Recognizing Human Action Efforts: An Adaptive Three-Mode PCA Framework[*]

James W. Davis   Hui Gao

Dept. Computer and Information Science
Ohio State University
Columbus OH 43210 USA
{jwdavis,gaoh}@cis.ohio-state.edu

Abstract

We present a computational framework capable of labeling the effort of an action corresponding to the perceived level of exertion by the performer (low – high). The approach initially factorizes examples (at different efforts) of an action into its three-mode principal components to reduce the dimensionality. Then a learning phase is introduced to compute expressive-feature weights to adjust the model's estimation of effort to conform to given perceptual labels for the examples. Experiments are demonstrated recognizing the efforts of a person carrying bags of different weight and for multiple people walking at different paces.

1. Introduction

There are sub-categorical properties of human actions that can be inferred by observation. For example people can tell the *gender* of a walker [8], estimate the *load* of a lifted box [11], and describe the *emotional* state of an action [10] from looking at simple point-light displays. The hypothesis is that regular visual cues (expressive features) in movements enable observers to reliably recognize the target properties. Our goal is to develop an efficient computational framework that can learn the expressive features indicative of the target property for recognition in accordance with human perceptual judgements of the actions. In this paper, we focus our work on recognizing perceptually-based action *efforts*. We define effort as the "perceived amount of exertion" by the performer, not the measured quantity of physical variables such as mass or speed. Therefore we wish to recognize *qualitative properties*, such as the observed heaviness of carried packages, the leisureness in walking styles, the strain in lifting, etc.

Models of action efforts are relevant to several domains and applications. Automatic video annotation of pedestrian walking pace and carrying load would provide more descriptive information to surveillance systems. Ergonomics could also benefit from effort analysis for evaluation and recommendation therapy. Similarly, visual monitoring of athletic training could help prevent costly sports injuries by recognizing the onset of fatigue (through characteristic changes in effort) during endurance workouts. A computer model of performance efforts could also be used to "warp" motion-capture animations into new efforts, and could potentially be used for searching digital motion libraries to find actions exhibiting a similar (or different) effort as the query example.

We begin with a review of related work (Sect. 2). Next we discuss the three-mode principal component analysis (PCA) model (Sect. 3). A three-mode PCA factorizes the action data (trajectories) across different efforts into a tri-modal separation of pose, time, and effort basis sets. A two-mode factorization with rasterized trajectory data however would only have two basis sets (pose-time, effort). The tri-modal basis enables us to easily emphasize certain trajectories (expressive features) to compute effort values that conform to labeled training data.

We then present an augmented three-mode model with expressive weights and a complementary learning algorithm using training data (Sect. 4). To collect the labeled training data, we introduce a perceptual matching task to measure the perceptual judgements for different action efforts (Sect. 5). We present experimental results recognizing the carrying efforts of a person and the walking paces for multiple people (Sect. 6), showing the advantage of the method over standard sum-of-squared error (SSE) matching. Lastly, we conclude with a summary of the research (Sect. 7).

2. Related Work

There has been much recent work in computer vision on detecting, tracking, and recognizing human actions. With regards to effort and style variation, a Parameterized-HMM was used by [18] 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 [12] for separating perceptual content and style parameters of non-action data. In [6], 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 [5] the motion regularities of walking motions of several people at different speeds were used to classify typical from atypical gaits. A two-mode PCA framework was described in [13] to linearly classify male and female walkers. Morphable models were employed in [7] to represent complex motion patterns by linear combinations of prototype sequences and used for movement analysis and synthesis. A method for recognizing skill was presented in [19] to determine the performance-level of skiers by ranking properties such as synchronous and smooth motion.

In relation to computer animation, a Fourier-based approach was used in [15] to generate human motion with emotional properties (e.g., a happy walk). An HMM with entropy minimization was used by [1] to generate different state-based animation styles. A factorization of motion-capture data for extracting person-specific motion signatures was described in [17]. A movement exaggeration model using measurements of the observability and predictability of joint angle trajectories was presented in [4]. In [3], the EMOTE character animation system using effort (and other) qualities was employed to generate natural synthetic gestures.

