International Journal of Humanoid Robotics
Vol. 4, No. 2 (2007) 365–385

World Scientific
www.worldscientific.com

INTERACTIVE HUMAN POSE AND ACTION RECOGNITION USING DYNAMICAL MOTION PRIMITIVES

ODEST CHADWICKE JENKINS*,‡, GERMÁN GONZÁLEZ SERRANO†,§
and MATTHEW M. LOPER*,¶

*Department of Computer Science, Brown University,
115 Waterman Street, Providence, RI 02912-1910, USA

†Computer Vision Lab, EPFL CH-1015 Lausanne, Switzerland
‡cjenkins@cs.brown.edu
§german.gonzalez@epfl.ch
¶matt@cs.brown.edu

Received 22 September 2006
Revised 21 March 2007

There is currently a division between real-world human performance and the decision making of socially interactive robots. This circumstance is partially due to the difficulty in estimating human cues, such as pose and gesture, from robot sensing. Towards bridging this division, we present a method for kinematic pose estimation and action recognition from monocular robot vision through the use of dynamical human motion vocabularies. Our notion of a motion vocabulary is comprised of movement primitives that structure a human's action space for decision making and predict human movement dynamics. Through prediction, such primitives can be used to both generate motor commands for specific actions and perceive humans performing those actions. In this paper, we focus specifically on the perception of human pose and performed actions using a known vocabulary of primitives. Given image observations over time, each primitive infers pose independently using its expected dynamics in the context of a particle filter. Pose estimates from a set of primitives inferencing in parallel are arbitrated to estimate the action being performed. The efficacy of our approach is demonstrated through interactive-time pose and action recognition over extended motion trials. Results evidence our approach requires small numbers of particles for tracking, is robust to unsegmented multi-action movement, movement speed, camera viewpoint and is able to recover from occlusions.

Keywords: Markerless motion capture; pose estimation; action recognition; motor primitives.

1. Introduction

Perceiving human motion and non-verbal cues is an important aspect of human–robot interaction.[1] For robots to become functional collaborators in society, they must be able to make decisions based on their perception of human state. Additionally, knowledge about human state is crucial for robots to learn control policies from direct observation of humans. Human state, however, encompasses a large and

diverse set of variables, including kinematic, affective, and goal-oriented information, that has proved difficult to model and infer. Part of this problem is that the relationship between such decision-related variables and a robot's sensor readings is difficult to infer directly.

Our greater view is that socially interactive robots will need to maintain estimates, or beliefs as probabilistic distributions, about all of the components in a human's control loop in order to make effective decisions during interaction. Humans make decisions to drive their muscles and affect their environment. A robot can only sense limited information about this control process. This information is often partial observations about the human's kinematics and appearance over time, such as images from a robot's camera. To estimate a human's decision making policy, a robot must attempt to invert this partial information back through its model of the human control loop, maintaining beliefs about kinematic movement, actions performed, decision policy, and intentionality.

As a step in this direction, we present a method for inferring a human's kinematic and action state from monocular vision. Our method works in a bottom-up fashion by using a vocabulary of predictive dynamical primitives, learned from previous work[2] as "action filters" working in parallel. Motion tracking is performed by matching predicted and observed human movement, using particle filtering[3,4] to maintain nonparametric probabilistic beliefs. For quickly performed motion without temporal coherence, we propose a "bending cone" distribution for extended prediction of human pose over larger intervals of time. State estimates from the action filters are then used to infer the linear coefficients for combining behaviors. Inspired by neuroscience, these composition coefficients are related to the human's cognitively planned motion, or "virtual trajectory," providing a compact action space for linking decision making with observed motion.

We present results from evaluating our motion and action tracking system to human motion observed from a single robot camera. Presented results demonstrate our methods ability to track human motion and action robust to performer speed and camera viewpoint with recovery from ambiguous situations, such as occlusion. A primary contribution of our work is interactive-time inference of human pose using sparse numbers of particles with dynamical predictions. We evaluate our prediction mechanism with respect to action classification over various numbers of particles and other prediction related variables. We highlight the application of our tracking results to humanoid imitation.

2. Background

2.1. *Motor primitives and imitation learning*

This work is inspired by the hypotheses from neuroscience pertaining to models of motor control and sensory-motor integration. We ground basic concepts for imitation learning, as described by Matarić,[5] in specific computational mechanisms for humanoids. Matarić's model of imitation consists of: (i) a selective attention

mechanism for extraction of observable features from a sensory stream, (ii) mirror neurons that map sensory observations into a motor repertoire, (iii) a repertoire of motor primitives as a basis for expressing a broad span of movement, and (iv) a classification-based learning system that constructs new motor skills.

Illustrated in Fig. 1, the core of this imitation model is the existence and development of computational mechanisms for mirror neurons and motor primitives. As proposed by Mussa-Ivaldi and Bizzi,[6] motor primitives are used by the central nervous system to solve the inverse dynamics problem in biological motor control. This theory is based on an equilibrium point hypothesis. The dynamics of the plant $D(x, \dot{x}, \ddot{x})$ is a linear combination of forces from a set of primitives, as configuration-dependent force fields (or attractors) $\phi(x, \dot{x}, \ddot{x})$:

$$D(x, \dot{x}, \ddot{x}) = c_i \sum_{i=1}^{K} \phi_i(x, \dot{x}, \ddot{x}), \tag{1}$$

where x is the kinematic configuration of the plant, c is a vector of scalar superposition coefficients, and K is the number of primitives. A specific set of values for c produces stable movement to a particular equilibrium configuration. A sequence

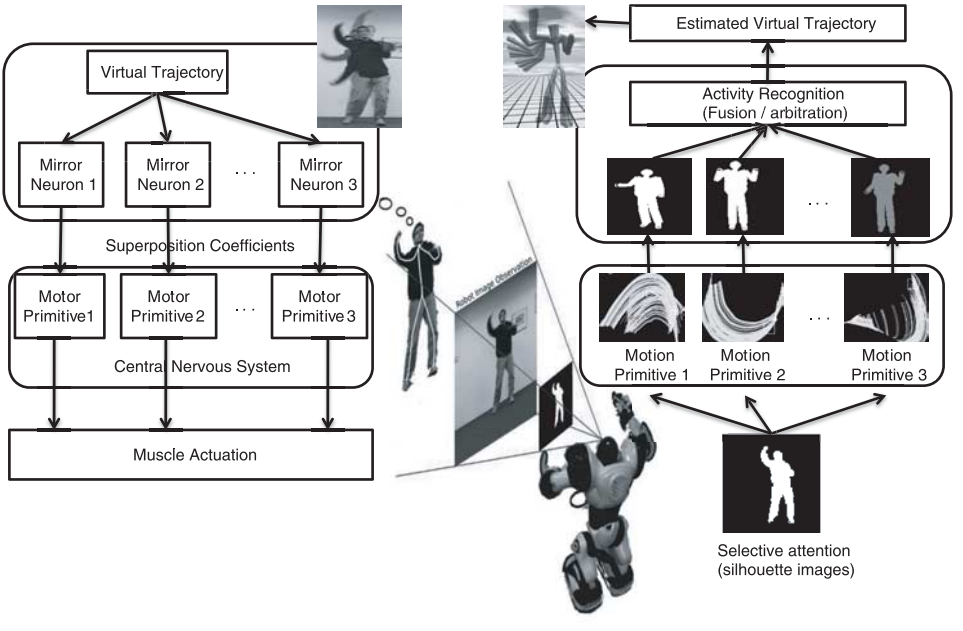


Fig. 1. A "toy" example of our approach to human state estimation and movement imitation. The movement of a human demonstrator assumed to be generated by virtual trajectory executed as a weighted superposition of motor primitives, predictive low-dimensional dynamical systems. For movement imitation, a particle filter for each primitive performs kinematic state (or pose) estimation. Pose estimates across the vocabulary are fused at each timestep and concatenated over time to yield an estimate of the virtual trajectory for the robot to execute.

of equilibrium points specifies a virtual trajectory[7] that can be use for control, as desired motion for internal motor actuation, or perception, to understand the observed movement of an external performer.

