ELSEVIER

Image and Vision Computing 21 (2003) 1001–1016

image AND vision COMPUTING

www.elsevier.com/locate/imavis

An expressive three-mode principal components model of human action style

James W. Davis*, Hui Gao

Computer Vision Laboratory, Department of Computer and Information Science, Ohio State University, 491 Dreese Lab, 2015 Neil Avenue, Columbus, OH 43210, USA

Received 13 September 2002; received in revised form 11 June 2003; accepted 26 June 2003

Abstract

We present a three-mode expressive-feature model for representing and recognizing performance styles of human actions. A set of style variations for an action are initially arranged into a three-mode data representation (body pose, time, style) and factored into its three-mode principal components to reduce the data dimensionality. We next embed tunable weights on trajectories within the sub-space model to enable different context-based style estimations. We outline physical and perceptual parameterization methods for choosing style labels for the training data, from which we automatically learn the necessary expressive weights using a gradient descent procedure. Experiments are presented examining several motion-capture walking variations corresponding to carrying load, gender, and pace. Results demonstrate a greater flexibility of the expressive three-mode model, over standard squared-error style estimation, to adapt to different style matching criteria.

Keywords: Action recognition; Style; Three-mode principal components; Motion analysis; Gesture recognition

1. Introduction

Human actions exhibit certain observable *styles*. For example, people can easily determine whether another person is walking 'in a hurry' as if late for a meeting, or if the person is just 'casually strolling along'. Other examples include seeing if a person is 'straining to carry' a heavy object rather than 'effortlessly holding it', or if someone is about to 'toss' a ball a short distance or preparing to 'throw' it a very long distance. People are quite adept at identifying such exertion movement qualities, even from impoverished visual cues such as point-lights and stick figures [14,34]. Furthermore, observable movement properties can be used to infer certain emotional qualities [31,40]. Other more internal physical origins can be the cause of some stylistic changes, such as the walking differences that occur between children and adults [17] or the walking styles due to gender [3,12,22,36]. Our goal is to develop an efficient computational framework capable of representing and recognizing such human action styles. We describe a multi-modal principal components method incorporating trainable feature weights to emphasize key feature trajectories during the style estimation process.

Computational models of action style are relevant to several important application areas. Automatic video annotation of descriptive scene behavior is a desired capability of advanced surveillance systems. Rather than just reporting 'OBJECT: IN-MOTION', it is desirable to generate more qualitative motion descriptions of the activity such as 'PERSON-WALKING: IN-A-HURRY' or 'PERSON-WALKING: CARRYING-HEAVY-OBJECT' which may help the system (or operator) to reason about higher-level behavior patterns. A system able to differentiate various action styles would also be relevant to ergonomic evaluation (e.g. for detecting improper heavy-lifting techniques to help reduce the (re-)occurrence of back injury), and for athletic training to help prevent injuries by recognizing the onset of fatigue (from movement) during endurance workouts. Another application is the re-use of motion-capture data by employing a style model to 'warp' animations into new styles [7,15]. Lastly, a style model could be used for searching digital motion libraries to find movements exhibiting a similar (or

* Corresponding author. Tel.: +1-614-292-1553; fax: +1-614-292-2911.
E-mail address: jwdavis@cis.ohio-state.edu (J.W. Davis).

0262-8856/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0262-8856(03)00138-0

different) style to the input query (e.g. 'Show me examples of people running *fast*').

One approach to recognizing action style variations is to match new movements to multitudes of training examples (of different style) based on their proximity in some global feature space. This approach would obviously be sensitive to features not directly associated with the style. Our belief is that certain visual cues have regularity across style variations that enable observers to reliably distinguish and recognize different style properties directly. Emphasizing these most expressive features during recognition would increase the ability of the system to reliably discriminate the style by focusing on the most predictive information. But which motion properties of an action give the strongest impression of certain stylistic changes? In this paper we present a weighted three-mode principal components model that learns the most expressive motion features (key features) needed for reliable recognition and matching of action style variations.

In our approach, training examples for an action style at different performance variations are placed into a 3D data cube representation, with each frontal plane corresponding to the motion trajectories of a movement at a particular style variation. The cube representation reflects the inherent three-mode nature of the data: body pose (mode-1), time (mode-2), and style (mode-3). As there typically exist a high amount of data redundancy, we reduce the dimensionality of the data cube into three small basis sets spanning the columns (body pose), rows (time), and slices (style). To achieve this multi-modal reduction, we apply a three-mode principal components factorization technique. This factorization is an extension of the standard matrix principal components method, and individually solves for a basis in each dimension by reshaping the data cube into three different 2D flattened matrix arrangements and applying standard principal components. A core cube is then solved which reconstructs the three basis sets back into the original data cube.

From data projections into the three-mode basis set, one could estimate the style for a new action using a standard least-squares match to the closest training example. To capitalize on the most predictive motion information (key features), we instead present a weighted-least-squares approach that is influenced by those motion trajectories that are most reflective of the assigned matching criterion. We outline a gradient descent method to learn the appropriate weight values given a collection of training examples with known style variations.

We present experimental results demonstrating the ability of the framework to model different stylistic actions of single and multiple people. We evaluate the approach with analysis and recognition of several motion-capture walking motions with style variations caused by (1) carrying load (light-to-heavy), (2) gender (male, female), and (3) walking pace (leisurely-to-quickly). Results demonstrate that the approach can successfully conform to different matching criteria of the action styles.

The remainder of this paper is presented as follows. In Section 2, we discuss related work on style analysis. The general three-mode factorization technique is presented in Section 3, and the expressive three-mode model is presented in Section 4. In Section 5, we describe the learning algorithm for the expressive weights using physical and perceptual matching criteria. In Section 6, we present experimental results demonstrating the approach. Lastly, we present the conclusion and future directions for this research in Section 7.

2. Related work

It is well known that certain types of biological movement patterns can be unambiguously recognized from their own organization of motion. People can easily recognize actions from limited types of input such as point-lights and blurred video [13,21]. But people can further differentiate stylistic action differences, such as the gender of the walker (even from viewing only two moving ankle points [22,23]) and the perceived exertion level of a person lifting a box [34]. Such perceptual motivations naturally lead to the desire to seek a computational model for addressing subtle, yet informative, stylistic motion variations.

There has been much recent work in computer vision on detecting, tracking, and recognizing human actions (see literature reviews [1,19,43]). With regards to style variation, a Parameterized-HMM was used by Ref. [44] to model spatial pointing gestures by adding a global variation parameter in the output probabilities of the HMM states. A bilinear model was used in Ref. [35] for separating perceptual content and style parameters, and was demonstrated with non-action two-mode examples including extrapolation of fonts to unseen letters and translation of faces to novel illuminates. In Ref. [17], an approach to discriminate children from adults based on variations in relative stride length and stride frequency over various walking speeds was presented. Additionally, in Ref. [16] the regularities in the walking motions for several people at different speeds were used to classify typical from atypical gaits. A two-mode principal components framework was described in Ref. [36] to linearly classify male and female walkers from trajectories of projection coefficients of the body pose. Morphable models were employed in Ref. [20] to represent complex motion patterns by linear combinations of prototype sequences and used for movement analysis and synthesis. A method for recognizing skill-level was presented in Ref. [45] to determine the ability of skiers by ranking properties such as synchronous and smooth motions.

In computer animation, a Fourier-based approach with basic and additional factors (walk; brisk) was employed in

Ref. [39] to generate human motion with different emotional properties (e.g. a happy walk). An HMM with entropy minimization was used by Ref. [7] to generate different state-based animation styles. An N-mode factorization of motion-capture data for extracting person-specific motion signatures was described in Ref. [42] to produce animations of people performing novel actions based on examples of other activities. The approach was also used to model faces under several changes, including illumination, expression, and pose [41]. A movement exaggeration model using measurements of the observability and predictability of joint-angle trajectories was presented in Ref. [15] to warp motions at one effort into increasing efforts using only selected trajectories. In Ref. [10], the EMOTE character animation system used the Effort and Shape components of Laban Movement Analysis to describe a parameterization for generating natural synthetic gestures with different expressive qualities.

A three-mode analysis of human movements over various performance efforts for biomechanic evaluation was presented in Ref. [29]. Arm segment velocities of 12 athletes throwing three different weighted balls were examined using a three-mode principal components factorization. The components themselves were manually inspected in an attempt to determine values signifying horizontal/vertical velocities, proximal/distal velocities, various throwing phases, and different skill levels of the throwers.

In this paper we present a different approach to modeling and recognizing style, where we use a three-mode principal components factorization of the data in which we embed trainable feature weights to bias the style estimation to a given matching criteria. The three-mode decomposition provides an explicit separation of the three natural modes into a low-dimensional set of components from which we can easily incorporate tunable weights on the motion trajectories. The approach can also capture meaningful temporal style variations in a computationally efficient manner, and does not require large training sets.

3. Three-mode principal components

Actions can be described as the changing body pose (mode-1) over time (mode-2). When considering stylistic actions, we therefore have a third action mode corresponding to the action style (mode-3). The data for multiple stylistic performances of a particular action can be naturally organized into a 3D cube Z (Fig. 1(a)), with the rows in each frontal plane Z_k comprised of the trajectories (segmented and normalized to a fixed length/size) for a particular style index k. The data for each variation k could alternatively be rasterized into a column vector and placed into an ordinary two-mode matrix (each column a motion example), but this simply ignores the underlying three-mode nature of the data (pose, time, style).

Many times it is preferable to reduce the dimensionality of large data sets for ease of analysis (or recognition) by describing the data as linear combinations of a smaller number of latent, or hidden, prototypes. Principal components analysis (PCA) and singular value decomposition (SVD) are standard methods for achieving this data reduction, and have been successfully applied to several *two-mode* (matrix) problems in computer vision (e.g. Refs. [5,6,26,28,38]). Three-mode factorization [37] is an extension of the traditional two-mode PCA/SVD in that it produces three orthonormal basis sets for a given data set represented as a 3D cube rather than a 2D matrix. As we will show, a three-mode action decomposition offers an efficient low-dimensional framework suitable to incorporating tunable weights on trajectories to drive the style estimation process to a context-based matching criteria.

