A Three-Mode Expressive Feature Model for Analysis and Recognition of Action Effort

James W. Davis Hui Gao Vignesh S. Kannappan
Dept. of Computer and Information Science
Ohio State University
Columbus, OH 43210 USA
{jwdavis,gaoh,kannappa}@cis.ohio-state.edu

Abstract

We present an expressive feature model for recognizing the performance effort of human actions. A set of low and high effort examples for an action are initially factorized into its three-mode principal components, followed by a learning phase to compute the expressive features required to bring the model estimation of effort into agreement with perceptual judgements. The approach is demonstrated using real and illusory movements.

1. Introduction

"Dynamic movements" are categories of movements performed under varying degrees of effort. Does his walk appear *leisurely*, or is he walking in a *hurry*? Does that package look *heavy* or *light* for her to carry? People are quite adept at identifying the amount of effort exerted by a person from subtle visual cues. For example, nuances in body motion can reveal how light or heavy a box is for a person to lift [16]. Our goal is to develop computational systems capable of identifying the "perceptual dynamics" associated with different performance efforts for the task of recognition. Intelligent machines capable of recognizing human action efforts are particularly relevant to automatic video surveillance, ergonomic evaluation, and sports analysis.

Instead of matching new movements to multitudes of training examples based on their proximity in some feature space, we present a system that learns the "expressive" motion features associated with a dynamic movement and computes a metric estimation of the effort. Our approach first constructs a three-mode factorization of low and high effort examples and then tunes the model using expressive features derived from additional labeled effort examples. A weighted-SSE minimization technique using perceptually-labeled training data is employed to approximate human judgements of effort similarity. We demonstrate the approach with the task of recognizing the amount of effort exerted by a person carrying a bag. Two movement illusions are also employed to further demonstrate the perceptual similarity of the model.

An important result demonstrated in this research is that people do not match motion sequences to examples by minimizing standard SSE of position or joint-angle trajectories. We present a three-mode weighted-SSE model can be used to produce results similar to human observations.

The remainder of this paper is described as follows. In Sect. 2, we present related work on effort and style analysis. Section 3 motivates the use of expressive (key) features. The overall framework is presented in Sect. 4, describing the three-mode factorization technique, the estimation of action effort, and the expressive feature learning algorithm. In Sect. 5, we outline the experiments performed to demonstrate the approach, followed by the results in Sect. 6. Lastly, we discuss the relevance of the approach to computational systems in Sect. 7, and present a conclusion of the research in Sect. 8.

2. Related Work

There has been much recent work in computer vision on the detection, tracking, and recognition of human actions (See reviews [1, 8]). In this paper, we present a general three-mode approach for the analysis and recognition of the performance effort of human actions.

The most related research addressing a three-mode analysis of human movements over various performance efforts is presented in [12]. Arm segment velocities of 12 athletes throwing three differently weighted balls were examined using a three-mode factorization. The components themselves were *manually* inspected in an attempt to determine loadings signifying horizontal/vertical velocities, proximal/distal velocities, various throwing phases, and different skill levels of the throwers.

Closely related to the recognition of action effort is analysis of style. A bilinear model was used in [17] for separating perceptual content and style parameters. A Fourier-based approach to generate human motion with emotional properties was described in [20]. A Hidden Markov Model (HMM) with entropy minimization was used by [3] to generate different state-based animation styles, and a Parameterized-HMM was used by [24] to model stylistic gesture variations. A factorization of motion capture data for extracting person-specific motion signatures was described in [22], and a movement exaggeration model using measurements of observability and predictability of joint angles was presented in [7].

3. Expressive Features as Key Features

The success of perception (machine or man) relies on the ability to construct model representations whose assumptions and constraints reflect the structure and regularity of the world. In the absence of any domain knowledge or preference for certain features, any two object classes share the same number of properties and thus cannot be distinguished when compared over all possible features (referred to as the ugly-duckling theorem [23]). A "key feature" is a property that can be reliably inferred in a particular context, where the likelihood of correctly indicating the property is high (few false targets) and the property itself has a significant prior probability of occurring [14].

With respect to recognizing action effort, do some properties of the movement vary consistently across effort to enable reliable discrimination of effort (e.g., which joint motions contribute the most to the overall percept of the action effort)? We refer to these key features as *expressive features*[1]. We will show that standard minimization of SSE over all trajectories is not the approach used by human observers when matching movements, but that certain trajectories are used to drive the perceptual rankings and therefore act as key features. We will present a model that can infer these expressive features to enable the model to make judgements similar to those of people.