Matarić's imitation model assumes the firing of mirror neurons specifies the coefficients for formation of virtual trajectories. Mirror neurons in primates[8] have been demonstrated to fire when a particular activity is executed, observed, or imagined. Assuming 1-to-1 correspondence between primitives and mirror neurons, the scalar firing rate of a given mirror neuron is the superposition coefficient for its associated primitive during equilibrium point control.

2.2. *Motion modeling*

While Matarić's model has desirable properties, there remain several challenges in its computational realization for autonomous robots that we attempt to address. Namely, what are the set of primitives and how are they parameterized? How do mirror neurons recognize motion indicative of a particular primitive? What computational operators should be used to compose primitives to express a broader span of motion?

Our previous work[2] addresses these computational issues through the unsupervised learning of motion vocabularies, which we now utilize within probabilistic inference. Our approach is close in spirit to work by Kojo *et al.*,[10] who define a "proto-symbol" space describing the space of possible motion. Monocular human tracking is then cast as localizing the appropriate action in the proto-symbol space describing the observed motion using divergence metrics. Schaal *et al.*[11] encode each primitive to describe the nonlinear dynamics of a specific trajectory with a discrete or rhythmic pattern generator. New trajectories are formed by learning superposition coefficients through reinforcement learning. While this approach to primitive-based control may be more biologically faithful, our method provides greater motion variability within each primitive and facilitates partially observed movement perception (such as monocular tracking) as well as control applications. Work proposed by Bentivegna *et al.*[12] and Grupen *et al.*[13,14] approach robot control through sequencing and/or superposition of manually crafted behaviors.

Recent efforts by Knoop *et al.*[15] perform monocular kinematic tracking using iterative closest point and the latest Swissranger depth sensing devices, capable of precise depth measurements. We have chosen instead to use the more ubiquitous passive camera devices and also avoid modeling detailed human geometry.

Many other approaches to data-driven motion modeling have been proposed in computer vision, animation, and robotics. The reader is referred to Refs. 2, 16 and 17 for broader coverage of these methods.

2.3. *Monocular tracking*

We pay particular attention to methods using motion models for kinematic tracking and action recognition in interactive-time. Particle filtering[3,4] is a well-established

means for inferring kinematic pose from image observations. Yet, particle filtering often requires additional (often overly expensive) procedures, such as annealing,[18] nonparametric belief propagation,[19,20] Gaussian process latent variable models,[16] POMDP learning[21] or dynamic programing,[22] to account for the high dimensionality and local extrema of kinematic joint angle space. These methods tradeoff real-time performance for greater inference accuracy. This speed-accuracy contrast is most notably seen in how we use our learned motion primitives[2] as compared to Gaussian process methods.[16,23] Both approaches use motion capture as probabilistic priors on pose and dynamics. However, our method emphasizes temporally extended prediction to use fewer particles and enable fast inference, whereas Gaussian process models aim for accuracy through optimization. Further, unlike the single-action motion-sparse experiments with Gaussian process models, our work is capable of inference of multiple actions, where each action has dense collections of motion. Such actions are generalized versions of the original training motion and allow us to track new variations of similar actions.

Similar to Ref. 24, our method aims for interactive-time inference on actions to enable incorporation into a robot control loop. Unlike Ref. 24, however, we focus on recognizing active motions, rather than static poses, robust to occlusion by developing fast action prediction procedures that enable online probabilistic inference. We also strive for robustness to motion speed by enabling extended look-ahead motion predictions using a "bending cone" distribution for dynamics. Yang *et al.*[25] define a discrete version with similar dynamics using Hidden Markov Model and vector quantization observations. However, such HMM-based transition dynamics are instantaneous with a limited prediction horizon, whereas our bending cone allows for further look-ahead with a soft probability distribution.

3. Dynamical Kinematic and Action Tracking

Kinematic tracking from silhouettes is performed via the steps in Fig. 2, those are: (i) global localization of the human in the image, (ii) primitive-based kinematic pose estimation and (iii) action recognition. The human localization is kept as an

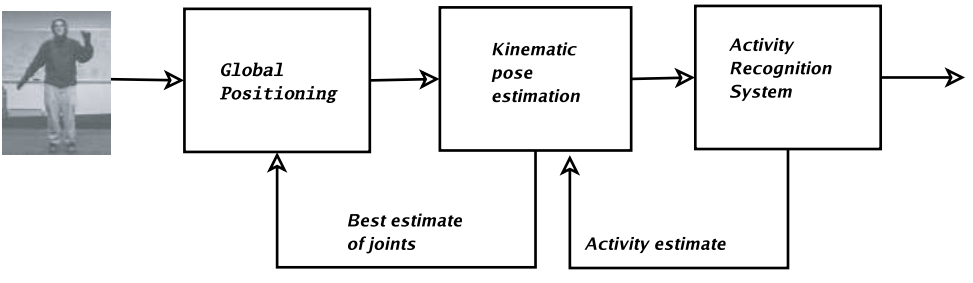


Fig. 2. Illustration of the three stages in our approach to tracking: image observations are used to localize the person in 3D, then infer kinematic pose, and finally estimate of activity/action. Estimates at each stage are used to form priors for the previous stage at the next timestep.

unimodal distribution and estimated using the joint angle configuration derived in the previous time step.

3.1. *Dynamical motion vocabularies*

The methodology of Ref. 2 is followed for learning dynamical vocabularies from human motion. We cover relevant details from this work and refer the reader to the citation for details. Motion capture data representative of natural human performance is used as input for the system. The data is partitioned into an ordered set of non-overlapping segments representative of "atomic" movements. Spatio-temporal Isomap[9] embed these motion trajectories into a lower dimensional space, establishing a separable clustering of movements into activities. Similar to Ref. 26, each cluster is a group of motion examples that can be interpolated to produce new motion representative of the underlying action. Each cluster is speculatively evaluated to produce a dense collection of examples for each uncovered action. A primitive B_i is the manifold formed by the dense collections of poses X_i (and associated gradients) in joint angle space resulting from this interpolation.

