Murdoch, Gender discrimination in biological motion displays based on dynamic cues, Proc. Roy. Soc. London Ser. B 258 (1994) 273–279.
[28] J. Michalak, N. Troje, J. Fischer, P. Vollmar, T. Heidenreich, D. Schulte, The embodiment of sadness and depression – gait patterns associated with dysphoric mood, Psychosom. Medicine 71 (2009) 580–587.
[29] G.W. Mu, G.D. Guo, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[30] J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, Int. J. Computer Vision 79 (3) (2008) 299–318.
[31] H. Ning, W. Xu, Y. Gong, T.S. Huang, Latent pose estimator for continuous action recognition, in: Proceedings of European Conference on Computer Vision, 2008, pp. 419–433.
[32] D. Ormoneit, H. Sidenbladh, M. Black, T. Hastie, D.J. Fleet, Learning and tracking human motion using functional analysis, in: Proceedings of IEEE Workshop on Human Modeling, Analysis and Synthesis, 2000.
[33] S. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[34] F. Pollick, J. Kay, K. Heim, R. Stringer, Gender recognition from point-light walkers, J. Experim. Psychol.: Human Percep. Perf. 31 (6) (2005) 1247–1265.
[35] S. Samangooei, M. Nixon, Performing content-based retrieval of humans using gait biometrics, Multimedia Tools Appl. (2009).
[36] S. Sarkar, J. Phillips, Z. Liu, I. Robledo, P. Grother, K. Bowyer, The human id gait challenge problem: data sets, performance, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 162–177.
[37] H. Sidenbladh, M. Black, D.J. Fleet, Stochastic tracking of 3d human figures using 2d image motion, in: Proceedings of European Conference on Computer Vision, vol. 2, 2000, pp. 702–718.
[38] L. Sigal, A. Balan, M. Black, Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Int. J. Computer Vision 87 (1) (2010) 4–27.
[39] C. Sminchisescu, A. Jepson, Generative modeling for continuous non-linearly embedded visual inference, in: Proceedings of International Conference on Machine Learning (ICML), 2004, pp. 759–766.
[40] C. Sminchisescu, A. Kanaujia, D. Metaxas, BM^3E: discriminative density propagation for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 2030–2044.
[41] G.W. Taylor, L. Sigal, D.J. Fleet, G.E. Hinton, Dynamical binary latent variable models for 3d human pose tracking, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 631–638.
[42] C. Thurau and V. Hlavac, Pose primitive based human action recognition in videos or still images, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[43] N. Troje, Decomposing biological motion: a framework for analysis and synthesis of human gait patterns, J. Vision 2 (5) (2002) 371–387.
[44] N. Troje, Retrieving information from human movement patterns, in: Understanding Events: How Humans See, Represent, and Act on Events, 2008, pp. 308–334.
[45] N. Troje, C. Westhoff, M. Lavrov, Person identification from biological motion: effects of structural and kinematic cues, Percept. Psychophys. 67 (4) (2005) 667–675.
[46] R. Urtasun, D.J. Fleet, P. Fua, 3D people tracking with Gaussian process dynamical models, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 238–245.
[47] R. Urtasun, D.J. Fleet, P. Fua, Motion models for 3d people tracking, Comput. Vision Image Understand. 104 (2–3) (2006) 157–177.
[48] Liang Wang, David Suter, Informative shape representations for human action recognition, in: Proceedings of International Conference on Pattern Recognition, vol. 2, 2006, pp. 1266–1269.

[49] Geert Willems, Tinne Tuytelaars, Luc Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proceedings of European Conference on Computer Vision, 2008.
[50] Weilong Yang, Yang Wang, Greg Mori, Recognizing human actions from still images with latent poses, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[51] J.-H. Yoo, D. Hwang, M. Nixon, Gender classification in human gait using support vector machine, Adv. Concepts Intell. Vision Syst. (2006).
[52] R. Zhang, C. Vogler, D. Metaxas, Human gait recognition, in: IEEE Workshop on Articulated and Nonrigid Motion, 2004.