4. Three-Mode Expressive Model

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 and principal components analysis are standard methods for achieving this data reduction, and have been successfully applied to several *two-mode* problems in computer vision (e.g., [19, 11, 2]). *Three-mode* factorization [18] is an extension of these traditional two-mode methods and offers a framework suitable to incorporating expressive features for efficient recognition of action efforts in a low-dimensional space.

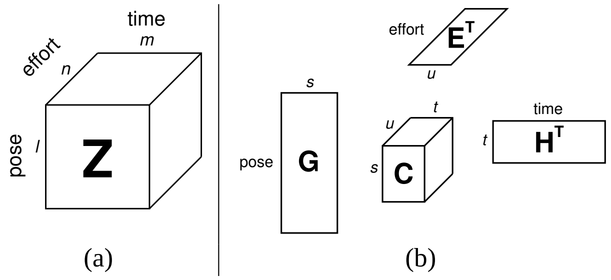


Figure 1: (a) Three-mode configuration of dynamic-effort movement data. (b) Three-mode factorization of data.

Dynamic movements are inherently three mode: body pose (mode 1), time (mode 2), and effort (mode 3). With the use of time-normalized trajectories as input, the data can be organized into a cube Z (See Fig. 1.a) with the rows in each frontal plane Z_k comprising all the trajectories for a particular effort index k. This data cube could be "flattened", or rasterized, into an ordinary two-mode matrix, but this simply ignores the underlying three-mode nature of the data.

Three-mode factorization decomposes the data cube Z into three orthonormal matrices G, H, and E that span the column (pose), row (time), and slice (effort) spaces (See Fig. 1.b). Typically, each mode needs only to retain its first few components to capture most of the fit to Z. Note that any two of the three basis sets cannot be produced within a single two-mode factorization. The core matrix C has three dimensions and represents the relationships of the components in G, H, and E for reconstructing the original data in Z. An alternating least-squares algorithm for solving the three-mode factorization is presented in [9]. A related tensor decomposition can be found in [21].

The three-mode factorization of Z can be concisely written in matrix form as

$$Z = GC(H^T \otimes E^T) \tag{1}$$

where $\otimes$ is the Kronecker product. Any frontal plane Z_k for a given effort index k (an action at a particular effort) can be formulated as

$$Z_k = G\left(\sum_{r=1}^{u} e_{kr} C_r\right) H^T \tag{2}$$

Therefore, we can reconstruct any frontal plane Z_k by choosing the correct e_{kr} component loadings from the effort mode E. To identify an unknown effort for a movement, we must solve for the proper e_{kr} loadings.

4.1. Estimating Action Effort

The three-mode factorization for each data element z_{ijk} of Z can be written as a summation of three-mode elements,

[1] An alternate meaning of expressive features is given by [6].

where the effort loadings can be isolated from the remaining factorized terms

$$\begin{aligned} z_{ijk} &= \sum_{p=1}^{s}\sum_{q=1}^{t}\sum_{r=1}^{u} g_{ip}h_{jq}e_{kr}c_{pqr} && (3)\\ &= \sum_{r=1}^{u} e_{kr}\left(\sum_{p=1}^{s}\sum_{q=1}^{t} g_{ip}h_{jq}c_{pqr}\right) && (4)\\ &= \sum_{r=1}^{u} e_{kr}\alpha_{ijr} && (5)\end{aligned}$$

If we have a nearly diagonal core (with $c_{pqr} \approx 0$ when $p \neq q$), we can further reduce the computations with $\alpha_{ijr} = \sum_{p=1}^{\min(s,t)} g_{ip}h_{jp}c_{ppr}$. The e_{kr} values in Eqn. 5 can be estimated using least-squares methods.

Human movement exhibits smooth and predictable regularity with changes in the dynamic condition [10]. Therefore only a few examples captured at distinct dynamic efforts may be all that is required to successfully model the actions. If we consider only two extreme efforts for an action (e.g., slow/fast walking, light/heavy lifting), the three-mode factorization of Z (after mean-subtraction along the effort dimension) is reduced to contain a single effort parameter e. Movements examined between these extreme efforts should not deviate considerably from this three-mode basis.