We define each primitive B_i as a gradient (potential) field expressing the expected kinematic behavior over time of the ith action. In the context of dynamical systems, this gradient field $B_i(x)$ defines the predicted direction of displacement for a location in joint angle space $\hat{x}[t]$ at time t[a]:

$$\begin{aligned}\hat{x}_i[t+1] &= f_i(x[t], u[t]) \\ &= u[t]B_i(x) = u[t]\frac{\sum_{x\in \texttt{nbhd}(x[t])} w_x \Delta_x}{\| \sum_{x\in \texttt{nbhd}(x[t])} w_x \Delta_x \|},\end{aligned} \tag{2}$$

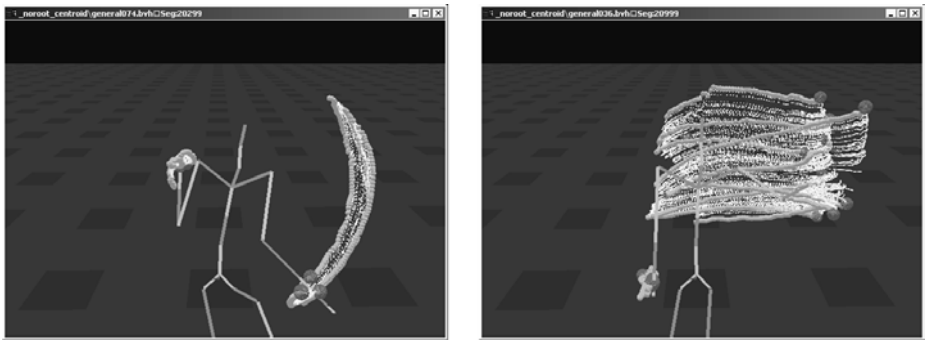
where $u[t]$ is a fixed displacement magnitude, Δ_x is the gradient of pose x[b] a motion example of primitive i, and w_x the weight[c] of x with respect to $x[t]$. Figure 3 shows examples of learned predictive primitives.

Given results in motion latent space dimensionality,[2,16] we construct a low dimensional latent space to provide parsimonious observables y_i of the joint angle space for primitive i. This latent space is constructed by applying Principal Components Analysis (PCA) to all of the poses X_i comprising primitive i and form the output equation of the dynamical system, such as in Ref. 27:

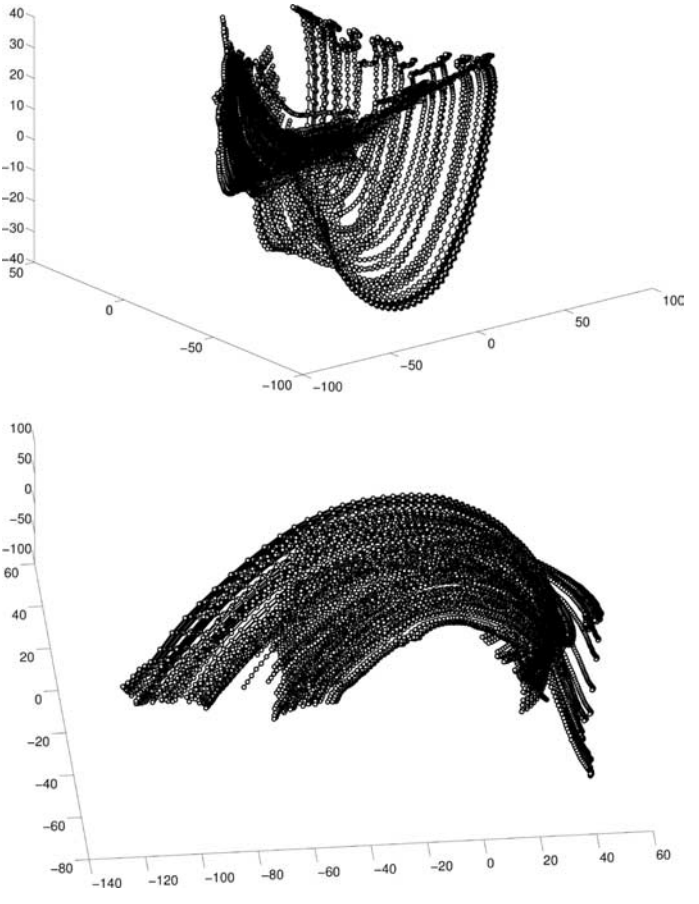
$$y_i[t] = g_i(x[t]) = A_i x[t], \tag{3}$$

where g_i is the latent space transformation and A_i is the expression of g_i as an affine transformation into the principal component space of primitive i.[d] Although

[a] $\texttt{nbhd}(\cdot)$ is used to identify the k-nearest neighbors in an arbitrary coordinate space, which we use both in joint angle space and the space of motion segments.
[b] The gradient is computed as the direction between y and its subsequent pose along its motion example.
[c] Typically, reciprocated Euclidean distance.
[d] $x[t]$ and $y_i[t]$ are assumed to be homogeneous in Eq. (3).

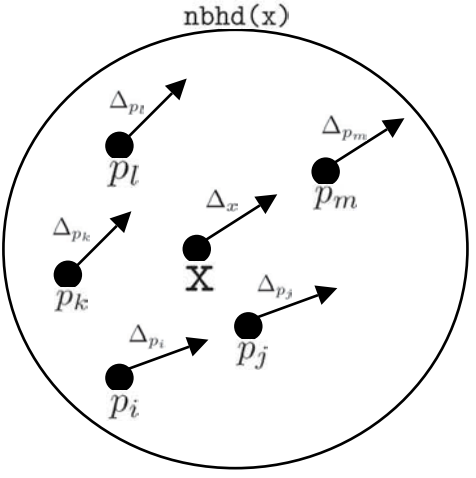


(a)



(b)

Fig. 3. (a) Kinematic endpoint trajectories for learned primitive manifolds, (b) corresponding joint angle space primitive manifolds (view from first three principal components), and (c) an instantaneous prediction example (illustrated as a zoomed-in view on a primitive manifold).



(c)

Fig. 3. (*Continued*)

other dimension reduction methods could provide greater parsimony, we chose a linear transform for g_i for inversion simplicity and evaluation speed. For each of our primitives, at least 95% of the variance of the pose manifold is preserved in this transformation, making A_i a reasonable approximation for the joint space manifold.

Given the preservation of variance in A_i, it is assumed that latent space dynamics, governed by $\tilde{f}_i$, can be computed in the same manner as f in joint angle space:

$$\frac{g_i^{-1}(\tilde{f}_i(g_i(x[t]), u[t])) - x[t]}{\|g_i^{-1}(\tilde{f}_i(g_i(x[t]), u[t])) - x[t]\|} \approx \frac{f_i(x[t], u[t]) - x[t]}{\|f_i(x[t], u[t]) - x[t]\|}. \tag{4}$$

3.2. *Kinematic pose estimation*

Kinematic tracking is performed by particle filtering[3,4] in the individual latent spaces created for each primitive in a motion vocabulary. We infer with each primitive individually and in parallel to avoid high-dimensional state spaces, encountered in Ref. 18. A particle filter of the following form is instantiated in the latent space of each primitive:

$$\begin{aligned} p(y_i[1:t] \mid z_i[1:t]) &\propto p(z[t] \mid g_i^{-1}(y_i[t])) \\ &\times \sum_{y_i} p(y_i[t] \mid y_i[t-1]) p(y_i[1:t-1] \mid z[1:t-1]), \end{aligned} \tag{5}$$

where $z_i[t]$ are the observed sensory features at time t and g_i^{-1} is the transformation into joint angle space from the latent space of primitive i.