3.1. Three-mode factorization

Before factorization of the data, we first *ij*-center the motions in Z by mean-subtraction of the trajectories along the style dimension. A three-mode factorization decomposes Z into three orthonormal matrices P, T, and S that span the column (pose), row (time), and slice (style)

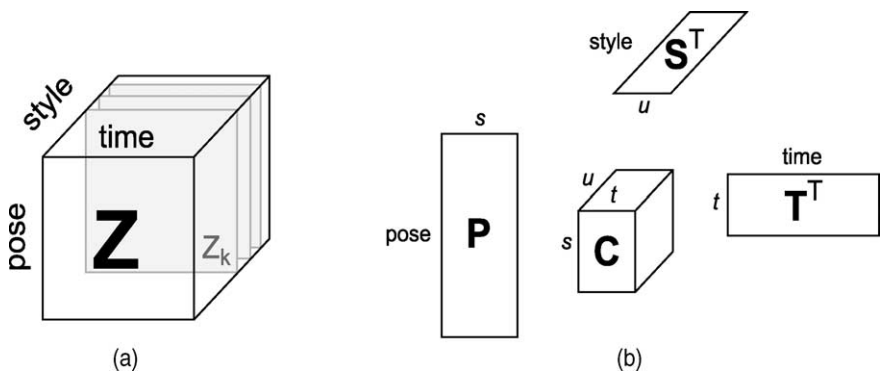


Fig. 1. (a) Three-mode arrangement of stylistic motion data. (b) Three-mode factorization of the data.

dimensions of the cube Z (Fig. 1(b)). The core C is a cube (much smaller in size than Z) and represents the complex relationships of the components in P, T, and S for reconstructing Z.

The three basis sets can be solved using three different 2D matrix flattening arrangements of Z

$$\text{Pose}: \ P = \text{colSpace}([Z_1|Z_2|\cdots|Z_n]) \tag{1}$$

$$\text{Time}: \ T = \text{colSpace}([|Z_1^{\mathrm{T}}|Z_2^{\mathrm{T}}|\cdots|Z_n^{\mathrm{T}}]) \tag{2}$$

$$\text{Style}: \ S = \text{rowSpace}([\vec{Z}_1|\vec{Z}_2|\cdots|\vec{Z}_n]) \tag{3}$$

where Z_k^{T} is the transpose of Z_k, and $\vec{Z}_k$ is the rasterized column vector of matrix Z_k (concatenation of motion trajectories for style k into a single column vector). The desired column and row spaces in Eqs. (1)–(3) can be found using standard two-mode SVD of the flattened matrix data. The resulting pose basis P is able to represent any body pose at any particular time and style (any column in Z). The time basis T represents a temporal trajectory of any feature (e.g. joint) position at any style (any row in Z). Lastly, the style basis S represents the style changes for any feature position at any particular time (any slice line in Z). Note that no two of the three basis sets can be produced within a single two-mode (matrix) SVD factorization. Typically, each mode needs only to retain its first few components (meeting some modal variance criteria) to capture most of the fit to Z when they are recombined.

The three-mode factorization of Z can be concisely written as $Z = PC(S^{\mathrm{T}} \otimes T^{\mathrm{T}})$, or in flattened matrix form as

$$[Z_1|Z_2|\cdots|Z_n] = P[C_1|C_2|\cdots|C_u](S^{\mathrm{T}} \otimes T^{\mathrm{T}}) \tag{4}$$

where $\otimes$ is the Kronecker product [25]. The core C (flattened) can be solved by re-arranging Eq. (4) as

$$C = [C_1|C_2|\cdots|C_u] = P^{\mathrm{T}}[Z_1|Z_2|\cdots|Z_n](S^{\mathrm{T}} \otimes T^{\mathrm{T}})^{\mathrm{T}} \tag{5}$$

where C need not be diagonal, as is required in two-mode PCA/SVD. Related methods for solving this three-mode factorization can be found in Refs. [24,41].

3.2. Comparison between two-mode and three-mode PCA/SVD

The three-mode factorization can be used to directly compute the traditional two-mode eigenvalues and projection operation if desired. To show a correspondence between three-mode and two-mode PCA/SVD for recognition, it is sufficient to derive the two-mode projection operation from the three-mode factorization.

It can be shown [25] that the squared summation of the rth frontal-plane elements of the three-mode core C (of cube Z) is equal to the rth squared singular values σ_r^2 (or equivalently the eigenvalues λ_r) of the rasterized flattened matrix $\vec{Z}$ (each column in $\vec{Z}$ is a rasterized motion vector $\vec{Z}_k$)

$$\lambda_r = \sigma_r^2 = \sum_i \sum_j C_{i,j,r}^2 \tag{6}$$

The SVD factorization of the two-mode matrix is $\vec{Z} = U\Sigma V^{\mathrm{T}}$, and the projection of the rasterized training examples onto the column space is $U^{\mathrm{T}}\vec{Z} = \sum V^{\mathrm{T}}$. Referring to Eq. (3), the style basis calculated for the cube Z is equivalent to the row basis V of the rasterized data in matrix $\vec{Z}$. Therefore, we can compute the two-mode projection coefficients ΣV^{T} from the singular values (eigenvalues) derived from the three-mode core (Eq. (6)) and the two-mode row basis $V = S$.

The advantage to a three-mode factorization over a two-mode decomposition is that we have a tri-modal separation of the data space from which we can easily weight certain dimensional variables (e.g. trajectories) with more influence when estimating the action style. The three-mode factorization also makes available useful basis sets for the feature trajectories and poses, which are not available from a single two-mode factorization of the rasterized data.

3.3. Three-mode estimation of style parameters

Any frontal plane Z_k (action at a particular style k) can be reconstructed in connection with Eq. (4) as

$$Z_k = P\left(\sum_{r=1}^{u} S_{kr} C_r\right) T^{\mathrm{T}} \tag{7}$$

where each example k has a corresponding set of S_{kr} component loadings from the style mode S. To determine the style for a new action within this framework, we therefore need only to estimate its corresponding style parameters in S. For this, a minimization of the reconstruction error can be used.

The three-mode reconstruction of each data element Z_{ijk} of Z can be written from Eq. (7) as a summation of particular three-mode basis elements, where the style loadings can be isolated from the remaining factored terms as

$$Z_{ijk} = \sum_{p=1}^{s} \sum_{q=1}^{t} \sum_{r=1}^{u} P_{ip} T_{jq} S_{kr} C_{pqr} \tag{8}$$

$$Z_{ijk} = \sum_{r=1}^{u} S_{kr} \left(\sum_{p=1}^{s} \sum_{q=1}^{t} P_{ip} T_{jq} C_{pqr} \right) \tag{9}$$

$$Z_{ijk} = \sum_{r=1}^{u} S_{kr} \alpha_{ijr} \tag{10}$$

Recall that indices i, j, and k correspond to elements in the respective pose, time, and style dimensions. If we have a nearly diagonal core (with $c_{pqr} \approx 0$ when $p \neq q$), we can further reduce the computations with $\alpha_{ijr} = \sum_{p=1}^{\min(s,t)} P_{ip} T_{jp} C_{ppr}$.

The style values for a new action $\hat{z}$ (after mean-subtraction with the model) can be estimated by

minimizing the least-squares reconstruction error for each element $\hat{z}_{ij}$ of $\hat{z}$

$$\mathscr{F} = \sum_i \sum_j \left(\hat{z}_{ij} - \sum_{r=1}^{u} \hat{s}_r \alpha_{ijr} \right)^2 \tag{11}$$

Setting the following partial derivatives to zero

$$\frac{\partial \mathscr{F}}{\partial \hat{s}_r} = -2 \sum_i \sum_j \left(\hat{z}_{ij} - \sum_{r=1}^{u} \hat{s}_r \alpha_{ijr} \right) \alpha_{ijr} = 0 \tag{12}$$

and re-arranging the terms in matrix format, we produce

$$\begin{bmatrix} \sum\sum \alpha_{ij1}^2 & \sum\sum \alpha_{ij2}\alpha_{ij1} & \cdots & \sum\sum \alpha_{iju}\alpha_{ij1} \\ \sum\sum \alpha_{ij1}\alpha_{ij2} & \sum\sum \alpha_{ij2}^2 & \cdots & \sum\sum \alpha_{iju}\alpha_{ij2} \\ & \vdots & & \\ \sum\sum \alpha_{ij1}\alpha_{iju} & \sum\sum \alpha_{ij2}\alpha_{iju} & \cdots & \sum\sum \alpha_{iju}^2 \end{bmatrix} \begin{bmatrix} \hat{s}_1 \\ \hat{s}_2 \\ \vdots \\ \hat{s}_u \end{bmatrix}$$

$$= \begin{bmatrix} \sum\sum z_{ij}\alpha_{ij1} \\ \sum\sum z_{ij}\alpha_{ij2} \\ \vdots \\ \sum\sum z_{ij}\alpha_{iju} \end{bmatrix} \tag{13}$$

The $\hat{s}_r$ style loadings can then be estimated using standard matrix inversion, where the α values and the $u \times u$ matrix can be pre-computed.

4. Estimation of style parameters using expressive features

The least-squares style estimation in Eq. (13) places equal prior emphasis on each motion trajectory in an action and minimizes based on those trajectories with the largest magnitude differences. Perhaps only select trajectories carry the salient information to distinguish the styles, while the differences in other trajectories may not correspond to the style changes. Furthermore, these salient trajectories may have smaller magnitude differences as compared to the other trajectories. What is needed is a method to place more emphasis on the important style-related trajectories. We refer to those key motion trajectories that reliably carry the impression of the styles as *expressive* trajectories.[1] To accommodate the notion of expressive features in our three-mode framework, we need to bias the three-mode style estimation process to emphasize the most style-expressive trajectories.

As all trajectories may not equally discriminate the action style (i.e. all may not be expressive key features), we augment the previous error function of Eq. (11) with positive expressibility weights $\mathscr{E}_i$ on each of the feature-i trajectories

$$\mathscr{F} = \sum_i \mathscr{E}_i \sum_j (\hat{z}_{ij} - \sum_{r=1}^{u} \hat{s}_r \alpha_{ijr})^2 \tag{14}$$

The weighted-least-squares estimation of the style is then

$$\begin{bmatrix} \sum_i \mathscr{E}_i \sum_j \alpha_{ij1}^2 & \cdots & \sum_i \mathscr{E}_i \sum_j \alpha_{iju}\alpha_{ij1} \\ & \vdots & \\ \sum_i \mathscr{E}_i \sum_j \alpha_{ij1}\alpha_{iju} & \cdots & \sum_i \mathscr{E}_i \sum_j \alpha_{iju}^2 \end{bmatrix} \begin{bmatrix} \hat{s}_1 \\ \vdots \\ \hat{s}_u \end{bmatrix}$$

$$= \begin{bmatrix} \sum_i \mathscr{E}_i \sum_j z_{ij}\alpha_{ij1} \\ \vdots \\ \sum_i \mathscr{E}_i \sum_j z_{ij}\alpha_{iju} \end{bmatrix} \tag{15}$$

For a given set of expressive weights $\mathscr{E}_i$, the above linear system can be used to solve for the target style values.