To estimate the effort value for a movement of unknown effort with this reduced model, we can solve an error function $\mathcal{F}$ using the sum-of-squares of the input and its formulated reconstruction

$$\mathcal{F} = \sum_i \sum_j \left(z_{ij} - e \cdot \alpha_{ij}\right)^2 \qquad (6)$$

As all trajectories may not equally discriminate the action effort (all may not be expressive features), we augment the error function with expressibility weights $\mathcal{E}_i$ for each of the feature-i trajectories

$$\mathcal{F}' = \sum_i \mathcal{E}_i \sum_j \left(z_{ij} - e \cdot \alpha_{ij}\right)^2 \qquad (7)$$

For minimizing $\mathcal{F}'$ to estimate the target effort parameter, we compute the derivative with respect to e and re-arrange to produce

$$\hat{e} = \frac{\sum_i \mathcal{E}_i \sum_j z_{ij}\alpha_{ij}}{\sum_i \mathcal{E}_i \sum_j \alpha_{ij}^2} \qquad (8)$$

Setting the $\mathcal{E}_i$ values to 1 in Eqn. 8 yields the standard least-squares estimation of e (also attainable from two-mode methods). Non-uniform expressive feature weights in Eqn. 8 can be used to compute effort values that do not necessarily correspond to a standard minimization of SSE (as will be demonstrated with human perceptual judgements of the actions).

4.2. Learning Expressive Features

To learn the expressive feature weights $\mathcal{E}_i$, we construct a second error function J comparing additional training examples with known e values (method to attain the e values will be described in Sect. 5.4) to the estimated $\hat{e}$ values computed with Eqn. 8.

For a set of k training motions $\tilde{Z}_k$ and their efforts $\tilde{e}_k$, we define the matching error as

$$\begin{aligned} J &= \sum_k \left(\tilde{e}_k - \hat{e}_k\right)^2 && (9)\\ &= \sum_k \left(\tilde{e}_k - \frac{\sum_i \mathcal{E}_i \sum_j \tilde{z}_{ijk}\alpha_{ij}}{\sum_i \mathcal{E}_i \sum_j \alpha_{ij}^2}\right)^2 && (10)\\ &= \sum_k \left(\tilde{e}_k - \frac{\sum_i \mathcal{E}_i B_{ijk}}{\sum_i \mathcal{E}_i A_{ij}}\right)^2 && (11)\end{aligned}$$

This non-linear arrangement of the $\mathcal{E}_i$ values can be solved using a fast iterative gradient descent algorithm [5] of the form

$$\mathcal{E}_i(n+1) = \mathcal{E}_i(n) - \eta(n) \cdot \frac{\partial J}{\partial \mathcal{E}_i} \qquad (12)$$

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

$$\begin{aligned} \frac{\partial J}{\partial \mathcal{E}_i} &= 2\sum_k \left(\tilde{e}_k - \frac{\sum_i \mathcal{E}_i B_{ijk}}{\sum_i \mathcal{E}_i A_{ij}}\right)\\ &\quad \cdot \frac{A_{ij}\sum_i \mathcal{E}_i B_{ijk} - B_{ijk}\sum_i \mathcal{E}_i A_{ij}}{\left(\sum_i \mathcal{E}_i A_{ij}\right)^2} && (13)\end{aligned}$$

The learning rate η is re-computed at each iteration to yield the best incremental update. A random-restart approach is employed to handle local minima. Following convergence of Eqn. 12, effort values can be estimated for new input movements using Eqn. 8.

5. Experiments

We analyzed our approach in the context of determining the "carrying effort" of a person holding a bag (in one hand) of increasing weight while walking on a treadmill.

Motion data of the movements were collected, and the lightest and heaviest carry were used to construct the three-mode basis. People were then asked to compare each of the carrying movements to a set of synthetic movements sampled from the three-mode model and choose the best match (mimicking the computer recognition process). Given these perceptual mappings, the expressive feature weights were automatically learned to tune the three-mode model to produce effort values similar to the human judgements. Two illusory movements were additionally tested with the learned model and compared to human observations.