The likelihood function $p(z[t] \mid g_i^{-1}(y_i[t]))$ can be any reasonable choice for comparing the hypothesized observations from a latent space particle and the sensor

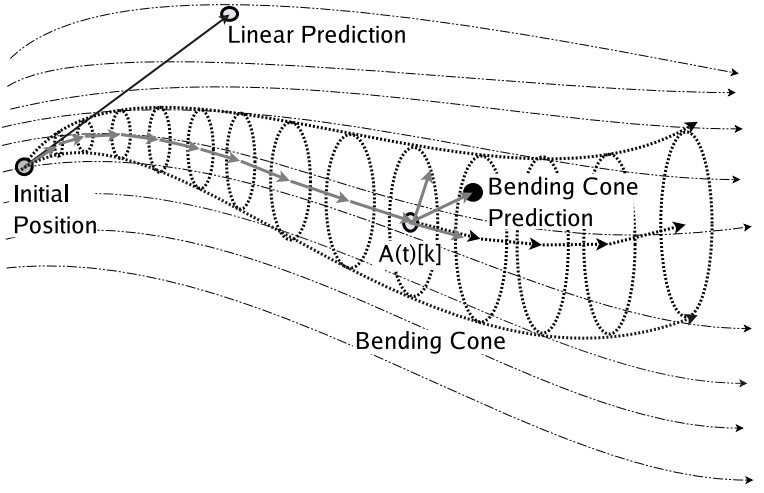


Fig. 4. Illustration of the predictive bending cone distribution. The thin dashed black lines indicate the flow of a primitive's gradient field. Linear prediction from the current pose $y_i(t)$ will lead to divergence from the gradient field as the prediction magnitude increases. Instead, we use a bending cone (in bold) to provide an extended prediction horizon along the gradient field. Sampling a pose prediction $y_i(t+1)$ occurs by selecting a cross-section $A(t)[k]$ and adding cylindrical noise.

observations. Ideally, this function will be monotonic with discrepancy in the joint angle space.

At first glance, the motion distribution $p(y_i[t] \,|\, y_i[t-1])$ could be given by the instantaneous "flow," as proposed by Ong *et al.*,[28] where a locally linear displacement with some noise is expected. However, such an assumption would require temporal coherence between the training set and the performance of the actor. Observations without temporal coherence cannot simply be accounted for by extending the magnitude of the displacement vector because the expected motion will likely vary in a nonlinear fashion over time. To address this issue, a "bending cone" distribution is used (Fig. 4) over the motion model. This distribution is formed with the structure of a generalized cylinder with a curved axis along the motion manifold and a variance cross-section that expands over time. The axis is derived from K successive predictions $\hat{y}_i[t]$ of the primitive from a current hypothesis $y_i[t]$ as a piecewise linear curve. The cross-section is modeled as cylindrical noise $\mathcal{C}(a, b, \sigma)$ with local axis a–b and normally distributed variance σ orthogonal to the axis.

The resulting parametric distribution, Eq. (6), is sampled by randomly selecting a step-ahead k and generating a random sample within its cylinder cross-section. Note that $f(k)$ is some monotonically increasing function of the distance from the cone origin; we used a linear function:

$$p(y_i[t] \mid y_i[t-1]) = \sum_{\hat{y}_i[t]}^{k} \mathcal{C}(\hat{y}_i[k+1], \hat{y}_i[k], f(k)). \tag{6}$$

3.3. *Action recognition*

For action recognition, a probability distribution across primitives of the vocabulary is created.[e] The likelihood of the pose estimate from each primitive is normalized into a probability distribution:

$$p(B_i[t] \mid z[t]) = \frac{p(z[t] \mid \bar{x_i}[t])}{\sum_B p(z[t] \mid \bar{x_i}[t])}, \tag{7}$$

where $\bar{x_i}[t]$ is the pose estimate for primitive i. The primitive with the maximum probability is estimated as the action currently being performed. Temporal information can be used to improve this recognition mechanism by fully leveraging the latent space dynamics over time.

The manifold in latent space is essentially an attractor along a family of trajectories towards an equilibrium region. We consider *attractor progress* as a value that increases as kinematic state progresses towards a primitive's equilibrium. For an action being performed, we expect its attractor progress will monotonically increase as the action is executed. The attractor progress can be used as a feedback signal into the particle filters estimating pose for a primitive i in a form such as:

$$p(B_i[t] \mid z[t]) = \frac{p(z[t] \mid \bar{x_i}[t], w_i[1:t-1])}{\sum_B p(z[t] \mid \bar{x_i}[t], w_i[1:t-1])}, \tag{8}$$

where $w_i[1:t-1]$ is the probability that primitive B_i has been performed over time.

4. Results

For our experiments, we developed an interactive-time software system in C++ that tracks human motion and action from monocular silhouettes using a vocabulary of learned motion primitives. Shown in Fig. 5, our system takes video input from a Fire-i webcam (15 frames per second, at a resolution of 120×160) mounted on an iRobot Roomba Discovery. Image silhouettes were computed with standard background modeling techniques for pixel statistics on color images. Median and morphological filtering were used to remove noisy silhouette pixels. An implementation of spatio-temporal Isomap[2] was used to learn motion primitives for performing punching, hand circles, vertical hand waving, and horizontal hand waving.

We utilize a basic likelihood function, $p(z[t] \,|\, g_i^{-1}(y_i[t]))$, that returns the similarity $R(A, B)$ of a particle's hypothesized silhouette with the observed silhouette image. Silhouette hypotheses were rendered from a cylindrical 3D body model to an binary image buffer using OpenGL. A similarity metric, $R(A, B)$ for two silhouettes A and B, closely related to the inverse of the Generalized Hausdorff

[e]We assume each primitive describes an action of interest.

Fig. 5. Robot platform and camera used in our experiments.

distance was used:

$$R(A,B) = \frac{1}{r(A,B) + r(B,A) + \epsilon}, \tag{9}$$

$$r(A,B) = \sum_{a \in A} \left(\min_{b \in B} \|a - b\| \right)^2. \tag{10}$$

This measure is an intermediate between undirected and generalized Hausdorff distance. ϵ is used only to avoid divide-by-zero errors. An example Hausdorff map for a human silhouette is shown in Fig. 4. Due to this silhouetting procedure, the robot must be stationary (i.e. driven to a specific location) during the construction of the background model and tracking process. As we are exploring in future work, this limitation could be relaxed through the use of other sensor modalities, such as stereo vision or time-of-flight ranging cameras.

To enable fast monocular tracking, we applied our system with sparse distributions (six particles per primitive) to three trial silhouette sequences. Each trial is designed to provide insight into different aspects of the performance of our tracking system.