3. Three-Mode Action Factorization

Actions can be described as the changing body pose (mode 1) over time (mode 2). When considering dynamic actions, we have a third mode corresponding to the effort (mode 3)[1]. The motion data (trajectories) for multiple effort examples of an action can be organized into a cube Z (see Fig. 1.a), with the rows in each frontal plane Z_k corresponding to the motion trajectories for a particular effort (indexed by k).

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. Singular value decomposition (SVD) and principal components analysis (PCA) are standard two-mode methods for achieving this data reduction. *Three-mode* factorization [14] is an extension of these traditional two-mode methods and offers a framework suitable to incorporating expressive weights on trajectories for efficient, perceptually-driven recognition of action efforts in a low-dimensional space.

---

[1]One could easily argue for additional modes spanning gender, age, etc.
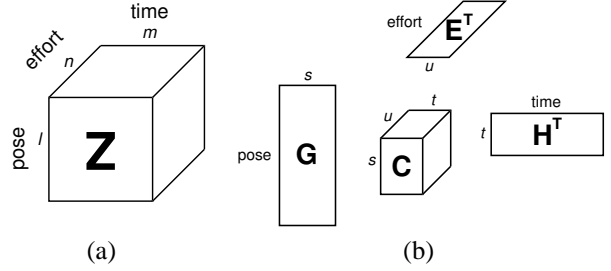
Figure 1. (a) Three-mode arrangement of data for an action at different efforts. (b) Three-mode factorization of the data.

The three-mode factorization of Z decomposes it into three orthonormal matrices G, H, and E that span the column (pose), row (time), and slice (effort) spaces (see Fig. 1.b). Following a similar tensor-based method as described in [16], the three bases can be solved using three different flattening arrangements of Z (after ij-centering)

$$G = \Lambda[\, Z_1 \mid Z_2 \mid \cdots \mid Z_n \,] \qquad (1)$$

$$H = \Lambda[\, Z_1^T \mid Z_2^T \mid \cdots \mid Z_n^T \,] \qquad (2)$$

$$E = \Lambda[\, \vec{Z}_1 \mid \vec{Z}_2 \mid \cdots \mid \vec{Z}_n \,]^T \qquad (3)$$

where $\vec{Z}_k$ is the rasterized column vector of matrix Z_k, and Λ is the column space operator. The desired column spaces in Eqns. 1-3 can be found using SVD. Typically, each mode needs only to retain its first few components (meeting some variance criteria) to capture most of the fit to Z.

The core C has three dimensions and represents the relationships of the components in G, H, and E for reconstructing Z. The three-mode factorization of Z can be concisely written as $Z = GC(E^T \otimes H^T)$, or in flattened matrix form as

$$[\, Z_1 \mid \cdots \mid Z_n \,] = G[\, C_1 \mid \cdots \mid C_u \,](E^T \otimes H^T) \quad (4)$$

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

$$[\, C_1 \mid \cdots \mid C_u \,] = G^T[\, Z_1 \mid \cdots \mid Z_n \,](E^T \otimes H^T)^T \quad (5)$$

The three-mode core C need not be diagonal, as is required in two-mode SVD. An additional method for solving the three-mode factorization using an alternating least-squares algorithm is described in [9].

Any frontal plane Z_k (action at a particular effort) can be reconstructed as

$$Z_k = G \left(\sum_{r=1}^{u} e_{kr} C_r \right) H^T \qquad (6)$$

Therefore, we can recover each training example by choosing the correct e_{kr} component loadings from the effort mode E. When recognizing the effort for an action, these effort loadings are what must be estimated.

If we consider only two efforts examples for an action (Z_1, Z_2), the effort mode reduces to a single component $E = [-1 \ \ 1]^T / \sqrt{2}$. The two effort examples Z_1 and Z_2 therefore differ only in a single effort parameter e, and thus Eqn. 6 is reduced to

$$Z_k = e_k G C H^T \qquad (7)$$

As human movements exhibit smooth and predictable regularity across increasing/decreasing efforts, only two diverse (or extreme) effort examples may be all that is required to successfully model the in-between range of efforts for the action. Hence, additional effort examples will be constrained to have effort values along this 1-D effort continuum. This reduced model with a single effort parameter will be used to simplify the algorithm for learning the expressive features.