5.1. Motion Capture

A Vicon-8 motion capture system with 14 video cameras was used to create a hierarchical skeleton of the body with 3-D joint-angle trajectories sampled at 30 Hz (Acclaim ASF/AMC format). The trajectories for the movements were lowpass filtered using a 5th order, zero-phase forward-and-reverse Butterworth filter with cut-off at 6 Hz. Two walk cycles were automatically extracted from each sequence using trajectory curvature peaks and averaged into a single walk cycle. The joint positions as seen from a camera placed between a front and side view of the person (45 degrees) were computed and rendered. Recent video tracking advances that could be applied to generate these joint positions include [15, 13, 4], and will be investigated in future work (the focus in this paper is movement representations for recognition).

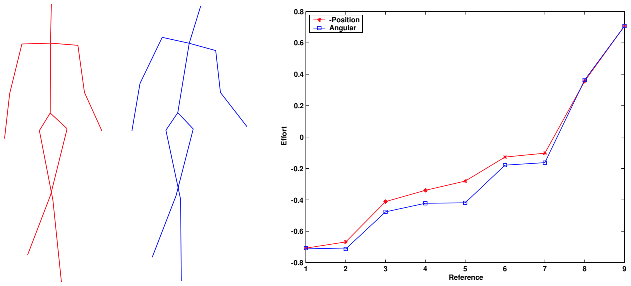
We captured 9 carrying sequences (carrying 0 – 40 lbs, in 5 lb increments) for a person walking on a treadmill at 1.4 MPH. The lowest (lightest) and highest (heaviest) carrying efforts are shown in Fig. 2.a. Additionally, we produced 22 "synthetic" movements by linearly interpolating and extrapolating a three-mode model created from the position data for the lightest and heaviest carry (producing 2 lighter, 15 interpolated, and 5 heavier). These synthetic movements will be used in the perceptual mapping task.

5.2. Input Representation

At this point we must decide on a representation for the movements to construct the three-mode basis for recognition. Rather than committing to any particular set of higher-level composite feature definitions, we represent movements more generally as sets of low-level motion trajectories. Figure 2.b compares the results of a standard least-squares effort estimation with the three-mode model for the carry examples using x-y joint positions and 2-D joint-angles. As both methods produce essentially the same results, we selected the 2-D angle representation as it has fewer degrees-of-freedom (10) and more invariants (e.g., translation, scale, and possibly rotation).

5.3. Illusory Movements

We additionally created an artificial low-effort carry $\check{Z}_{low}$ and high-effort carry $\check{Z}_{high}$ from manually altering the motion capture data. These two movements, when compared to the 22 synthetic model movements using a non-expressive effort estimation (Eqn. 8 with all $\mathcal{E}_i = 1$), both map to the same synthetic movement (#11 of 22). However, these artificial movements perceptually appear quite distinctly as light and heavy efforts: $\check{Z}_{low} \rightarrow$ LOW-EFFORT and $\check{Z}_{high} \rightarrow$ HIGH-EFFORT. These illusions will be used to further demonstrate the perceptually-based behavior of the model that was trained only using the real motions.



(a) (b)

Figure 2: (a) Light (left) and heavy (right) carry figures. A bag is held in the right hand. (b) Position vs. angular effort estimation.

5.4. Perceptual Mapping Task

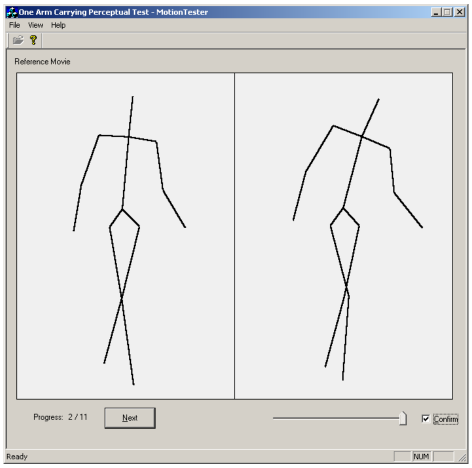
Nine people were given the task of matching the carrying movements to the set of synthetic model motions to provide a mapping for Eqn. 12 to learn the expressive features. Each person was capable of distinguishing side-by-side examples of extreme light and heavy carrying movements prior to the matching task.

A computer program was implemented to conduct the mapping task, and is shown in Fig. 3. On the left, one of 11 carry movies (9 real, 2 illusory) is shown as a "reference" movie. On the right, one of the 22 synthetic model movies is shown. Below the synthetic movie, a slider bar is provided for the user to quickly and easily seek through the possible synthetics. Moving the slider bar to the left or right instantly displays lighter or heavier looking movies, respectively. The reference and synthetic movies are synchronized and played repeatedly. The reference movies are selected in random order for each trial. The slider bar is initially set to a random position for each reference movie.