In the first trial (termed multi-action), the actor performs multiple repetitions of three actions (hand circles, vertical hand waving, and horizontal hand waving) in sequence. As shown in Fig. 7, reasonable tracking estimates can be generated from as few as six particles. As expected, we observed that the Euclidean distance between our estimates and the ground truth decreases with the number of particles used in the simulation, highlighting the tradeoff between the number of particles and accuracy of the estimation.

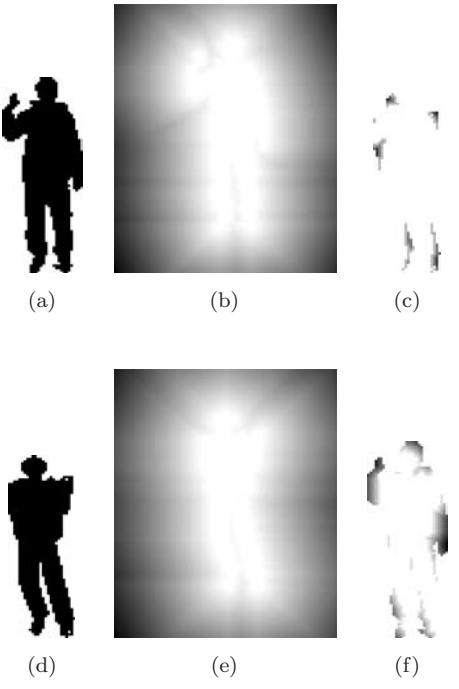


Fig. 6. Likelihood function used in the system: (a) is the silhouette A extracted from the camera and (d) is the synthetic silhouette B generated from a pose hypothesis. (b) and (e) are the respective Hausdorff distance transforms, showing pixels with larger distances from the silhouette as dark. (c) and (f) illustrate the sums $r(A, B)$, how silhouette A relates B, and $r(B, A)$, silhouette B relates to A. These sums are added and reciprocated to assess the similarity of A and B.

To explore the effects of the number of particles, we ran our tracking system on the multi-action trial using powers-of-two number particles between 1 and 1,024 for each action. The bending cone for these trials are generated using 20 predictions into the future and the noise aperture is $\pi/6$, which increases in steps of $20\pi/6$ per prediction. Shown in Fig. 8, the action classification results from the system for each trial were plotted in the ROC plane for each action. The ROC plane plots each trial (shown as a labeled dot) in 2D coordinates where the horizontal axis is the "false positive rate," percentage of frames incorrectly labeled as a given action, and the vertical axis is the "true positive rate," percentage of correctly labeled frames. Visually, plots in the upper left-hand corner are indicative of good performance, points in the lower right-hand corner indicate bad performance, and points along the diagonal indicate random performance.

The ROC plots for the numbers of particles indicate the classifier works better with more particles, but not always. Trials using 1,024 particles per action is never the closest point to the upper-left corner. The most noticeable benefit to having more particles is when occlusion was present. In particular, the Circle action does not introduce occlusion when performed in profile, where as the Horizontal Waving

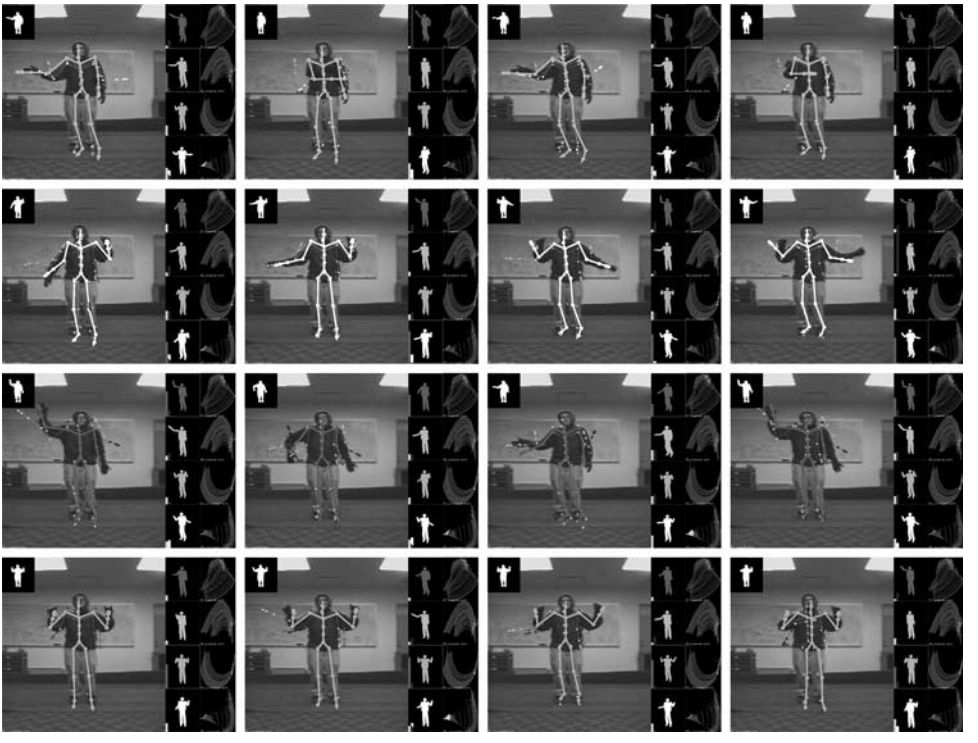


Fig. 7. Tracking of motion sequence containing three distinct actions performed in sequence without stopping. Each row shows the recognition of individual actions for waving a hand across the body (top row), bottom-to-top in a circular fashion (second and fourth row) and top-to-bottom (third row). The kinematic estimates are shown with a thick-lined stick figure; the color of the stick figures represents the action recognized. Each image contains a visualization of the dynamical systems and pose estimates for each action.

motion does require occlusion in its performance. Consequently, the Circle trials all plot near the ROC upper-left corner and the Horizontal Waving ROC trials fork biomodally in classification performance.

ROC plots were also generated for varying bending cone prediction length and noise aperture, also shown in Fig. 8. Plotted in the middle row, the variations in bending cone length were varied between 10 and 100 predictions in increments of 10, with an additional trial using three predictions. For these trials, the number of particles was fixed to 64 and the noise aperture to Pi/6. Plotted in the bottom row, the variation in the bending cone noise aperture were varied between 0 and $\pi/2$ in increments of $0.1 \times \pi/2$. The number of particles and bending cone length were fixed at 64 and 20, respectively. These plots indicate variations similar to those in the numbers of particles plots. Specifically, good performance results regardless of the bending cone parameters when the action has little or no occlusion ambiguity, but performance drops off when such ambiguity is present. However, in these