4. Expressive Three-Mode Model

The reduced three-mode factorization (Eqn. 7) can be expanded for each data element z_{ijk} of Z as a summation of three-mode elements, where the effort parameter can be isolated from the remaining factored terms as

$$z_{ijk} = e_k \sum_{p=1}^{s} \sum_{q=1}^{t} g_{ip} h_{jq} c_{pq} = e_k \alpha_{ij} \qquad (8)$$

Note that α_{ij} is pre-computable, and for a nearly diagonal core, it can be further reduced to $\alpha_{ij} = \sum_{p=1}^{\min(s,t)} g_{ip} h_{jp} c_{pp}$.

To solve for the effort parameter of an unknown test input $Z_{\hat{k}}$, an SSE error function $\mathcal{F}$ of the form

$$\mathcal{F} = \|Z_{\hat{k}} - eGCH^T\|^2 = \sum_i \sum_j (z_{ij} - e \cdot \alpha_{ij})^2 \qquad (9)$$

can be minimized. Traditional squared-error techniques place equal prior emphasis on the different terms during minimization. But all trajectories may not have the same discrimination power with respect to the action effort. Given training data of different effort examples labeled according to some criteria (e.g., perceptual judgements of the effort), certain features (of motion and/or pose) will likely be more informative of the effort than others. Therefore we would like to place more emphasis on those expressive features (position trajectories) during the effort estimation process to produce results that most closely match the recognition criteria.

Following this concept, we augment $\mathcal{F}$ with expressibility weights $\mathcal{E}_i$ on each trajectory

$$\hat{\mathcal{F}} = \sum_i \mathcal{E}_i \sum_j (z_{ij} - e \cdot \alpha_{ij})^2 \qquad (10)$$

To minimize $\hat{\mathcal{F}}$ for estimating the target effort parameter, we set $\frac{\partial \hat{\mathcal{F}}}{\partial e} = 0$ and re-arrange it to produce

$$\hat{e} = \frac{\sum_i \mathcal{E}_i \sum_j z_{ij} \alpha_{ij}}{\sum_i \mathcal{E}_i \sum_j \alpha_{ij}^2} \qquad (11)$$

Setting the expressive weights $\mathcal{E}_i = 1$ in Eqn. 11 yields the a standard SSE (least-squares) estimation of effort. This is also equivalent to using standard two-mode PCA with rasterized (rank-1 tensor) data to recover the effort parameter (projection coefficient). With non-uniform expressive weights, the approach can adapt to the specifics of labeled training data.

4.1. Learning

The next task is to learn appropriate values for the expressive feature weights to compute efforts that correspond to the effort values assigned to the training data. Given a set of K training effort examples, we first use the two extreme efforts to construct the reduced three-mode model. Then we define the matching error of the training labels $\bar{e}_k$ with the *computed* expressive model efforts $\hat{e}_k$ as

$$
\begin{aligned}
J &= \sum_k (\bar{e}_k - \hat{e}_k)^2 \qquad &(12)\\
&= \sum_k \left(\bar{e}_k - \frac{\sum_i \mathcal{E}_i \sum_j z_{ijk} \alpha_{ij}}{\sum_i \mathcal{E}_i \sum_j \alpha_{ij}^2} \right)^2 \qquad &(13)\\
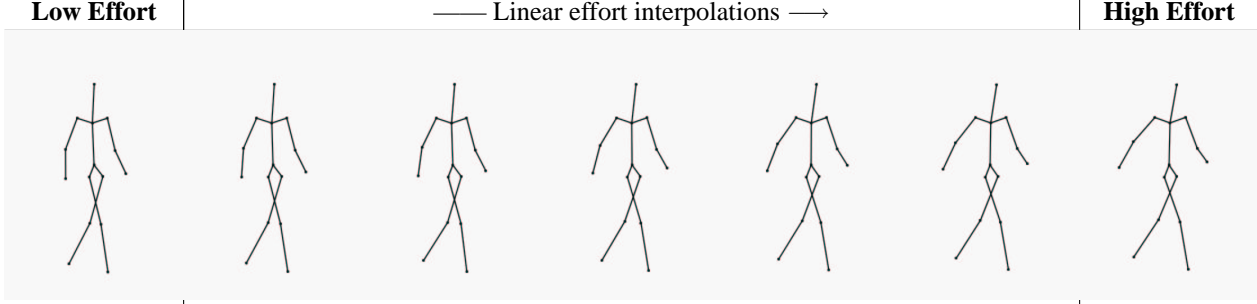&= \sum_k \left(\bar{e}_k - \frac{\sum_i \mathcal{E}_i B_{ik}}{\sum_i \mathcal{E}_i A_i} \right)^2 \qquad &(14)
\end{aligned}
$$