Once the user has made a choice of which synthetic movie most closely resembles the reference movie, the person selects a confirmation box and clicks the NEXT button to load the next reference movie. The program records the synthetic movie choice for each reference movie. Two complete trials were required, with the first trial used only to familiarize the person with the program and the movies.

5.5. Learning Perceptual Features

After the perceptual task is completed, each reference movie (for each person) is mapped to a particular synthetic movie generated from the original three-mode model. For each selected synthetic movie, there exists a known e value (used to generate the movie). The mean and standard deviation for the reference-synthetic e mappings of the nine people are then computed (not including the illusory movies). Using the mean e values, Eqn. 12 is converged to the ex-



Progress: 2 / 11

☑ Confirm

Figure 3: Screen-shot of perceptual task program.

pressive feature weights needed to bring the model estimation of effort (Eqn. 8) into alignment with the human judgements of the movements.

6. Results

We present the results of the perceptual matching task in Fig. 4.a. Rather than a smooth mapping from low (light) to high (heavy) effort, the 7 lower-effort carry movements visually appeared similar, yet distinct from the remaining 2 higher-effort carry movements. The average mapping correlation of pairwise subjects was $\rho = 0.91$ (SD 0.05). The standard least-squares effort estimate is also shown for comparison. There is a noticeable difference around reference movies 6 and 7.

The mean e values determined from the perceptual mapping task were used in Eqn. 12 to learn the expressive feature weights. The method consistently converged to the same trajectory weights, with the lowerback (.36), right-hip (.15), neck (.59), and right-elbow (1.0) determined to be the expressive features (See Table 1). These results are meaningful as they relate to the increased leaning of the body and straightening of the carry arm as weight was added to the bag. The left counterbalance arm was not found to be an expressive feature even though it had considerable deviation from low to high effort (thus affecting the standard least-squares estimation). In Fig. 4.b, the efforts estimated with the learned three-mode model are compared with the perceptual means.

We additionally tested the sensitivity of the gradient descent algorithm by adding random noise to the perceptual mean values (within .5 SD for each reference movie). The normalized average of the expressive weights computed for 100 random-augmented mappings showed a similar result as the original perceptual means, but introduced very small weights for the left-hip (.03), left-shoulder (.02), and right-shoulder (.09). To further demonstrate that standard least-squares estimation is not sufficient to produce perceptually-valid results, we turn to the illusory movements. Figure 4.c illustrates that the non-expressive least-squares estimation maps the illusory movements to basically the same effort in the three-mode basis. Perceptually, the result is quite different. The illusory movements were perceived to be significantly different (Mann-Whitney U test: U=81, $p < 0.0003$). The results of the learned model more closely resemble the perceptual choices than do the standard least-squares results, even though the illusory data was not used to train the model. This supports our hypothesis that all motion features are not equal during motion recognition.

	Angular Expressive Weights									
	1	2	3	4	5	6	7	8	9	10
(a)	.36	0	0	.15	0	.59	0	0	0	1
(b)	.35	.03	0	.35	0	.61	.02	0	.09	1

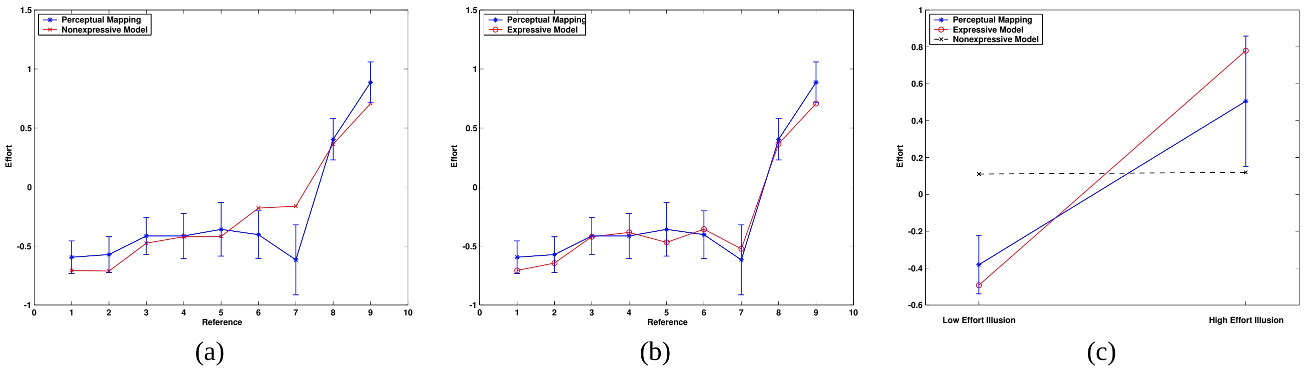
Table 1: Learned expressive weights (normalized) for the carrying examples using (a) perceptual means and (b) perceptual means with random noise trials.