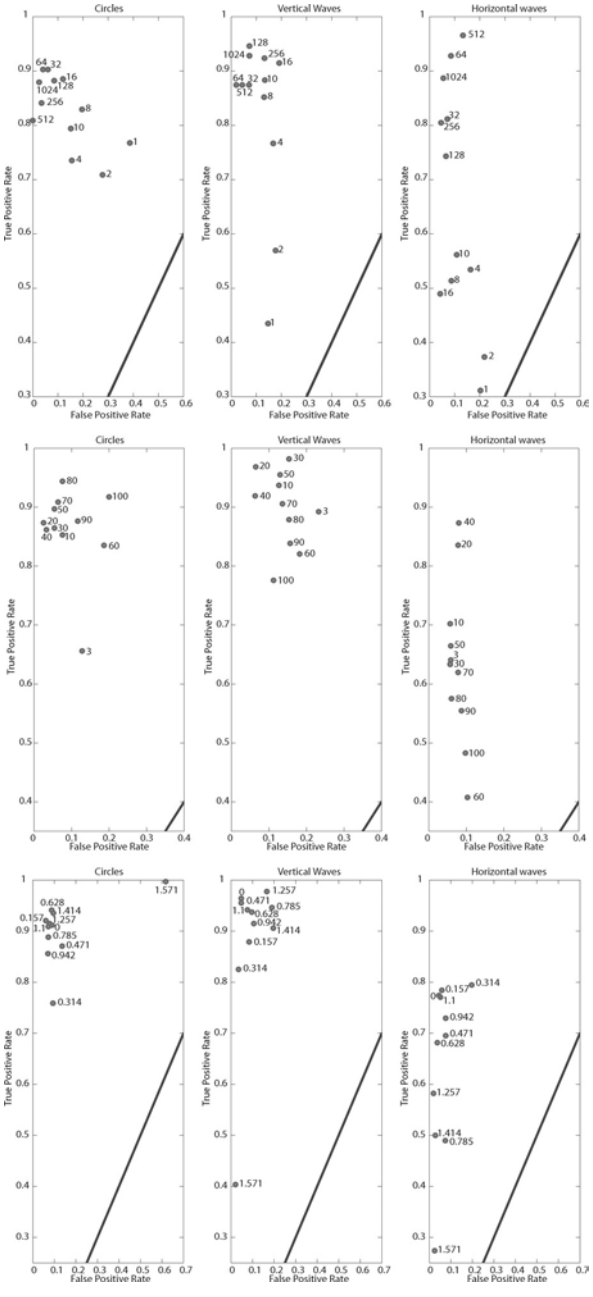


Fig. 8. ROC plots of action classification on the multi-action sequence. Columns breakdown by the action performed: Circle action (left), Vertical Waving action (center), and Horizontal Waving action (right). Columns show the effect of varied numbers of particles (top, varied between 1 and 1024), bending cone prediction length (middle, varied between 3 and 100), and bending cone noise aperture (bottom, varied between 0 and $\pi/2$). Each plot shows for each trial (plotted as a labeled point) the false positive rate (horizontal axis) and true positive rate (vertical axis). Note the difference in scales among rows.

trials, increased prediction length and noise aperture does not necessarily improve performance. It is surprising that with zero aperture (that is, staying fixed in the manifold), the classifier does not perform that badly. Instead, there are sweet spots between 20–40 predictions and under 0.15π noise aperture for the bending cone parameters. Although including more particles is always increase accuracy, we have

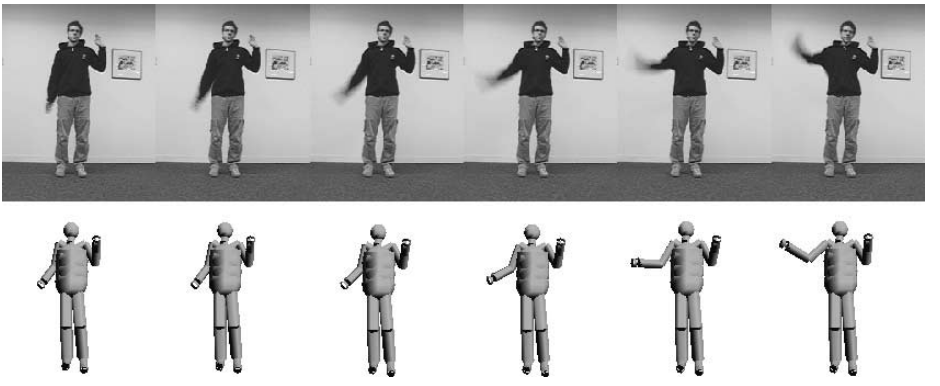


Fig. 9. Tracking of a fast waving motion. Observed images (top) and pose estimates from the camera view (bottom).

Fig. 10. Illustrations of a demonstrated fast moving "punch" movement (left) and the estimated virtual trajectory (right) as traversed by our physically simulated humanoid simulation.

not yet explored how the sweet spots in the bending cone parameters could change as the numbers of particles vary.

In trial two (fast-wave motion), we analyzed the temporal robustness of the tracking system. The same action is performed at different speeds, ranging from slow (hand moving at $\approx 3\,\mathrm{cm/s}$) to fast motion (hand moving at $\approx 6\,\mathrm{m/s}$). The fast motion is accurately predicted as seen in Fig. 9. Additionally, we were able to track a fast moving punching motion (Fig. 10) and successfully execute the motion with our physics-based humanoid simulation. Our simulation system is described in Ref. 29.

In trial three (overhead-view), viewpoint invariance was tested with video from a trial with an overhead camera, shown in Fig. 11. Even given limited cues from the silhouette, we are able to infer the horizontal waving of an arm. Notice that the arm estimates are consistent throughout the sequence.

Using the above test trials, we measured the ability of our system to recognize performed actions to provide responses similar to mirror neurons. In our current system, an action is recognized as the pose estimate likelihoods normalized over

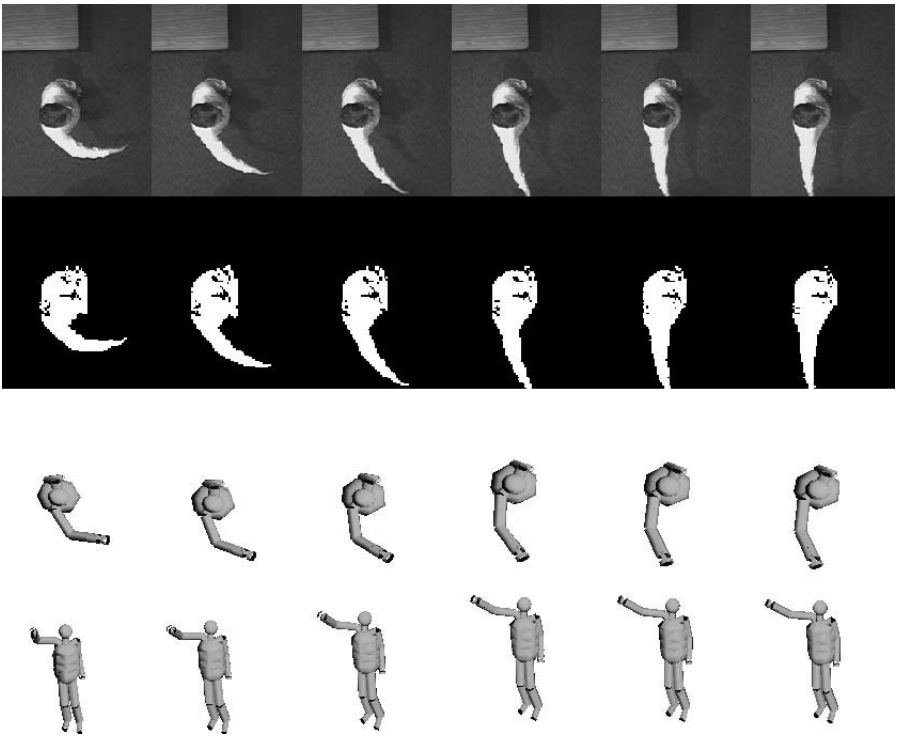


Fig. 11. A sequence of pose estimates for a reaching motion. Observed silhouettes (second from top) can be compared with our pose estimates from the camera view (second from bottom) and from overhead (bottom).