The non-linear arrangement of the expressive weights in Eqn. 14 can be solved using a fast iterative gradient descent algorithm [2] of the form

$$\mathcal{E}_i(n+1) = \mathcal{E}_i(n) - \eta(n) \cdot \frac{\partial J}{\partial \mathcal{E}_i} \qquad (15)$$

with the gradients $\frac{\partial J}{\partial \mathcal{E}_i}$ computed over the K training examples as

$$
\begin{aligned}
\frac{\partial J}{\partial \mathcal{E}_i} &= 2 \sum_k \left(\bar{e}_k - \frac{\sum_j \mathcal{E}_j B_{jk}}{\sum_j \mathcal{E}_j A_j} \right) \\
&\cdot \frac{A_i \sum_j \mathcal{E}_j B_{jk} - B_{ik} \sum_j \mathcal{E}_j A_j}{\left(\sum_j \mathcal{E}_j A_j \right)^2} \qquad (16)
\end{aligned}
$$

Low Effort | —— Linear effort interpolations ⟶ | **High Effort**

Figure 2. Walking poses at frame #1 in a series of synthetic motions created with different effort values. Increasing arm swing and stride are visible across the increasing walking pace.

The learning rate η is re-computed at each iteration to yield the best incremental update. The expressive weights are initialized to $\mathcal{E}_i = 1$ (SSE formulation) and confined to be positive. Following convergence of Eqn. 15, effort values for new examples can now be estimated with Eqn. 11 using the learned expressive weights.

5. Perceptually-Labeled Training Data

To label the efforts of the training data for learning the expressive weights, we conducted a series of perceptual tests with stick-figure motions in which people were asked to match different effort examples to synthetic motions with known effort values. The synthetic motions were produced by linearly interpolating/extrapolating the two extreme effort examples used to create the reduced three-mode model (see Fig. 2). The synthetic motions were used as a linear "ruler" to label the efforts for the training data.

A screen-shot of the computer display for the matching task is shown in Fig. 3, with one *unlabeled* effort example on the left and one synthetic (*labeled*) example on the right. By using the left and right arrow keys on the keyboard, the user is asked to seek through the synthetics to find the best example that gives a similar overall feeling of the effort shown in the unlabeled example on the left. After the user confirms the choice, the program records the matching result and loads the next example. The unlabeled examples were presented in random order and looped in synchronization with the displayed synthetic example. No time restriction was enforced during the matching task.

The mean and standard deviation of the assigned effort values for the unlabeled examples were then computed from the matching results of multiple people who took the test. We set the effort label $\bar{e}_k$ to the (perceptual) mean of the synthetic effort values chosen for training example k. We also computed an influence factor $\omega(\sigma_k)$ to give more importance in the learning algorithm to those examples hav-
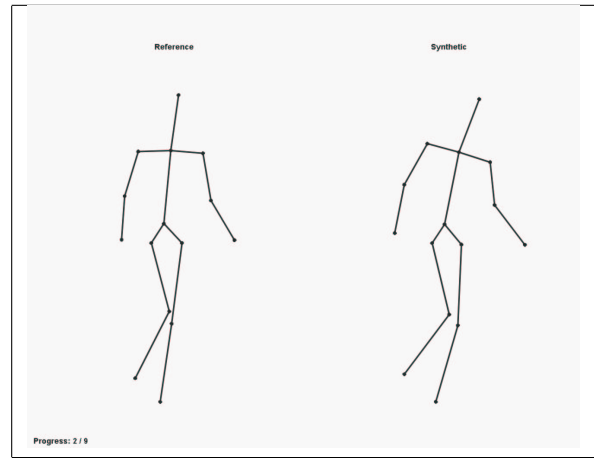
Figure 3. Perceptual matching task display.