We desire to learn the expressive weight values needed to map the training data to some given set of style values. Under different contexts, we may desire different style associations for the motions. But even for a small number of training examples, there would be multiple style parameters $(s_1 \cdots s_u)$ for each example, which makes it difficult to gauge/assign these values to examples for a learning process. As there typically exists strong regularity within an action style, we expect that a single style parameter can be used to simply describe the range of variation. Therefore we consider a reduced model having only a single style parameter, which also simplifies the learning of the expressive weights and the computation of style.

4.1. Reduced three-mode model

To reduce the style mode in the previous three-mode model, and yet account for as much variance as possible, we focus on the two extreme style variations, Z_1 and Z_2 (e.g. slowest/fastest walking or lightest/heaviest carrying).

First we compute the ij-mean (along the style dimension) for the two extreme motions. This mean is subtracted from all of the training data to center the three-mode model on the two extremes. We then assign a (2×1) style basis $S = [-1, 1]^{\mathrm{T}}/\sqrt{2}$ to account only for the two extreme motions. For many continuous style variations (e.g. walking pace from *slow-to-fast*), we expect the variations that occur between these two extremes to exhibit smooth and predictable regularity [27]. Hence the remaining style variations in the training data between these two extremes are expected to be well-approximated with a style value in the range $\frac{-1}{\sqrt{2}} \leq \hat{s} \leq \frac{1}{\sqrt{2}}$.

The remaining P and T basis sets have no restriction on the number of training examples, therefore all of the training

[1] An alternate meaning of expressive features is given by Ref. [11] that refers to the principal eigenvectors.

data (mean-subtracted) can be used to compute P and T. The core is solved as in Eq. (5), but using only the two extreme motion examples (mean-subtracted) to match the dimensionality in S

$$C = P^{\mathrm{T}}[Z_1|Z_2](S^{\mathrm{T}} \otimes T^{\mathrm{T}})^{\mathrm{T}} \tag{16}$$

The core C is therefore reduced from a cube to a matrix.

The new reduced model with one (2×1) style component has only 1 style parameter, as shown in

$$Z_k = P(s_k C)T^{\mathrm{T}} = s_k PCT^{\mathrm{T}} \tag{17}$$

$$Z_{ijk} = s_k \alpha_{ij} \tag{18}$$

which are the new reduced versions of the full style equations of Eqs. (7) and (10). With only two extreme styles, $Z_1 = \frac{-1}{\sqrt{2}} PCT^{\mathrm{T}}$ and $Z_2 = \frac{1}{\sqrt{2}} PCT^{\mathrm{T}}$.

If we have the style ranges of multiple people in the training set, two 'prototype' extreme motions can be formed using the average of the lowest extreme Z_1 and the average of the highest extreme Z_2 from the set of multiple people. These two extreme prototypes can then be used in the above formulation. We will demonstrate this multi-person approach in the experiments.

4.2. Reduced style estimation

Recalling the expressive error function of Eq. (14), we can construct and solve a new error function using this reduced model (using Eq. (18)) to estimate the style value $\hat{s}$ for a movement $\hat{z}$ with

$$\hat{\mathscr{F}} = \sum_i \mathscr{E}_i \sum_j (\hat{z}_{ij} - \hat{s} \cdot \alpha_{ij})^2 \tag{19}$$

For minimizing $\hat{\mathscr{F}}$ to estimate the target style parameter $\hat{s}$, we again set the partial derivatives to zero and re-arrange to produce

$$\hat{s} = \frac{\sum_i \mathscr{E}_i \sum_j \hat{z}_{ij} \alpha_{ij}}{\sum_i \mathscr{E}_i \sum_j \alpha_{ij}^2} \tag{20}$$

$$\hat{s} = \sum_i \hat{\mathscr{E}}_i \sum_j \hat{z}_{ij} \alpha_{ij} \tag{21}$$

$$\hat{s} = \sum_i \hat{\mathscr{E}}_i \Delta_i \tag{22}$$

where $\Delta_i = \sum_j \hat{z}_{ij} \alpha_{ij}$. As the denominator $\sum_i \mathscr{E}_i \sum_j \alpha_{ij}^2$ in Eq. (20) is a constant for a given set of $\mathscr{E}_i$, we fold this term into the new expressive weights $\hat{\mathscr{E}}_i$.

Setting the new expressive weights to $\hat{\mathscr{E}}_i = (\sum_m \sum_n \times \alpha_{mn}^2)^{-1}$ in Eq. (22) yields the standard sum-of-squared-error (SSE) least-squares estimation of the style parameter. This is also equivalent to using standard two-mode PCA/SVD with rasterized data to recover the style parameter (the projection coefficient). However, with non-uniform expressive weights, to emphasize certain trajectories, the approach is capable of producing other desired style estimations. The approach is therefore not limited to the standard minimization of SSE solution (as is the case with standard PCA/SVD).

5. Learning expressive weights

The remaining task is to determine the appropriate values of the expressive weights $\hat{\mathscr{E}}_i$ within the reduced three-mode model to compute styles that correspond to some preferred values. We outline a method on how these weights can be efficiently learned from labeled training examples.

Our approach is based on minimizing an error function comparing the computed style values (using Eq. (22)) with the given training style values. With a set of $(K \geq 2)$ training motions (K different variations), we first construct the reduced three-mode basis. We then construct an error function of the K examples, comparing the computed style values $\hat{s}_k$ with the desired style values $\bar{s}_k$ pre-assigned to those examples

$$J = \sum_k (\bar{s}_k - \hat{s}_k)^2 \tag{23}$$

$$J = \sum_k (\bar{s}_k - \sum_i \hat{\mathscr{E}}_i \Delta_{ik})^2 \tag{24}$$

The expressive weights in Eq. (24) can be solved linearly, but this requires as many style variations (K) as there are expressive weights (feature trajectories). Instead, to accommodate smaller training sets, we employ a fast iterative gradient descent algorithm [9] of the form

$$\hat{\mathscr{E}}_i(n+1) = \hat{\mathscr{E}}_i(n) - \eta(n) \cdot \frac{\partial J}{\partial \hat{\mathscr{E}}_i} \tag{25}$$

with the gradients $\partial J / \partial \hat{\mathscr{E}}_i$ computed over the K training examples

$$\frac{\partial J}{\partial \hat{\mathscr{E}}_i} = -2 \sum_k \Delta_{ik} \left(\bar{s}_k - \sum_j \hat{\mathscr{E}}_j \Delta_{jk} \right) \tag{26}$$

The learning rate η is re-computed at each iteration (via interpolation of the error function [9]) to yield the best incremental update.

The general gradient descent algorithm determines a local minimum for a multi-parameter error function by searching through the 'parameter space' to find the minimum error. The algorithm evaluates the error function with the current parameter values (in our case, the $\hat{\mathscr{E}}_i$) and then determines a change in those parameters that decreases the error. Updating the parameters by a small amount in the opposite direction of the positive error gradient reduces the error function. This updating process is repeated until the process converges or reaches a maximum number of iterations.

In our approach, the expressive weights are initialized to $\hat{\mathscr{E}}_i = (\sum_m \sum_n \alpha_{mn}^2)^{-1}$ (the default SSE formulation) and confined to be positive. Following termination of Eq. (25),

the style variation for a motion can then be estimated using the learned expressive weights in Eq. (22).

5.1. Pre-assignment of style to training data

A numeric assignment of the style parameters $\bar{s}_k$ to the K variations is required before the training phase. In the case of smoothly changing (continuous) variations, such as walking at different speeds (from *slow-to-fast*), the two extreme variations are assigned the default values of $\bar{s}_1 = \frac{-1}{\sqrt{2}}$ and $\bar{s}_2 = \frac{1}{\sqrt{2}}$. The remaining $(K-2)$ variations are then assigned values in the range $\bar{s}_1 \leq \bar{s}_k \leq \bar{s}_2$, according to their relationship to the extreme variations. For a binary (discrete) style change with no in-between variations and multiple training examples, such as labeling walkers as *male* or *female*, all K training actions are assigned to either $\bar{s}_1 = \frac{-1}{\sqrt{2}}$ or $\bar{s}_2 = \frac{1}{\sqrt{2}}$, depending on their known state. We conform style variations to one of these two scenarios (continuous, discrete). We now describe two methods of determining the actual style values to be used in training the expressive weights.

5.1.1. Physical parameterization

A measured physical quantity of the action or actor can be used as the target style variation. For example, one could assign the style values for a lifting action based on the amount of weight of the object being lifted. For a person lifting five progressively heavier objects in turn (e.g. 10, 20, 30, 40, 50 lbs), the variation mapping of $\bar{s}_k$ will be evenly distributed as $\frac{-1}{\sqrt{2}}, \frac{-1}{2\sqrt{2}}, 0, \frac{1}{2\sqrt{2}}, \frac{1}{\sqrt{2}}$. The trained model would then be able to estimate the actual amount of weight of new objects being lifted by the person.

We note, however, that such *dynamic* physical style assignments, as mapping to the actual weight being lifted, are typically only valid for the one particular individual used to train the system. The actual body movement changes due to the different weights will obviously be relative to the strength and build of the person. A 10 lb object may seem fairly light for an adult to lift, but it may be extremely difficult for a young child. Hence, if the framework employs these dynamic physical mappings to style, the resulting model will generally not be valid across multiple people. However, non-dynamic physical properties, such as labeling a person as a *male* or *female*, can be used within a general multi-person model (as will be demonstrated). For any physical parameterization to be successful, there must exist a high correlation between the physical style labels and the observable movement changes.

From these physical style values, the expressive weights can be adapted (using Eq. (25)) such that the model produces these desired style variations using Eq. (22). New test motions outside of the trained variation range can still be identified via extrapolation of the physical-value mappings.

5.1.2. Perceptual parameterization

An alternate method of assigning style variations to training motions is to have multiple people observe the target actions to determine a 'perceptual' style rating.

In the continuous style variation case, the two extreme motions can be linearly interpolated (and/or extrapolated) to create synthetic in-between style variations. These synthetic motions therefore have known style values linearly sampled between $\frac{-1}{\sqrt{2}}$ and $\frac{1}{\sqrt{2}}$. A perceptual matching task involving the K real training variations and the synthetic examples can then be performed. Here the observer is asked to select which of the synthetic motions *appears* most similar to each real motion. With multiple observers matching the motions, we can assign the numeric value of the average (or the most common) style value of the synthetic motions perceptually matched to each real motion.

