Tracking Human Motion and Actions for Interactive Robots

*

Odest Chadwicke Jenkins
Dept. of Computer Science
Brown University
Providence, RI, USA
cjenkins@cs.brown.edu

German González
Computer Vision Lab
EPFL
Lausanne, Switzerland
gerg@cs.brown.edu

Matthew Maverick Loper
Dept. of Computer Science
Brown University
Providence, RI, USA
matt@cs.brown.edu

ABSTRACT

A method is presented for kinematic pose estimation and action recognition from monocular robot vision through the use of dynamical human motion vocabularies. We propose the utilization of dynamical motion vocabularies towards bridging the decision making of observed humans and information from robot sensing. Our motion vocabulary is comprised of learned primitives that structure the action space for decision making and describe human movement dynamics. Given image observations over time, each primitive infers on pose independently using its prediction density on movement 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 tracking and action recognition over extended motion trials. Results evidence the robustness of the algorithm with respect to unsegmented multi-action movement, movement speed, and camera viewpoint.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Measurement

Keywords

Human Tracking, Markerless Motion Capture, Action Recognition, Human-Robot Interaction

*(Produces the permission block, and copyright information). For use with SIG-ALTERNATE.CLS. Supported by ACM.

HRI'07, March 10–12, 2007, Arlington, Virginia, USA.

1. INTRODUCTION

Perceiving human motion and non-verbal cues is an important aspect of human-robot interaction. 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 beliefs 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 infer on human decision making, 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 [10] as "action filters" working in parallel. Motion tracking is performed by matching predicted and observed human movement, using particle filtering [9, 25] to maintain probabilistic beliefs. For quickly performed motion without temporal coherence, we propose a "bending cone" distribution for predicting far ahead in 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 the ability of our system to track human motion and action robust to performer speed and camera viewpoint with the ability to