ing smaller standard deviations in the perceived effort (i.e., having more consistent choices across people). The new matching error function is

$$J = \sum_k \omega(\sigma_k) \cdot \left(\bar{e}_k - \frac{\sum_i \mathcal{E}_i B_{ik}}{\sum_i \mathcal{E}_i A_i} \right)^2 \qquad (17)$$

with the influence factor $\omega(\sigma_k) = \exp(-\sigma_k^2/.25)$.

6. Experiments

We demonstrate the potential of our framework with experiments modeling and recognizing the carrying effort of a person and the walking pace of multiple people.

As the focus of this work is a representation for motion recognition, we tested the approach with trajectories attained from a Vicon-8 motion-capture system (future work will incorporate a video-based human body tracker). A limited skeleton with 15 joint positions was generated (see Fig.

3). Two cycles for each action effort were automatically extracted (using trajectory curvature peaks), averaged, and time-normalized to a fixed duration using spline interpolation (42 frames for carrying, 50 frames for walking). Time-normalization is required by PCA and does not affect our approach (i.e., we are not measuring absolute speeds).

The 3-D motion trajectories were rendered in 2-D at 30 Hz for the perceptual matching task. All motions were viewed from a 45 degree orthographic camera. For training and recognition, the 2-D trajectories were automatically converted into Local-Position-Space (LPS) trajectories to achieve certain view invariants. Each 2-D joint position (e.g., r-hand) was converted by **1)** setting the local axis origin at its parent joint (r-elbow), **2)** placing the "up" direction of local Y axis along the parent's link (r-elbow:r-shoulder), and **3)** scaling the joint position by a body-centric factor (e.g., torso length). LPS is a simple 2-D hierarchical representation with local rotation invariance, tolerance to different body scales, and use of spatial coordinates. In this work, we examined 20 x-y LPS trajectories (from 10 limb and torso joints).

6.1. Carrying Effort

Our first experiment was to model the visual effort of a person carrying packages of different weight. The motions were captured of the person walking on a treadmill at a constant speed (1.4 mph) while carrying (in one hand) 9 differently loaded bags (0 – 40 lbs).

Ten subjects participated in the perceptual matching task. The average correlation coefficient of the example-to-synthetic perceptual mappings (all pairwise combinations of subjects) was $r = .8$ (SD .1). When building the three-mode model of Z_{carry} ($20 \times 42 \times 2$ cube of LPS trajectories of the lightest and heaviest carry) with an 85% variance criterion in each mode, G, H, and E had 4, 4, and 1 components respectively. This basis captured 99% of the overall data variance. The learning algorithm (after 1500 iterations) produced non-zero expressive weights for trajectories in the right-knee, head, left-elbow, right-elbow, and right-hand (nearly zero). Several trajectory weights (15/20) were set to zero and therefore were not considered as informative of the effort (see Fig. 4.a).

In Fig. 4.b, we show the effort matches between the 9 unlabeled training examples and the synthetic motions using the perceptual effort means, standard SSE effort estimations, and expressive effort computations. We again mention that the non-expressive SSE result is the same as a two-mode PCA projection using rasterized motion data. The results show a noticeable difference mainly at examples #6 and #7. Upon inspection, the counter-balancing left-arm was basically ignored in the perceptual match to the synthetics, even though the arm had considerable deviation across


(a)