If instead we have a two-class discrete case (e.g. gender), the perceptual task is slightly different. Instead of matching between real and synthetic motions, the observer is presented with each motion (individually) and asked to label it as class-1 ($\frac{-1}{\sqrt{2}}$) or class-2 ($\frac{1}{\sqrt{2}}$). For example, the observer may be asked if the displayed motion looks more like a *male* or a *female*. This result may be quite different from the true physical label (as we will demonstrate).

With a perceptual style value assigned to each of the motions (either by synthetic matching or classification), we then proceed to adapt the expressive weights to bias the reduced three-mode model to produce the assigned style values for the motions. Given the set of K training motions and their assigned perceptual styles $\bar{s}_k$, we slightly alter the previous matching error function J to reflect the strength of the perceptual matching results with

$$J_{\mathrm{p}} = \sum_k \omega(\sigma_k) \cdot (\bar{s}_k - \sum_i \hat{\mathscr{E}}_i \Delta_{ik})^2 \tag{27}$$

where $\omega(\sigma_k)$ is a function of the perceptual matching consistency for each motion k. This consistency function is used to give more influence during minimization to those examples having more reliable matches across the observers (e.g. having a smaller standard deviation of the synthetic choices). The corresponding perceptual gradient is then

$$\frac{\partial J_{\mathrm{p}}}{\partial \hat{\mathscr{E}}_i} = -2 \sum_k \omega(\sigma_k) \Delta_{ik} (\bar{s}_k - \sum_j \hat{\mathscr{E}}_j \Delta_{jk}) \tag{28}$$

which is used as before in the gradient descent procedure (Eq. (25)) to determine the appropriate expressive weights for the data.

6. Experiments

We present experiments to demonstrate the potential of the expressive three-mode framework for modeling and recognizing different human action style variations. We test the generality of the expressive model by examining

different matching criteria (physical, perceptual) and showing the capability of the model to produce useful expressive feature weights for estimating the desired styles. Specifically, we test the approach with several walking variations that are due to (1) light-to-heavy carrying load (physical, perceptual), (2) different gender (physical), and (3) increased pace (perceptual). Both single-person and multi-person training sets are evaluated.

As the focus of this work is a representation for movement/action recognition, and not on the initial body tracking mechanism, we tested the approach with motion trajectories of people collected with a motion-capture system. The carrying load and walking pace examples were collected with an in-house 14-camera Vicon-8 system, and the gender data was provided from the collection used in Ref. [36]. Trajectories of the x, y image coordinates for the joints of the hands, elbows, shoulders, hips, knees, feet, and head were selected (or generated automatically with software) from a collection of markers strategically placed on the body (Fig. 2). In future work, we will attempt to incorporate a video-based human body tracker, such as one of the approaches presented in Refs. [4,8,18,30,33], to extract these joint positions automatically. To help facilitate a smoother transition from motion-capture data to video tracking output, we selected only those joint positions that are typically used in the articulated figure models of the cited trackers. Also, we used the 2D position (image) trajectories, rather than relying completely on full 3D position or joint-angles.

The motion-capture trajectories were initially lowpass filtered using a 5th-order, zero-phase forward-and-reverse Butterworth filter with cut-off at 6 Hz. The translation component for each action was removed from each trajectory by subtracting the mean root translation during the action. Two action cycles for each carry and pace example were automatically extracted using trajectory curvature peaks (of the leg motion), averaged, and time-normalized to a fixed duration using spline interpolation. A total of 15 joint positions (30 $x-y$ trajectories) were used. For the gender data provided to us, one cycle was extracted (due to non-regular translational shifts that occurred over multiple cycles) and 13 (of the 15) joint locations were used. All data were saved as 2D (orthographic) joint-position trajectories at selected views.

6.1. Physical style variation experiments

For the physical parameterization experiments, a physical property directly associated with the changing style was recorded for the actions. For each example, a relative style metric between the two extremes $(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ was assigned. These values were then used to adapt the expressive weights (using Eq. (25)) so the reduced three-mode model could produce the desired physical style values for the given motions (using Eq. (22)).

6.1.1. Carrying load

Our first experiment was to model the style changes that occur for a single person walking while carrying packages of different (increasing) weight. The person walked on a treadmill at a constant speed (1.4 mph) while carrying (with one hand) nine differently loaded bags in succession

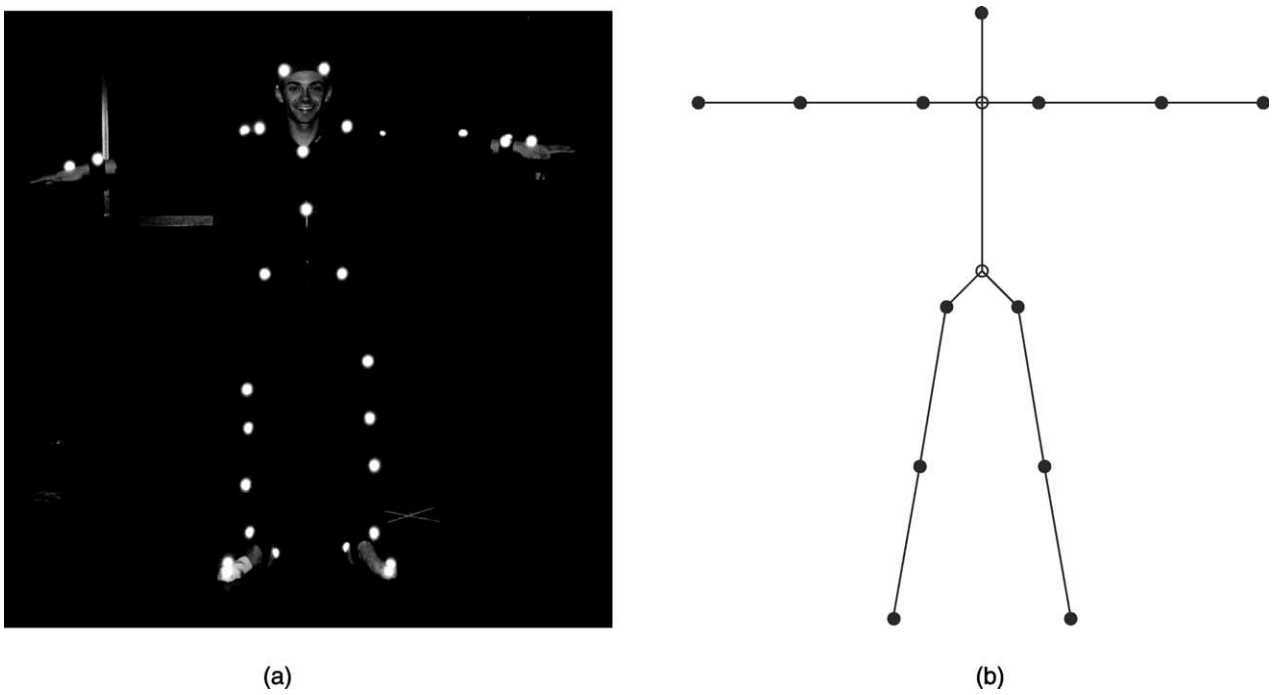


Fig. 2. Motion-capture. (a) Person in T-pose configuration suited with reflective markers. (b) Limited body-skeleton derived from the motion-capture data.

(0–40 lbs). The joint-position trajectories for one complete walk cycle at each carrying effort were rendered at a 45° view angle to best show the body changes, and the trajectories were time-normalized to 42 frames (duration chosen from the slowest cycle-time of the carry examples).

The lightest and heaviest carry motions (0, 40 lbs) were then used as the extreme motions needed to create the reduced three-mode basis. With an 85% modal variance criterion using the full training set, P and T were of dimension (30×10) and (42×5), having 10 and 5 components, respectively. By default in the reduced model, the style mode had a single (2×1) component $S=[-11]^{\mathrm{T}}/\sqrt{2}$. The core C had the size (10×5). The size of the full training set was $9\times(30\times 42)$. The reduced three-mode modal captured 99.2% of the overall data variance in the two carry extremes.

To learn the expressive weights needed to capture the actual physical carrying load of the person, the style values for the nine carry examples were evenly distributed between the extreme values of $\frac{-1}{\sqrt{2}}$ and $\frac{1}{\sqrt{2}}$. Running the gradient descent learning algorithm of Eq. (25) (for 1500 iterations, with SSE initial values of $\hat{\mathscr{E}}_i=(\sum_m\sum_n\alpha_{mn}^2)^{-1}=7.72\times 10^{-4}$) produced the 30 expressive feature weights shown in Fig. 3. Approximately one-third of the weights were zero. The ordering of the 30 weights correspond to x–y joint positions {ROOT:(1–2), LEFT-LEG/HIP:(3–8), RIGHT-LEG/HIP:(9–14), HEAD-NECK:(15–18), LEFT-ARM:(19–24), RIGHT-ARM:(25–30)}.

The target style values and the non-expressive SSE result (using $\hat{\mathscr{E}}_i=7.72\times 10^{-4}$) are shown in Fig. 4(a). The default SSE estimation was quite different from the desired styles (average error = 0.1424). In Fig. 4(b), we present the resulting expressive estimation after adapting the expressive weights to the training data. The model conformed quite well to the training data, producing styles similar to the target values (average error = 0.0198).

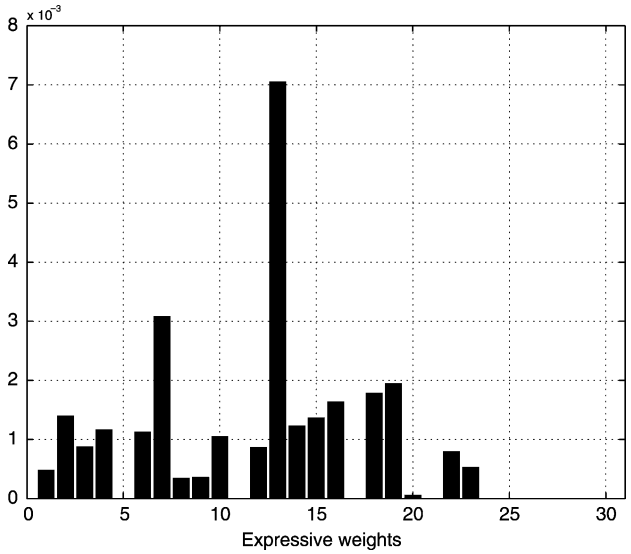


Fig. 3. Expressive weights for the physical carrying load experiment.