7. Computational Relevance

As with other principal component approaches, this recognition method provides a fast and efficient computation of action effort. Additionally, the approach shows that standard SSE computations do not produce perceptual results. Once the system has been trained, the recognition of movement effort is computed using Eqn. 8. The α_{ij} values and the denominator of this equation can be computed off-line prior to recognition. Since only those feature-i trajectories with high expressibility weights are used to determine the effort, trajectories with low weight can be removed from the computation, reducing the total amount of processed data to only the most expressive trajectories for the category. This has an added advantage in that the approach can therefore accommodate occlusions of non-expressive trajectories. To verify the action category, the distance-from-feature-space (DFFS) residual [19] could be examined using the same three-mode model.

8. Conclusion

We presented an expressive feature model for analyzing and recognizing action effort. Initially, a set of low and high effort examples for an action are factorized into its three-mode principal components. Using a perceptually-based mapping of real to model-generated synthetic movements, a weighted-SSE minimization technique learned the expressive motion trajectories needed for the model to produce ef-



(a) (b) (c)

Figure 4: (a) Perceptual (mean ± 1SD) and least-squares effort estimation for real motions. (b) Perceptual and learned efforts of real motions. (c) Perceptual, learned, and least-squares efforts of illusory movements.

fort labels similar to human judgements.

The approach was demonstrated with carrying examples and two illusory movements to demonstrate the improvement of the three-mode weighted-SSE technique over a non-expressive SSE approach. Future work includes investigating an auto-mapping procedure that can mimic the human matching process, examining additional action categories (walking, running, throwing, lifting, etc.), incorporating natural video input, and modeling the effort regularities across multiple people. Perhaps the approach could be extended to additionally recognize different people.

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: a review. In *Nonrigid and Articulated Motion Workshop*, pages 90–102. IEEE, 1997.

[2] M. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. In *Proc. Comp. Vis. and Pattern Rec.*, pages 561–567, 1997.

[3] M. Brand and A. Hertzmann. Style machines. In *Proc. SIGGRAPH*, pages 183–192. ACM, July 2000.

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. Comp. Vis. and Pattern Rec.*, pages 8–15, 1998.

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

[6] Y. Cui, D. Swets, and J. Weng. Learning-based hand sign recognition using SHOSLIF-M. In *Proc. Int. Conf. Comp. Vis.*, pages 631–636. IEEE, 1995.

[7] J. Davis and V. Kannappan. Expressive features for movement exaggeration. In *Proc. SIGGRAPH*. ACM, 2002.

[8] D. Gavrila. The visual analysis of human movement: a survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[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] T. McMahon. *Muscles, Reflexes, and Locomotion*. Princeton Univ. Press, 1984.

[11] H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *Int. J. of Comp. Vis.*, 14(1):5–24, 1995.

[12] R. Neal, C. Snyder, and P. Kroonenberg. Individual differences and segment interactions in throwing. *Human Movement Sci.*, 10:653–676, 1991.

[13] V. Pavlovic, J. Rehg, T. Cham, and K. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *Proc. Int. Conf. Comp. Vis.*, pages 94–101, 1999.

[14] W. Richards, A. Jepson, and J. Feldman. Priors, preferences and categorical percepts. In *Perception as Baysian Inference*, pages 93–122. Cambridge Univ. Press, 1996.

[15] R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. In *Proc. Comp. Vis. and Pattern Rec.*, pages 721–727. IEEE, 2000.

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

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

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

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

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

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

[22] M. Vasilescu. Human motion signatures for character animation. In *Proc. SIGGRAPH*. ACM, 2001.

[23] S. Watanabe. *Pattern Recognition: Human and Mechanical*. Wiley, New York, 1985.

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