(b)
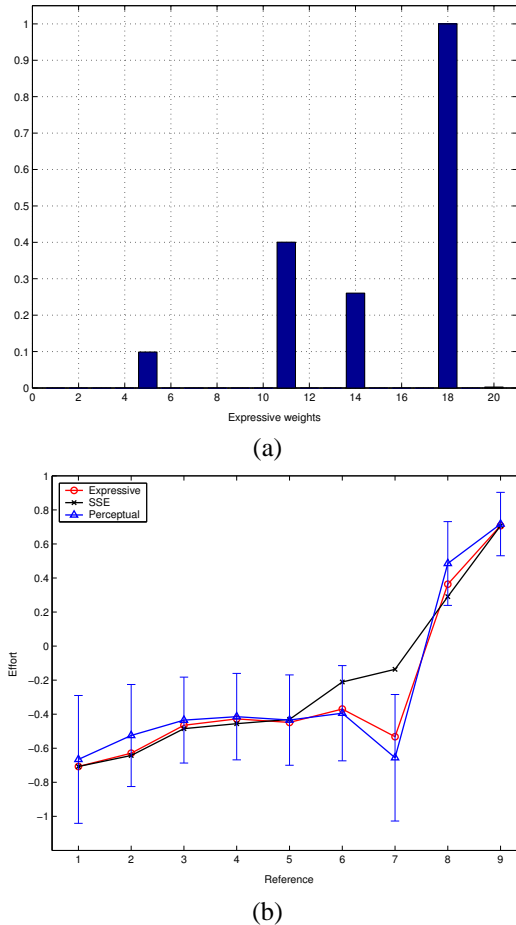
Figure 4. Carry results. (a) Learned expressive weights (normalized). (b) Perceptual (mean $\pm$ 1 SD), SSE, and expressive effort estimations.

effort. The perceptual matching process is willing to sacrifice certain joint motions in order to attain a more "global" alignment of two movements. As a result of the learning algorithm, the expressive matches closely resembled the perceptual choices (see Fig. 4.b).

6.2. Walking Pace of Multiple People

We next demonstrated the framework using the efforts of multiple people. Our goal was to recognize the qualitative walking pace (leisurely – quickly) rather than absolute speed[2]. The training set was comprised of three people, each with five different walking paces (recorded using a treadmill at speeds ranging between 2.0 mph and each

---

[2]Actual speed is not a global indicator of walking pace for different people (consider child vs. adult walking).
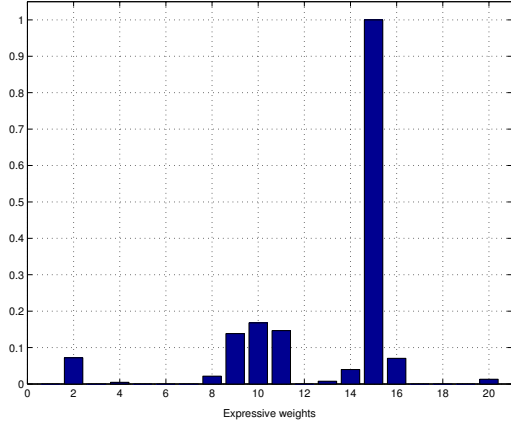
Figure 5. Learned expressive weights (normalized) for walking pace.

person's walk-run transition).

We first constructed low and high effort "prototypes" using the mean of the slowest walk and the mean of the fastest walk of the three people (computed using a mean skeleton with averaged joint-angle trajectories). Given these prototypes, the perceptual matching task and the reduced three-mode factorization can be performed as with the single person case.

The same 10 people from the previous perceptual experiment were asked to match all 15 walking examples to the synthetic walking motions generated from the two mean-prototypes (see Fig. 2). The examples were automatically scaled to a fixed height in the display to accommodate the different person statures. The average correlation coefficient of the example-to-synthetic perceptual mappings was again $r = .8$ (SD .1). With an 85% modal variance criterion for Z_{walk} ($20 \times 50 \times 2$ cube of LPS trajectories of the two mean prototypes), G, H, and E had 5, 5, and 1 components respectively. This basis captured 98% of the overall data variance. The resulting non-zero expressive weights from the learning algorithm corresponded to trajectories in the left-knee, left-foot (nearly zero), right-foot, torso, left-elbow, left-hand, and right-hand (nearly zero). Nearly half of the expressive weights were zero (see Fig. 5).

We compared the results for each walker separately, examining the perceptual, SSE, and expressive efforts. For Person-1, the perceptual and expressive efforts were very similar but the SSE efforts were quite different from the perceptual choices (see Fig. 6.a). The advantage of the expressive model is well illustrated. For Person-2, all effort methods produced essentially the same results (see Fig. 6.b). Lastly, the perceptual and expressive efforts for Person-3 were again similar, but the SSE results were somewhat different (see Fig. 6.c).