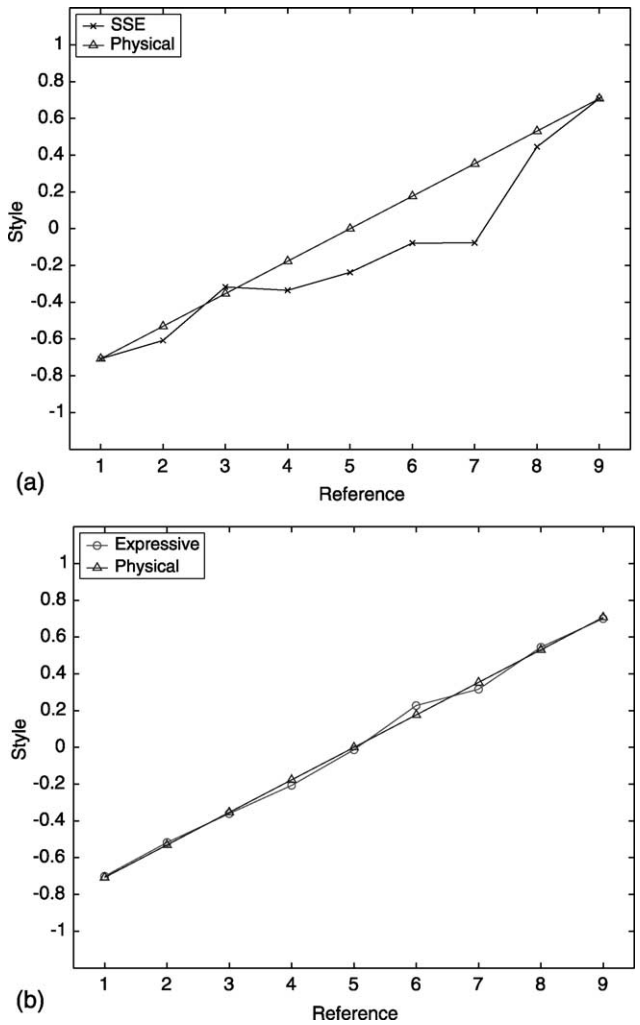


Fig. 4. Physical carrying load experiment results. (a) Physical and SSE styles. (b) Physical and expressive styles.

6.1.2. Gender of walker

In our second physical parameterization experiment, we examined the style differences of multiple male and female walkers (two-state discrete style variation). The data set for this experiment included 20 males and 20 females, each walking at a comfortable pace on a treadmill. The motion trajectories of one cycle at the frontal view (determined to be the most discriminating view [36]) were automatically extracted and time-normalized to 50 frames. To build the reduced three-mode model, the prototype method (as described in Section 4.1) was used in which the male motions were averaged together to form a 'male prototype' and the female motions were averaged into a 'female prototype' (each person was first height-normalized). All of the data were used to compute the P and T basis sets, and the prototypes were used to solve for the core C. With an 85% modal variance criterion for the full training set, P and T were of dimension (26×16) and (50×5),

respectively (C was therefore of size (16×5)). The size of the full training set was $40 \times (26 \times 50)$. The reduced three-mode model captured 98.57% of the overall data variance in the two gender prototypes.

To learn the expressive weights from the training data, all of the females were assigned to a style of $\frac{-1}{\sqrt{2}}$ and all of the males were assigned a value of $\frac{1}{\sqrt{2}}$. The gradient descent learning algorithm (1500 iterations, with SSE initial values of $\hat{\mathscr{E}}_i = 1.2119$) produced the 26 expressive feature weights shown in Fig. 5. The ordering of the 26 weights correspond to {HEAD:(1–2), RIGHT-ARM:(3–8), LEFT-ARM:(9–14), RIGHT/LEFT-HIP:(15–18), RIGHT-LEG:(19–22), and LEFT-LEG:(23–26)}.

The target two-class style values and the non-expressive SSE result for the 40 people are shown in Fig. 6(a). The SSE estimation of the gender was quite different from the true label (average error $= 0.9423$). In Fig. 6(b), we present the expressive estimation for the gender data. The expressive model adapted to the data to produce styles much more similar to the true genders (average error $= 0.2990$).

Thresholding the estimated style values at zero (midway between the male–female assignments) produced a 7.5% classification error of male and female walkers for the expressive model. Thresholding the SSE estimation produced a much larger 27.5% error. To examine the generalization capability of the model (and to avoid overfitting), we also performed a leave-one-out cross-validation of the data. We computed 40 different models (each using 39 examples by leaving one example out of the training set) and averaged the resulting expressive weights from the 40 models. The average cross-validation training error with the expressive model was 6.54% (SSE average training error was 26.92%). For the testing samples, the average classification error for the expressive model was 22.50% (SSE average testing error was 30.00%). Using the averaged set of expressive weights from cross-validation on the complete set of training data produced the same classification as the previous result (7.5% error).

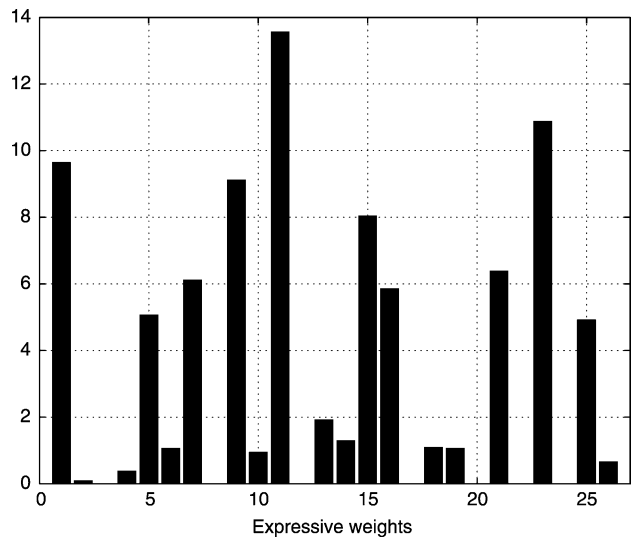


Fig. 5. Expressive weights for the physical gender experiment.

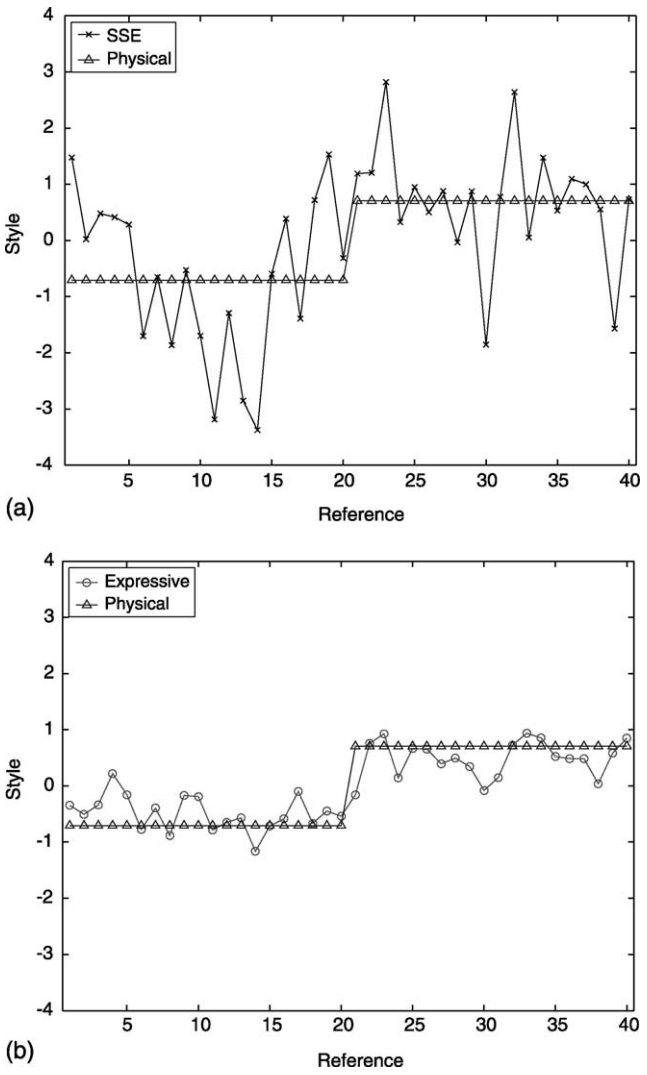


Fig. 6. Physical gender experiment results. (a) Physical and SSE styles. (b) Physical and expressive styles.

6.2. Perceptual style variation experiments

We next assigned style values to the training data using perceptual matching labels rather than employing an actual physical property. In these experiments, observers were asked to match examples of carrying load (carrying light-to-heavy objects) and walking pace (walking leisurely-to-quickly). A computer program was developed to facilitate the perceptual matching task for assigning style values to the motion data.

In the computer display (Fig. 7), two motions are presented side-by-side to the observer, with the motions synchronized and looped. The motion on the left of the display, referred to as a 'reference' motion, is one of the examples we wish to assign a perceptual style label. The motion displayed on the right is a synthetic motion linearly

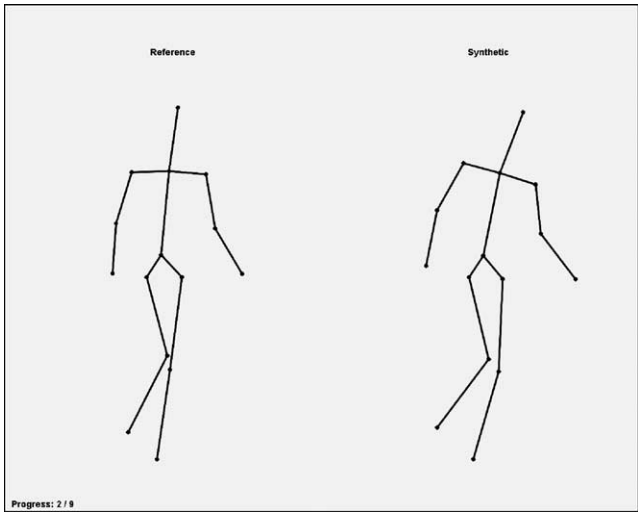


Fig. 7. Screen-shot of perceptual matching program. The user manipulates the motion on the right to make it look as similar as possible to the motion on the left.

interpolated/extrapolated from the two extreme reference style variations. As the synthetics are linearly derived from the two extremes, they have known style values. The task for the observer is to seek through the various synthetic motions to find the best match to the displayed reference motion. The left and right arrow keys on the keyboard enable the user to quickly and easily seek through all possible synthetics (generated online). Once the user has made a selection of which synthetic motion most closely resembles the reference action, the person confirms this choice by pressing the spacebar (changing the color of the synthetic motion to blue) and hits the enter key to load the next random reference and synthetic motions.