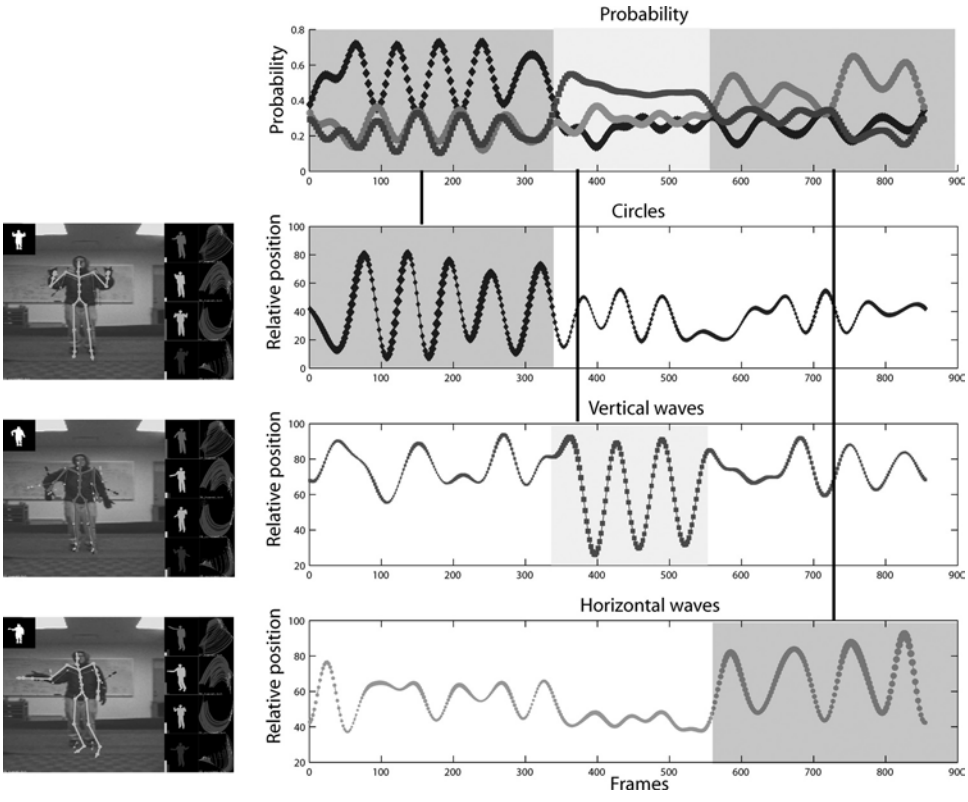


Fig. 12. An evaluation of our action recognition system over time with a 3-action motion performing "hand circles," "horizontal waving," and "vertical waving" in sequence. The first row reflects the relative likelihood (idealized as mirror neuron firing) for each primitive with background sections indicating the boundary of each action. Each of the subsequents rows shows time on the x-axis, attractor progress on the y-axis, and the width of the plot marker indicates the likelihood of the pose estimate.

all of the primitives into a probability distribution, as shown in Fig. 12. Temporal information can be used to improve this recognition mechanism by fully leveraging the latent space dynamics over time. The manifold in latent space is essentially an attractor along a family of trajectories. A better estimator of action would consider *attractor progress*, monotonic progress towards to equilibrium region of an action's gradient field. We have analyzed preliminary results from observing attractor progress in our trials, as shown in Fig. 12. For an action being performed, its attractor progress is monotonically increasing. If the action is performed repeatedly, we can see a periodic signal emerge, as opposed to the noisier signals of the action not being performed. These results indicate that we can use attractor progress as a feedback signal to further improve an individual primitive's tracking performance.

Because of their attractor progress properties, we believe that we can analogize these action patterns into the firing of an idealized mirror neurons. The firing of our

artificial mirror neurons provide superposition coefficients, as in Ref. 30. Given real-time pose estimation, online movement imitation could be performed by directly executing the robot's motor primitives weighted by these coefficients. Additionally, these superposition coefficients could serve as input into additional inference systems to estimate the human's emotional state for providing an affective robot response.

In our current system, we use the action firing to arbitrate between pose estimates for forming a virtual trajectory. While this is a simplification of the overall goal, our positive results for trajectory estimation demonstrate our approach is viable and has promise for achieving our greater objectives. As future work, we will extend the motion dynamics of the vocabulary into basis behaviors using our complementary work in learning behavior fusion.[30]

5. Conclusion

We have presented a neuro-inspired method for monocular tracking and action recognition for movement imitation. Our approach combines vocabularies of kinematic motion learned offline with online estimation of a demonstrator's underlying virtual trajectory. A modular approach to pose estimation is taken for computational tractability and emulation of structures hypothesized in neuroscience. Our current results suggest our method can perform tracking and recognition from partial observations at interactive rates. Our current system demonstrates robustness with respect to the viewpoint of the camera, the speed of performance of the action, and recovery from ambiguous situations.

Acknowledgments

This research was supported by grants from the Office of Naval Research (Award N000140710141) and National Science Foundation (Award IIS-0534858). The authors are grateful to Chris Jones and iRobot Corporation for support with the Roomba platform, Brian Gerkey the Player/Stage Project, and RoboDynamics Corporation for Roomba interfaces, and Rod Beresford.

References

1. T. Fong, I. Nourkbaksh and K. Daoutenhahn, A survey of socially interactive robots: Concepts, design and applications, Carnegie Mellon University Robotics Institute, Pittsburgh, PA, Technical Report CMU-RI-TR02-29 (November 2002).
2. O. C. Jenkins and M. J. Matarić, Performance-derived behavior vocabularies: Data-driven acqusition of skills from motion, *Int. J. Humanoid Robot.* **1**(2) (2004) 237–288.
3. M. Isard and A. Blake, Condensation — Conditional density propagation for visual tracking, *Int. J. Comput. Vision* **29**(1) (1998) 5–28.
4. S. Thrun, W. Burgard and D. Fox, *Probabilistic Robotics* (MIT Press, Cambridge, MA, 2005).
5. M. J. Matarić, Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics, in *Imitation in Animals and Artifacts*, eds. C. Nehaniv and K. Dautenhahn (MIT Press, 2002), pp. 392–422.