We also tested two additional walking motions from each person (one at 1.6 mph and one midway between 2.0 mph and the walk-run transition) to evaluate the expressive effort estimation for motions not included in the training set. The expressive effort values for these motions closely matched the perceptual results (see Fig. 6.d-f). For both paces of Person-1 and the second pace of Person-2, the non-expressive SSE effort estimation results were significantly different from the desired perceptual choices.

7. Conclusion

We presented an approach for modeling and recognizing action efforts using an efficient three-mode PCA framework that gives more influence to key expressive trajectories learned from a perceptual effort-matching task.

The approach initially factorizes a set of low and high effort examples of an action into its three-mode principal components. We then augment the standard least-squares estimation of effort within this basis to include expressive weights (one for each trajectory) to bias the computations with the trajectories most indicative of the effort. Labeled training data were used in a gradient descent learning algorithm to solve for the expressive weight values needed to produce the desired effort estimations. To collect the effort-labeled training data, a perceptual matching task was conducted that mapped the efforts of synthetically-generated data to real examples.

The approach was demonstrated with experiments examining the efforts of a person carrying bags of different weight and for multiple people walking at different paces. We showed that a standard SSE (two-mode PCA) method does not always conform to human preferences in these examples, and that our three-mode model with the learned expressive weights can be used to produce perceptually-similar effort results.

In future work, we plan to incorporate a video-based body tracking algorithm and broaden the range of actions to include non-periodic activities (e.g., throwing, lifting, jumping). We will also extend the framework to recognize other sub-categorical properties (e.g., gender and emotion) and build a higher-level action style analysis system. Lastly, we are seeking to unify the three-mode model with an action classification method to produce a single framework for recognizing human actions and efforts.

Acknowledgments
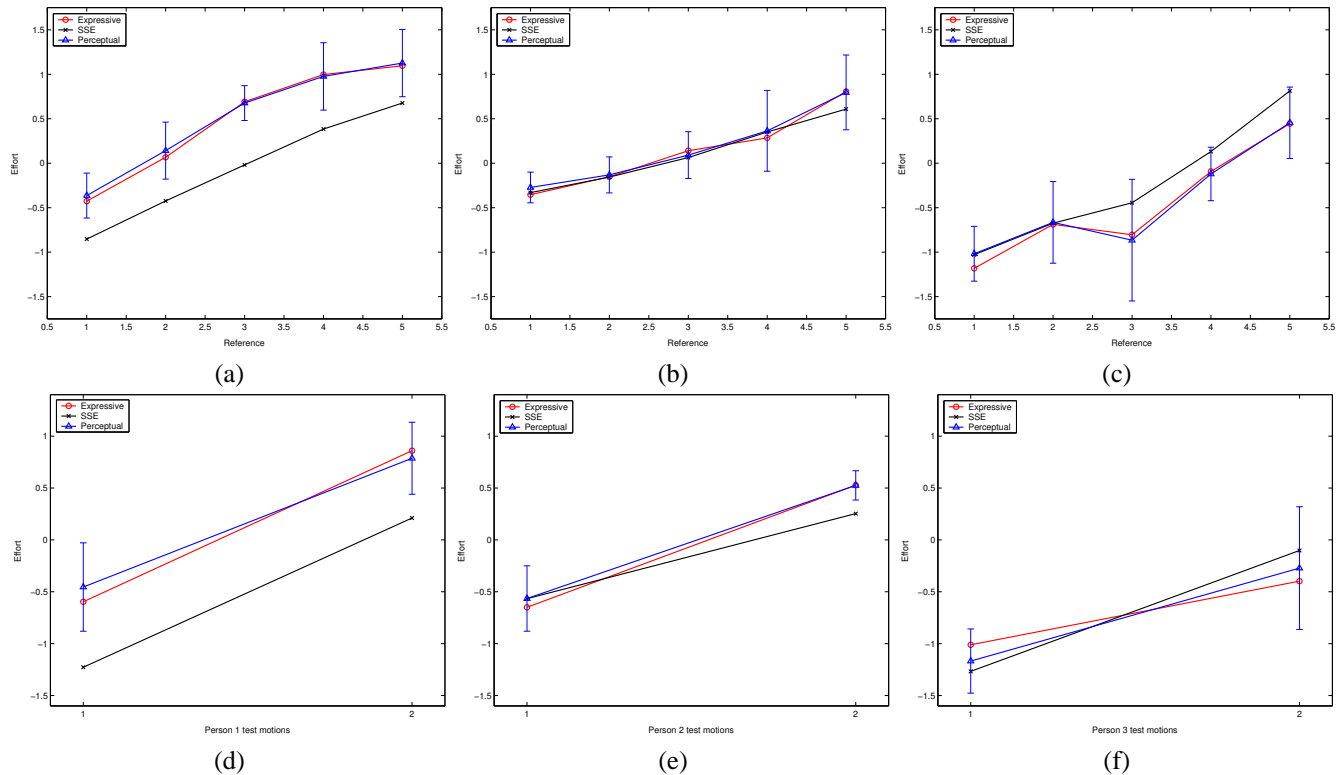
Figure 6. Walking pace results showing perceptual (mean $\pm$ 1 SD), SSE, and expressive effort estimations. Results of training data for (a) Person-1, (b) Person-2, and (c) Person-3. Results for testing data of (d) Person-1, (e) Person-2, and (f) Person-3.