The program records the style value of the synthetic motion selected (determined from the interpolation/extrapolation amount) for each reference action. The display was generated at a 1280×1024 resolution on a $19''$ monitor using C++ and OpenGL with anti-aliasing. The user was seated a typical working distance to the computer.

Since the reference motions will not have a perfect correspondence with the synthetic motions (except at the two extremes), the person will have to match the motions based on some subjective measure of similarity. Our goal is to tune the expressive weights to reflect this unknown perceptual similarity process.

6.2.1. Carrying Load

For the carrying load perceptual matching task, the same data used in the previous physical carry parameterization experiment are judged by multiple observers for similarity with their synthetic counterparts. Fig. 8 shows a sample of the synthetic versions generated between the two carry extremes. The interpolation/extrapolation was produced with a sampling interval of $\frac{2}{14\sqrt{2}}$ (producing 13 interpolations between the two extremes). In this experiment, the carried object was not displayed so the observer would concentrate only on the changing body movements (rather than on the movement of the bag).

Ten people were given the task of perceptually matching the carrying movements to the set of synthetic motions to provide the needed training labels. From the resulting reference-synthetic perceptual choices of the 10 observers, the mean and standard deviation of the 10 selected synthetic style values for each reference motion were computed. The average correlation coefficient of the reference-synthetic values for pairwise observers was $r = 0.8$ (SD 0.1), suggesting that the observers were fairly consistent in perceiving a similar carrying effort for the movements.

We employed the reduced three-mode basis from the previous physical carrying load experiment and trained the model with the perceptually labeled examples. Using the perceptual means $\bar{s}_k$ and standard deviations σ_k for each carrying example, the gradient descent learning algorithm with the perceptual gradient function (Eq. (28)) was run (for 1500 iterations, with the initial SSE values of $\hat{\mathscr{E}}_i = 7.72 \times 10^{-4}$) to determine the appropriate expressive weights for the training data. We used the weighting factor $\omega(\sigma_k) = \exp(-\sigma_k^2/0.25)$ in Eq. (28) to give more emphasis to those examples with the most consistent perceptual matches (this weighting factor is also used in the following experiment). The resulting 30 expressive weights are shown in Fig. 9. As 19 out of 30 weights are zero, the focus on expressive features significantly reduced the amount of data needed to capture the perceived variations.

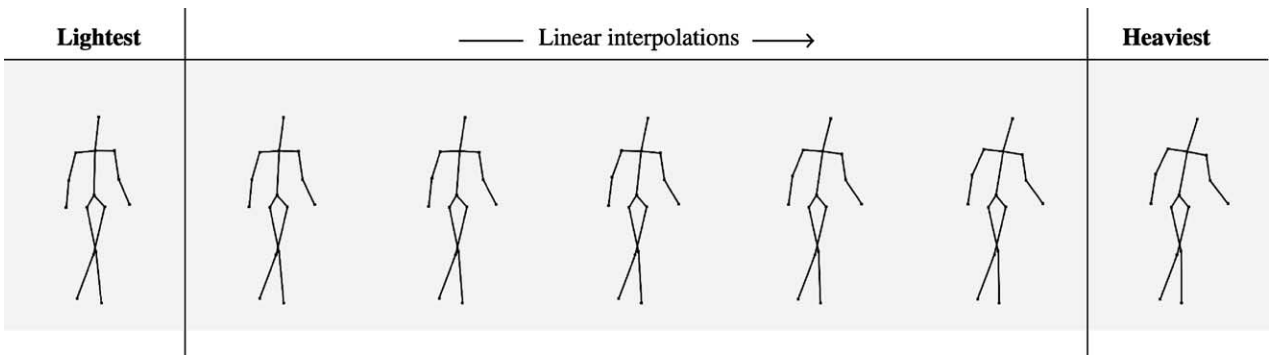


Fig. 8. Multiple efforts for carrying (object carried in the person's right hand is not shown). The real lightest and heaviest carry actions are used to generate the synthetic linear interpolations. One frame from each interpolated sequence is shown.

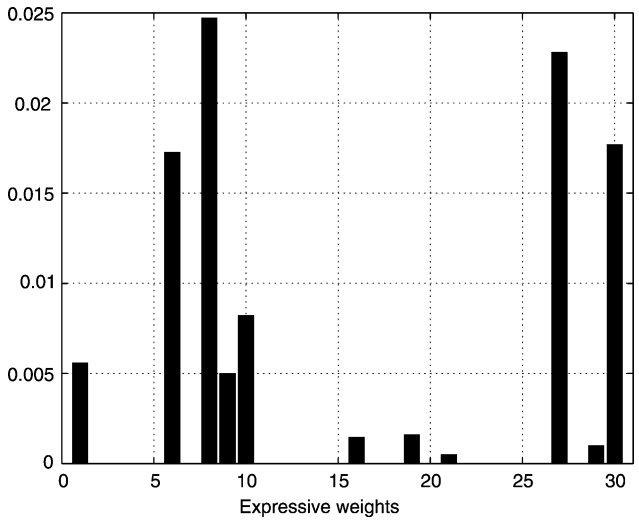


Fig. 9. Expressive weights for the perceptual carry experiment.

In Fig. 10(a), we show the perceptual style values for the reference motions (style means with one standard deviation) and the default SSE style estimation. Note that the perceptual styles appear sub-categorical to the observers (with no smooth trend as in the physical parameterization of carry load). The reference motions #8 and #9 (two heaviest carries) were perceived quite differently from the remaining lighter carry motions (#1–#7). In Fig. 10(b), we show the perceptual results along with the output of the expressive model trained with the perceptual style means and standard deviations. The results show that our model could adapt well to the perceptual ranking of the effort styles (in addition to the previous physical parameterization). The average errors for the SSE and expressive models were 0.1624 and 0.0442, respectively.

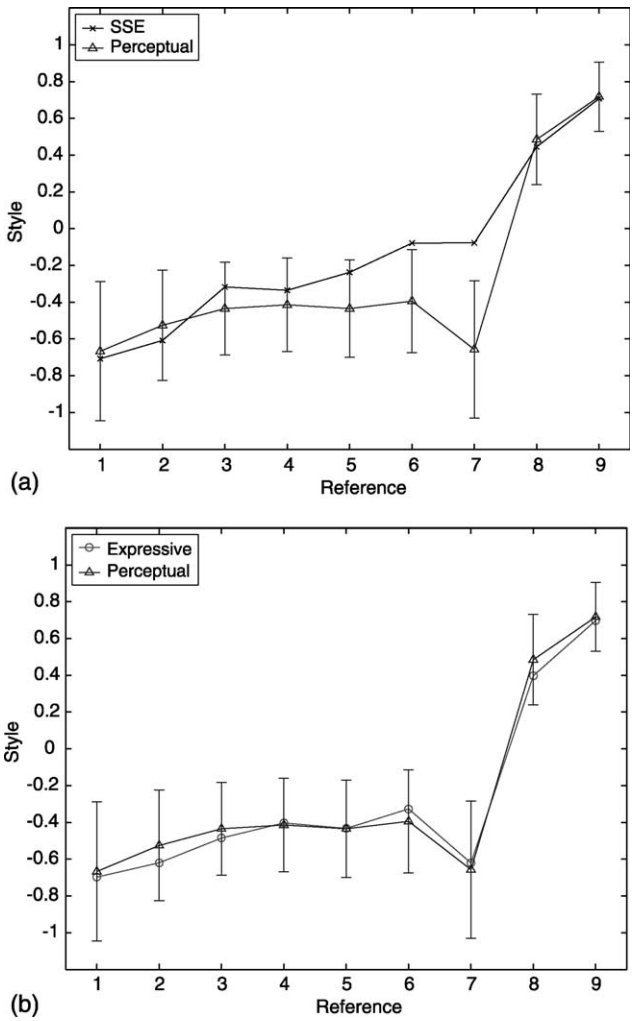


Fig. 10. Perceptual carry experiment results. (a) Physical and SSE styles. (b) Physical and expressive styles.

6.2.2. Walking pace

To further evaluate the framework using perceptual styles, we examined different walking paces of multiple people to address the dynamic changes that occur with increased walking speed. Rather than mapping to the actual physical walking speed, we instead focus on the observer's judgement of the pace, such as determining if the person is walking 'leisurely' or walking 'in a hurry'. Actual walking speed is not a global indicator for the walking pace of different people. Consider a child walking alongside an adult. Both are moving at the same ground speed, but the child's pace will be much higher to keep up with the adult. Thus, speed and cycle time do not generally correspond to the pace.

The training set was comprised of five different walking speeds for each of three people walking on a treadmill at speeds ranging between 2.0 mph and each person's natural walk-run transition. Two additional motions (one slower at 1.6 mph and one midway between the 2.0 mph and the walk-run transition) were collected for further testing. The motions were time-normalized (to 50 frames) to remove the influence of the actual cycle time and speed of the walkers.

To generate the synthetic motions for the perceptual test and to construct the reduced three-mode basis from multiple people, we use the prototype method outlined in Section 4.1. The minimum-pace prototype was formed by averaging[2] the three people walking at 2.0 mph, and the maximum-pace prototype was formed by averaging the motions just before their respective walk-run transitions. Example pace synthetics between the prototypes are shown in Fig. 11. As in the previous perceptual experiment, the motions were rendered in real-time at 30 Hz at a 45° camera view, but the figure heights were normalized (by the image height of the person) to accommodate for different person statures.

Ten observers (from the previous experiment) were given the task of perceptually matching the reference walking pace movements of the three people (15 total) to the set of synthetic walking motions generated from the two

[2] In this experiment, we averaged the motion-capture skeleton (limb lengths) and joint-angles rather than the joint positions.

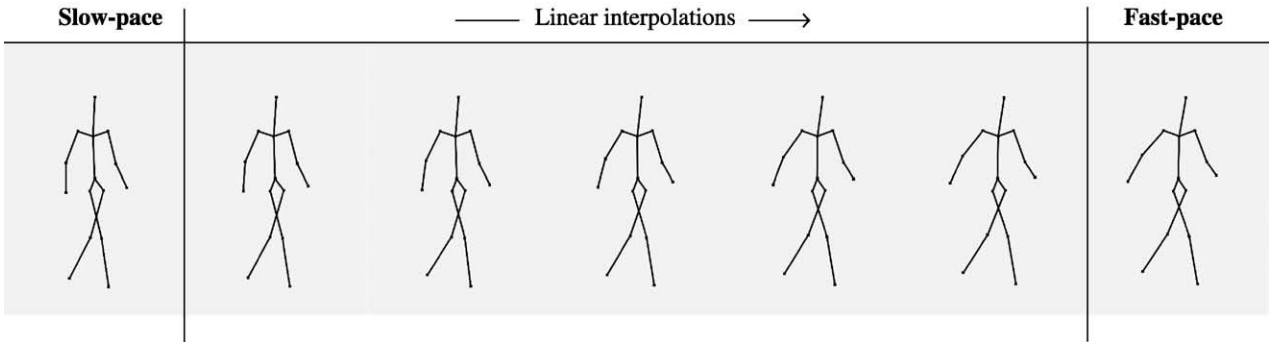


Fig. 11. Action styles for the mean-prototype slowest and fastest walking pace with linearly interpolated examples. One frame from each interpolated sequence is shown.

prototypes. The mean and standard deviation of the 10 synthetic pace values selected for each reference motion were computed. The average correlation coefficient of the reference-synthetic choices for pairwise observers was $r = 0.8$ (SD 0.1). The observers were also asked to match the remaining two motions from each walker to the synthetics.