6. F. Mussa-Ivaldi and E. Bizzi, Motor learning through the combination of primitives, *Phil. Trans. R. Soc. Lond. B* **355** (2000) 1755–1769.
7. N. Hogan, The mechanics of posture and movement, *Biol. Cybernet.* **52** (1985) 315–331.
8. G. Rizzolatti, L. Gadiga, V. Gallese and L. Fogassi, Premotor cortex and the recognition of motor actions, *Cogn. Brain Res.* **3** (1996) 131–141.
9. O. C. Jenkins and M. J. Matarić, A spatio-temporal extension to isomap nonlinear dimension reduction, in *Int. Conf. Machine Learning* (*ICML*), Banff, Alberta, Canada, (ACM Press, 2004), p. 56.
10. N. Kojo, T. Inamura, K. Okada and M. Inaba, Gesture recognition for humanoids using proto-symbol space, in *IEEE Int. Conf. Humanoid Robots* (*Humanoids 2006*), Detroit, USA (IEEE Press, 2006), pp. 76–81.
11. A. J. Ijspeert, J. Nakanishi and S. Schaal, Trajectory formation for imitation with nonlinear dynamical systems, in *IEEE Intelligent Robots and Systems* (*IROS 2001*), Maui, Hawaii, USA (IEEE Press, 2001), pp. 752–757.
12. D. C. Bentivegna and C. G. Atkeson, Learning from observation using primitives, in *IEEE Int. Conf. Robotics and Automation*, Seoul, Korea (IEEE Press, 2001), pp. 1988–1993.
13. R. A. Grupen, M. Huber, J. A. Coehlo Jr. and K. Souccar, A basis for distributed control of manipulation tasks, *IEEE Expert* **10**(2) (1995) 9–14.
14. R. Platt, A. H. Fagg and R. R. Grupen, Manipulation gaits: Sequences of grasp control tasks, in *IEEE Conf. Robotics and Automation*, New Orleans, LA, USA (IEEE Press, 2004), pp. 801–806.
15. S. Knoop, S. Vacek and R. Dillmann, Sensor fusion for 3D human body tracking with an articulated 3D body model, in *IEEE Int. Conf. Robotics and Automation*, Orlando, Florida, USA (IEEE Press, 2006), pp. 1686–1691.
16. R. Urtasun, D. J. Fleet, A. Hertzmann and P. Fua, Priors for people tracking from small training sets, in *Int. Conf. Computer Vision* (*ICCV*), Vol. 1, Beijing, China (IEEE Press, 2005), pp. 403–410.
17. L. Kovar and M. Gleicher, Automated extraction and parameterization of motions in large data sets, *ACM Trans. Graph.* **23**(3) (2004) 559–568.
18. J. Deutscher, A. Blake and I. Reid, Articulated body motion capture by annealed particle filtering, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Vol. 2, Hilton Head, SC, USA (IEEE Press, 2000), pp. 126–133.
19. L. Sigal, S. Bhatia, S. Roth, M. J. Black and M. Isard, Tracking loose-limbed people, in *Computer Vision and Pattern Recognition* (*CVPR*), Vol. 1, Washington D. C., USA (IEEE Press, 2004), pp. 421–428.
20. E. B. Sudderth, A. T. Ihler, W. T. Freeman and A. S. Willsky, Nonparametric belief propagation, in *CVPR*, Vol. 1, Madison, Wisconsin, USA (IEEE Computer Society, 2003), pp. 605–612.
21. T. Darrell and A. Pentland, Active gesture recognition using learned visual attention, in *Advances in Neural Information Processing Systems*, eds. D. S. Touretzky, M. C. Mozer and M. E. Hasselmo, Vol. 8 (MIT Press, 1996), pp. 858–864.
22. D. Ramanan and D. A. Forsyth, Automatic annotation of everyday movements, in *Neural Information Processing Systems* (*NIPS*), Vancouver and Whistler, Canada (MIT Press, 2003), pp. 1547–1554.
23. J. Wang, D. J. Fleet and A. Hertzmann, Gaussian process dynamical models, in *Neural Information Processing Systems* (*NIPS*), Vancouver and Whistler, Canada (MIT Press, 2003), pp. 1441–1448.

24. E. Huber and D. Kortenkamp, A behavior-based approach to active stereo vision for mobile robots, *Eng. Appl. Artif. Intell.* **11** (1998) 229–243.
25. H. D. Yang, A. Y. Park and S. W. Lee, Gesture spotting and recognition for human–robot interaction, *IEEE Trans. Robot.* **23** (2007) 256–270.
26. C. Rose, M. F. Cohen and B. Bodenheimer, Verbs and adverbs: Multidimensional motion interpolation, *IEEE Comput. Graph. Appl.* **18**(5) (1998) 32–40.
27. N. R. Howe, M. E. Leventon and W. T. Freeman, Bayesian reconstruction of 3D human motion from single-camera video, *Advances in Neural Information Processing Systems* (*NIPS*), Vancouver and Whistler, Canada (MIT Press, 2000).
28. E. Ong, A. Hilton and A. Micilotta, Viewpoint invariant exemplar-based 3D human tracking, *Comput. Vision Image Understand.* **104**(2) (2006) 178–189.
29. P. Wrotek, O. Jenkins and M. McGuire, Dynamo: Dynamic data-driven character control with adjustable balance, in *Proc. Sandbox Symp. Video Games*, Boston, MA, USA (ACM Press, 2006).
30. M. Nicolescu, O. Jenkins and A. Olenderski, Learning behavior fusion estimation from demonstration, in *IEEE Int. Symp. Robot and Human Interactive Communication (RO-MAN 2006)*, Hatfield, United Kingdom (IEEE Press, 2006), pp. 340–345.

Odest Chadwicke Jenkins has been an Assistant Professor of Computer Science at Brown University since 2004. Prof. Jenkins earned his Ph.D. in Computer Science at the University of Southern California (2003) at the Center for Robotics and Embedded Systems, M.S. in Computer Science at Georgia Tech (1998) and B.S. in Computer Science and Mathematics at Alma College (1996). Prof. Jenkins' research interests pertain primarily to human–robot interaction and robot learning and also include humanoid robotics, machine learning, computer vision, computer animation, motion capture systems, autonomous agents and interactive entertainment. His dissertation concerned model-free methods for motion capture and analysis of kinematic motion for modeling perceptual-motor primitives.

Germán González Serrano was born in Madrid in 1982. He received his M.S. degree in Interactive Systems Engineering from the Kungliga Tekniska Högskolan (KTH), Sweden and his engineering degree in Telecommunications Engineering from the Universidad Politécnica de Madrid, Spain, both in 2006. From 2005 to 2006, he was at Brown University as a research assistant. He is currently pursuing his Ph.D. at the Ecole Polytechnique Fédérale de Lausanne, in Switzerland. His research interests range from image processing to human–robot interaction through the use of natural language. In 2006, he received the best Swedish master thesis in artificial intelligence award, given by the Swedish Artificial Intelligence Society.

Matthew Maverick Loper is currently pursuing a Ph.D. in Computer Science at Brown University. He earned his B.A. in Economics and Psychology at Connecticut College in 1996. His interests include markerless human pose recognition, gesture recognition, reconstruction of geometry from images, camera calibration, video mosaics and texture synthesis.