Computing Center for the Arts and Design (ACCAD) for access to the Vicon-8 Motion-Capture Studio.

References

[1] M. Brand and A. Hertzmann. Style machines. In *Proc. SIG-GRAPH*, pages 183–192. ACM, July 2000.

[2] R. Burden and J. Faires. *Numerical Analysis*. PWS, Boston, 1993.

[3] D. Chi, M. Costa, L. Zhao, and N. Badler. The EMOTE model for effort and shape. In *Proc. SIGGRAPH*, pages 173–182. ACM, 2000.

[4] J. Davis and V. Kannappan. Expressive features for movement exaggeration. In *SIGGRAPH Conference Abstracts and Applications*, page 182. ACM, 2002.

[5] J. Davis and S. Taylor. Analysis and recognition of walking movements. In *Proc. Int. Conf. Pat. Rec.*, pages 315–318, 2002.

[6] J. Davis. Visual categorization of children and adult walking styles. In *Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication*, pages 295–300, 2001.

[7] M. Giese and T. Poggio. Morphable models for the analysis and synthesis of complex motion patterns. *Int. J. of Comp. Vis.*, 38(1), 2000.

[8] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from dynamic point-light display. *Perception & Psychophysics*, 21(6):575–580, 1977.

[9] P. Kroonenberg and J. Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97, 1980.

[10] F. Pollick, H. Paterson, A. Bruderlin, and A. Stanford. Perceiving affect from arm movement. *Cognition*, 82(2):B51–61, 2001.

[11] S. Runeson and G. Frykholm. Visual perception of lifted weight. *J. of Exp. Psych.*, 7(4):733–740, 1981.

[12] J. Tenenbaum and W. Freeman. Separating style and content. *Advances in Neural Information Processing Systems*, 10:662–668, 1997.

[13] N. Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *J. of Vision*, 2:371–387, 2002.

[14] L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometricka*, 31(3):279–311, 1966.

[15] M. Unuma, K. Anjyo, and R. Takeuchi. Fourier principles for emotion-based human figure animation. In *Proc. SIG-GRAPH*, pages 91–96. ACM, 1995.

[16] M. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: TensorFaces. In *Proc. European Conf. Comp. Vis.*, pages 447–460, 2002.

[17] M. Vasilescu. Human motion signatures for character animation. In *SIGGRAPH Conference Abstracts and Applications*, page 200. ACM, 2001.

[18] A. Wilson and A. Bobick. Parametric Hidden Markov Models for gesture recognition. *IEEE Trans. Patt. Analy. and Mach. Intell.*, 21(9):884–900, 1999.

[19] M. Yamamoto, T. Kondo, T. Yamagiwa, and K. Yamanaka. Skill recognition. In *Proc. Int. Conf. on Auto. Face and Gesture Recognition*, pages 604–609, 1998.