With an 85% modal variance criterion using the 15 training examples (height-normalized), P and T were of size (30×10) and (50×4), respectively. The core C was of size (10×4). The size of the complete training set was $15 \times (30 \times 50)$. The combined basis sets captured 94.3% of the overall data variance in the walking pace prototypes. The 30 expressive weights produced by the perceptually-based gradient descent learning algorithm (1500 iterations, initialized with SSE values of $\hat{\mathscr{E}}_i = 2.2022$) are shown in Fig. 12.

We compared the perceptual pace matches between the reference and synthetic motions separately for each person in the training set. For Person-1, the perceptual and expressive results were very similar, but the SSE estimations were different (shifted) from the desired perceptual labels (Fig. 13(a)). The advantage of the adaptable expressive model is well illustrated. For Person-2, the SSE estimation method is still different, though now within the ± 1 SD range (Fig. 13(b)). Lastly, for Person-3, the SSE and expressive styles were very similar except at the mid-pace examples of #3 and #4 (Fig. 13(c)). The average error for the SSE and expressive methods for the 15 examples were 0.1950 and 0.0417, respectively.

We also tested the two additional walking motions collected from each person to evaluate the expressive style estimation method for new motions of the training people. In Fig. 14(a)–(c), the expressive values (learned only from the previous training examples) and SSE estimations closely matched the perceptual results for all three people, though the average expressive error (0.1004) was less than the average SSE error (0.1638).

6.3. Discussion of results

We presented several experiments examining the ability of the expressive three-mode framework to adapt to both physical and perceptual values assigned to the style variations of the general walking category. The successful adaptation of the model to the different matching criteria demonstrates the potential of the method to learn the necessary key features of the motions needed to produce the desired style output.

We examined both physical and perceptual variations for a single person carrying objects of different weight, showing how the physical and perceptual criteria are quite different (linear vs. sub-categorical), but that the expressive model was capable of tuning itself to either criteria. Thus an advantage of this model is that one set of expressive weights can be used to model the physical load and another set of weights can be used to recognize the observed heaviness, though only one three-mode basis is required. The different recognition contexts are reflected only in the expressive weights.

We also examined the approach to model walking variations due to gender. The results showed an impressive 7.5% classification error on a set of 40 walkers. As gender recognition has been an active research domain for several years, we feel our model has merit for further analysis in this area. Depending on the input representation selected for

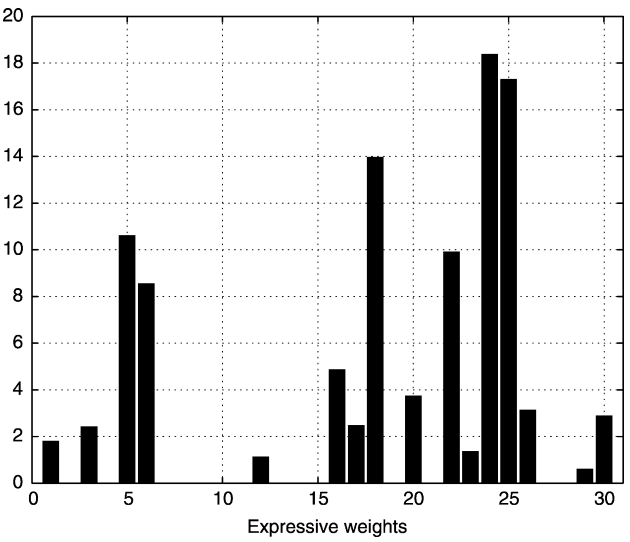
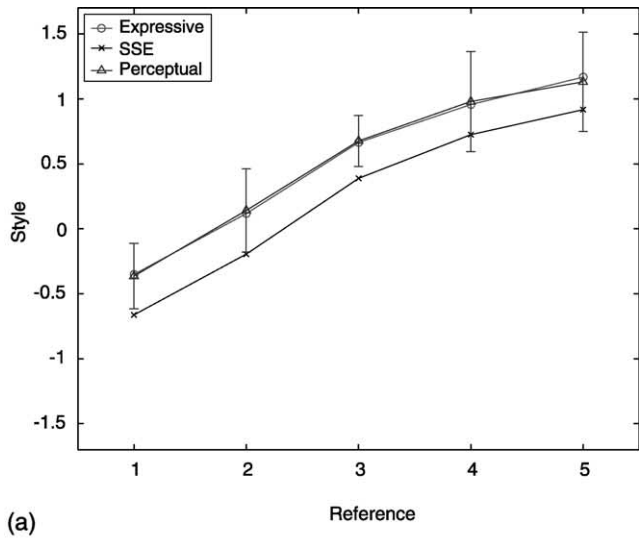
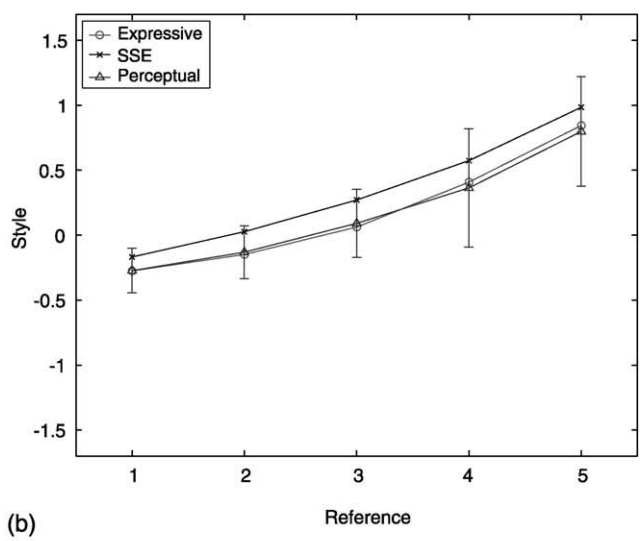


Fig. 12. Expressive weights for perceptual pace experiment.





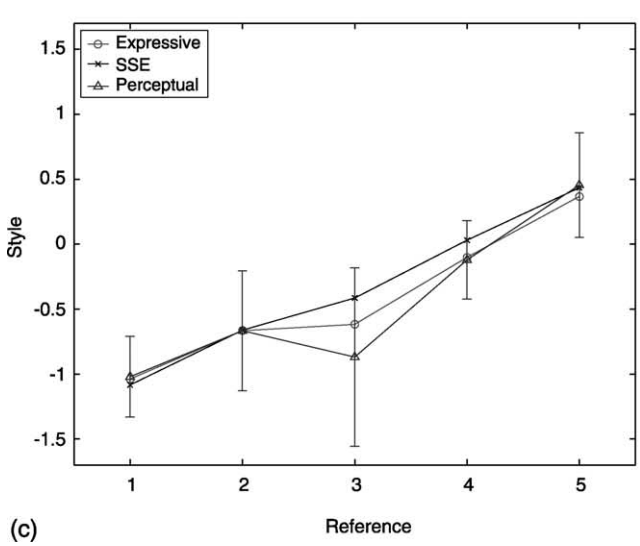
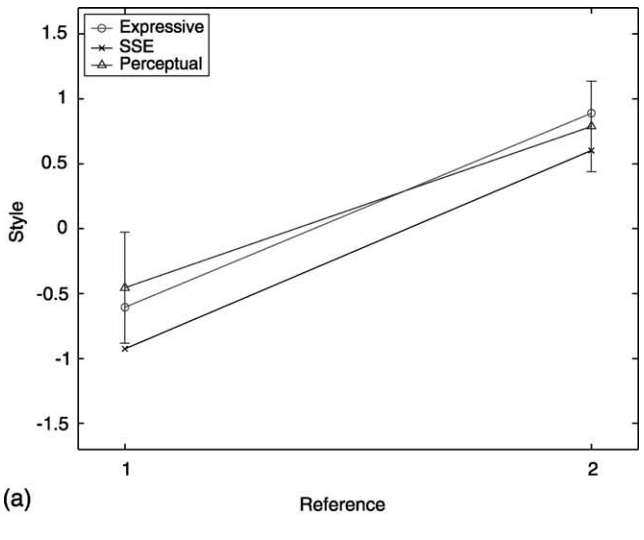
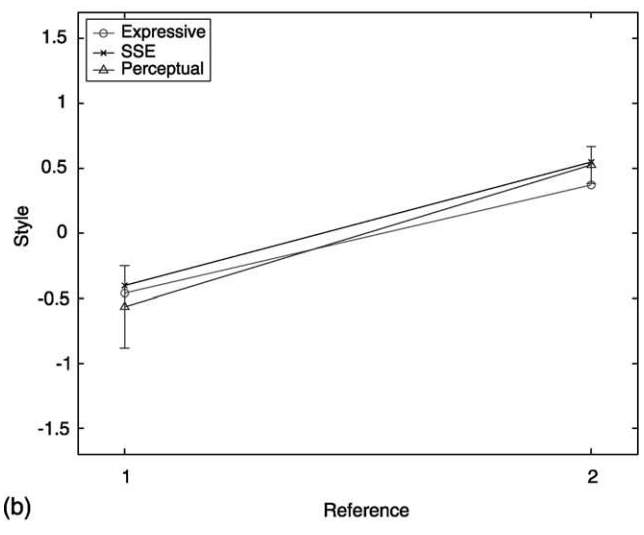


Fig. 13. Perceptual (mean ± 1 SD), SSE, and expressive pace estimation for the five walking motions of (a) Person-1, (b) Person-2, and (c) Person-3.





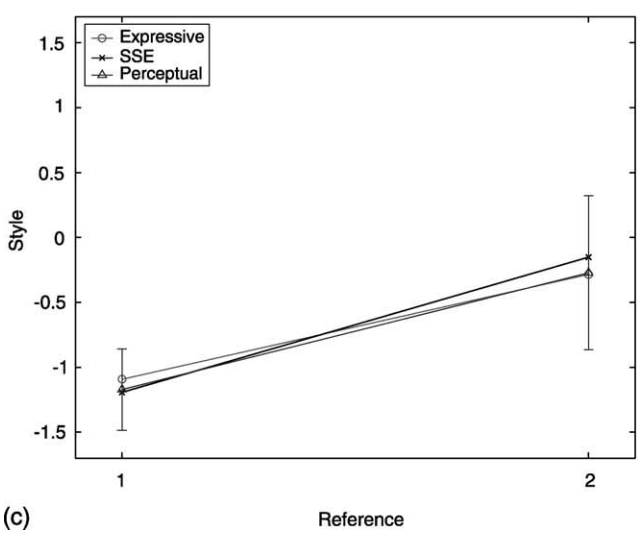


Fig. 14. Perceptual (mean ± 1 SD), SSE, and expressive pace estimation for two new (non-training) walking motions of (a) Person-1, (b) Person-2, and (c) Person-3.

the motion data, our model may provide insight to which features may be used by human observers during the gender classification task.

Lastly, we demonstrated the framework's ability to capture the general style changes associated with the walking pace of multiple walkers. The results again clearly illustrated the advantage of the expressive model over a standard SSE estimation.

7. Conclusion and future directions

We presented an approach for modeling and recognizing different action styles using an efficient three-mode principal components framework that gives more influence to key expressive trajectories learned from style-labeled training data.

The approach initially factors training examples with different style variations into a reduced three-mode principal components model to provide individual basis sets for the body poses, temporal trajectories, and styles. An advantage of this multi-modal basis set is that it offers a low-dimensional decomposition of the data suitable for incorporating expressive weights on trajectories to bias the model estimation of style to desired values. We presented a weighted-least-squares style estimation method in this three-mode sub-space with expressive weights to emphasize those trajectories most indicative of the style.

To learn the values of the expressive weights, we presented two types of style parameterization of the training data. Physical parameterization manually assigns style values that correspond to a known physical property of the motion, such as the actual carrying load or gender. Perceptual parameterization is accomplished by using a perceptual matching task to attain the observable correspondences of real motions to synthetic variations having known style values. The style-labeled training data are then used in a gradient descent learning algorithm to solve for the expressive weight values needed to align the model estimation of style to the assigned training values. Thus instead of matching a new motion to several exemplars for recognition, our low-order expressive model is used to directly compute a metric value of style for the motion.

The approach was examined with several walking style variations that were due to carrying load (physical, perceptual), gender (physical), and pace (perceptual). We showed that our model with expressive weights can be used to adapt to different style parameterizations, and therefore demonstrated more flexibility than a standard SSE-based style estimation.

In future work, we plan to incorporate a video-based body tracking algorithm to locate and track the same motion-captured body joints used in this research, and evaluate the approach with occlusion and tracking errors. We will additionally broaden the range of actions to include other non-periodic activities (e.g. throwing, lifting), which will in turn require a method of robust temporal segmentation of the actions for proper time normalization. We are currently examining curvature-based approaches such as in Refs. [2,32] to automatically detect action keyframes for the segmentation. We demonstrated the expressive three-mode approach with experiments using physical male/female gender labels, but we will additionally examine a perceptually-based gender assignment to the training data to compare results using the true gender with results using the perceived gender. We are also interested in the relation of our framework and its applicability to modeling the Effort factor of Laban Movement Analysis.

Acknowledgements

This research was supported by the NSF Faculty Early Career Development (CAREER) Award IIS-0236653 and OBR Hayes Doctoral Incentive Fund Grant Program Fellowship. We additionally thank the OSU Advanced Computing Center for the Arts and Design (ACCAD) for access to the Vicon-8 Motion-Capture Studio, and thank N. Troje at the BioMotionLab of the Ruhr-University in Bochum, Germany for supplying the gender motion-capture data used in these experiments.

References

[1] J. Aggarwal, Q. Cai, Human motion analysis: a review, In Nonrigid Articulated Motion Wkshp, IEEE (1997) 90–102.
[2] A. Ali, J. Aggarwal, Segmentation and recognition of continuous human activity, In Proc. Wkshp. Detect. Recogn. Events Video, IEEE (2001) 28–35.
[3] C. Barclay, J. Cutting, L. Kozlowski, Temporal and spatial factors in gait perception that influence genderrecognition, Percept. Psychophys. 23 (2) (1978) 145–152.
[4] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, Human activity recognition using multidimensional indexing, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1091–1104.
[5] M. Black, Y. Yacoob, A. Jepson, D. Fleet, Learning parameterized models of image motion, In Proc. Comput. Vis. Pattern Rec. (1997) 561–567.
[6] A. Bobick, J. Davis, An appearance-based representation of action, In Proc. Int. Conf. Pattern Rec. (1996) 307–312.
[7] M. Brand, A. Hertzmann, Style machines, In Proc. SIGGRAPH, ACM July (2000) 183–192.
[8] C. Bregler, J. Malik, Tracking people with twists and exponential maps, In Proc. Comput. Vis. Pattern Rec. (1998) 8–15.
[9] R. Burden, J. Faires, Numerical Analysis, PWS, Boston, 1993.
[10] D. Chi, M. Costa, L. Zhao, N. Badler, The EMOTE model for effort and shape, In Proc. SIGGRAPH, ACM (2000) 173–182.
[11] Y. Cui, D. Swets, J. Weng, Learning-based hand sign recognition using SHOSLIF-M, In Proc. Int. Conf. Comput. Vis., IEEE (1995) 631–636.
[12] J. Cutting, D. Proffitt, L. Kozlowski, A biomechanical invariant for gait perception, J. Exp. Psych. 4 (3) (1978) 357–372.
[13] J. Davis, A. Bobick, The representation and recognition of action using temporal templates, In Proc. Comput. Vis. Pattern Rec., IEEE (1997) 928–934.

[14] J. Davis, H. Gao, V. Kannappan, A three-mode expressive feature model of action effort, In Wkshp. Motion Video Comput., IEEE (2002) 139–144.
[15] J. Davis, V. Kannappan, Expressive features for movement exaggeration, In SIGGRAPH Conf. Abstr. Appl., ACM (2002) 182.
[16] J. Davis, S. Taylor, Analysis and recognition of walking movements, In Proc. Int. Conf. Pattern Rec. (2002) 315–318.
[17] J. Davis, Visual categorization of children and adult walking styles, In Proc. Int. Conf. Audio Video-based Biometric Person Authentication (2001) 295–300.
[18] S. Dockstader, K. Bergkessel, A. Tekalp, Feature extraction for the analysis of gait and human motion, In Proc. Int. Conf. Pattern Rec. (2002) 5–8.
[19] D. Gavrila, The visual analysis of human movement: a survey, Comput. Vis. Image Understanding 73 (1) (1999) 82–98.
[20] M. Giese, T. Poggio, Morphable models for the analysis and synthesis of complex motion patterns, Int. J. Comput. Vis. 38 (1) (2000).
[21] G. Johansson, Visual perception of biological motion and a model for its analysis, Percept. Psychophys. 14 (2) (1973) 201–211.
[22] L. Kozlowski, J. Cutting, Recognizing the sex of a walker from dynamic point-light display, Percept. Psychophys. 21 (6) (1977) 575–580.
[23] L. Kozlowski, J. Cutting, Recognizing the gender of walkers from point-lights mounted onankles: some second thoughts, Percept. Psychophys. 23 (5) (1978) 459.
[24] P. Kroonenberg, J. Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1) (1980) 69–97.
[25] P. Kroonenberg, Three-Mode Principal Component Analysis Theory and Applications, DSWO Press, Leiden, 1983.
[26] N. Li, S. Dettmer, M. Shah, Visually recognizing speech using eigensequences, In: Motion-Based Recognition, Kluwer Academic Press, Dordrecht, Norwell, MA. 1997, pp. 345–371.
[27] T. McMahon, Muscles, Reflexes, and Locomotion, Princeton University Press, Princeton, NJ, 1984.
[28] H. Murase, S. Nayar, Visual learning and recognition of 3D objects from appearance, Int. J. Comput. Vis. 14 (1) (1995) 5–24.
[29] R. Neal, C. Snyder, P. Kroonenberg, Individual differences and segment interactions in throwing, Hum. Movement Sci. 10 (1991) 653–676.
[30] V. Pavlovic, J. Rehg, T. Cham, K. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models, In Proc. Int. Conf. Comput. Vis. (1999) 94–101.
[31] F. Pollick, H. Paterson, A. Bruderlin, A. Stanford, Perceiving affect from arm movement, Cognition 82 (2) (2001) B51–B61.
[32] C. Rao, M. Shah, View-invariant representation and learning of human action, In Proc. Wkshp. Detect. Recogn. Events Video, IEEE (2001) 55–63.
[33] R. Rosales, S. Sclaroff, Inferring body pose without tracking body parts, In Proc. Comput. Vis. Pattern Rec., IEEE (2000) 721–727.
[34] S. Runeson, G. Frykholm, Visual perception of lifted weight, J. Exp. Psych. 7 (4) (1981) 733–740.
[35] J. Tenenbaum, W. Freeman, Separating style and content, Adv. Neural Inf. Process. Syst. 10 (1997) 662–668.
[36] N. Troje, Decomposing biological motion: a framework for analysis and synthesis of human gait patterns, J. Vis. 2 (2002) 371–387.
[37] L. Tucker, Some mathematical notes on three-mode factor analysis, Psychometricka 31 (3) (1966) 279–311.
[38] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cogn. Neurosci. 3 (1) (1991) 71–86.
[39] M. Unuma, K. Anjyo, R. Takeuchi, Fourier principles for emotion-based human figure animation, In Proc. SIGGRAPH, ACM (1995) 91–96.
[40] M. Unuma, R. Takeuchi, Generation of human motion with emotion, In Proc. Comput. Anim. (1991) 77–88.
[41] M. Vasilescu, D. Terzopoulos, Multilinear analysis of image ensembles: TensorFaces, In Proc. Eur. Conf. Comput. Vis. (2002) 447–460.
[42] M. Vasilescu, Human motion signatures for character animation, In SIGGRAPH Conf. Abstr. Appl., ACM (2001) 200.
[43] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recogn. 36 (2003) 585–601.
[44] A. Wilson, A. Bobick, Parametric Hidden Markov Models for gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21 (9) (1999) 884–900.
[45] M. Yamamoto, T. Kondo, T. Yamagiwa, K. Yamanaka, Skill recognition, In Proc. Int. Conf. Auto. Face Gesture Recogn. (1998) 604–609.