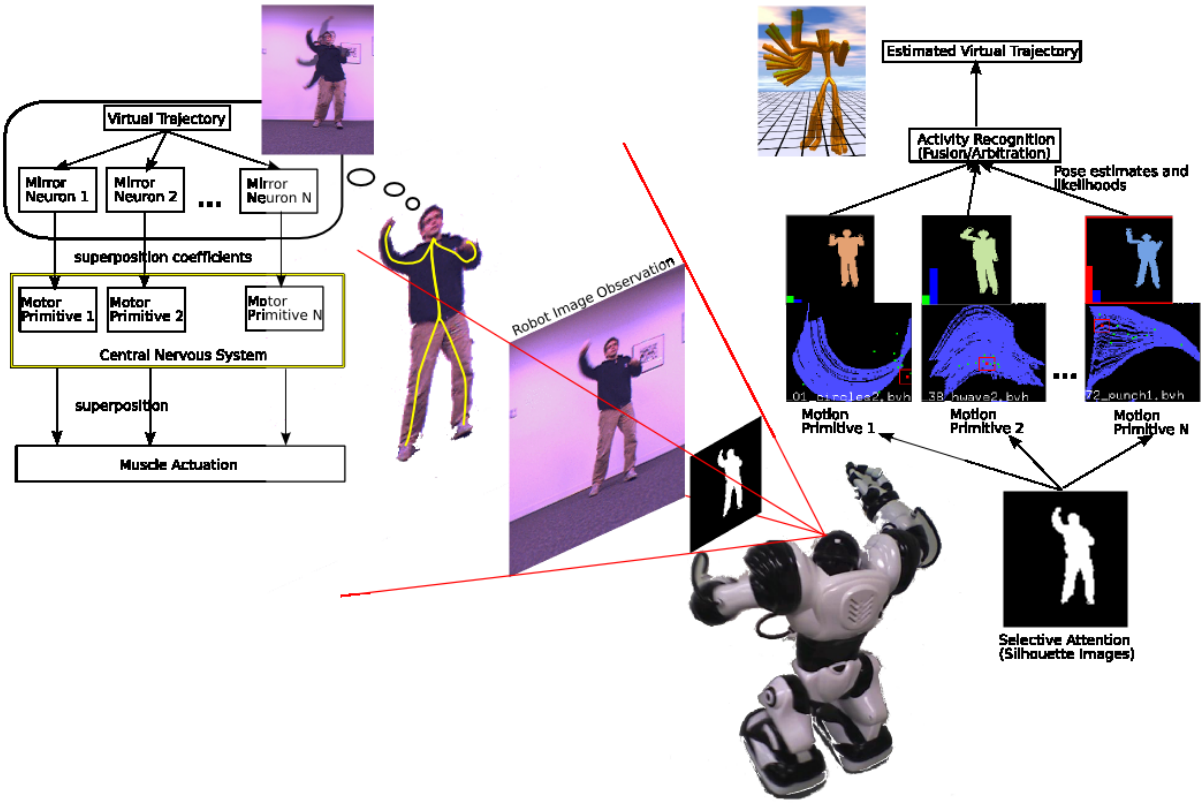

Figure 1: A "toy" example of our approach to human state estimation and movement imitation. The movement of a human demonstrator assumed to be generated by virtual trajectory executed as a weighted superposition of motor primitives, predictive low-dimensional dynamical systems. For movement imitation, 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.

recover from occlusion. Our system achieves interactive-time using sparse numbers of particles. We highlight the application of our tracking results to humanoid imitation.

2. BACKGROUND

2.1 Motor Primitives and Imitation Learning

The 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 [15], in specific computational mechanisms for humanoids. Matarić's model of imitation consists of: 1) a selective attention mechanism for extraction of observable features from a sensory stream, 2) mirror neurons that map sensory observations into a motor repertoire, 3) a repertoire of motor primitives as a basis for expressing a broad span of movement, and 4) a classification-based learning system that constructs new motor skills.

Illustrated in Figure 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 [16], 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) $\phi(x, \dot{x}, \ddot{x})$:

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

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 of equilibrium points specifies a virtual trajectory [5] of motion desireds for internal motor actuation or observed from 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 [21] have been demonstrated to fire when a particular activity is executed, observed, or imagined. Assuming 1-1 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 does a 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 [10] address 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. [13], 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. [8] 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. [1] and Grupen et al. [4, 19] approach robot control through sequencing and/or superposition of manually crafted behaviors.

Recent efforts by Knoop et al. [12] 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 other papers [10, 26, 14] 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 [9, 25] 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 [3], nonparametric belief propagation [23, 24], Gaussian process regression [26], POMDP learning [2] or dynamic programming [20], to account for the high dimensionality and local extrema of kinematic joint angle space. These methods tradeoff real-time performance for greater inference accuracy. Similar to Huber and Kortenkamp [7], interactive-time inference on actions to enable incorporation into a robot control loop. Unlike [7], 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.

3. DYNAMICAL KINEMATIC AND ACTION TRACKING

Kinematic tracking from silhouettes is performed via the steps in Figure 2, those are: 1) global localization of the human in the image, 2) primitive-based kinematic pose estimation and 3) action recognition. The human localization is kept as an unimodal distribution and estimated using the joint angle configuration derived in the previous time step.

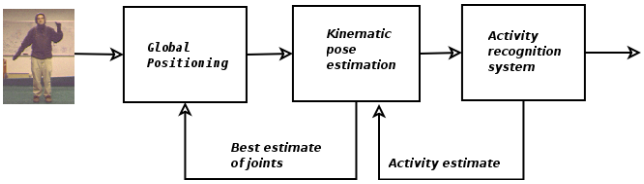

Figure 2: Illustration of the three stages in our approach to tracking: image observations are used to localize the person in 3D, then infer kinematic pose, and finally estimate of activity/action. Estimates at each stage are used to form priors for the previous stage at the next timestep.

3.1 Dynamical Motion Vocabularies

The methodology of Jenkins and Matarić [10] 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 [11] embed these motion trajectories into a lower dimensional space, establishing a separable clustering of movements into activities. Similar to [22], 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 i^{th} 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[1]:

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

where $u[t]$ is a fixed displacement magnitude, Δ_x is the gradient of pose x[2] a motion example of primitive i, and w_x the weight[3] of x w.r.t. $x[t]$. Figure 3 shows examples of learned predictive primitives.

Given results in motion latent space dimensionality [26, 10], 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

[1] `nbhd()` 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.
[2] The gradient is computed as the direction between y and its subsequent pose along its motion example.
[3] Typically reciprocated Euclidean distance

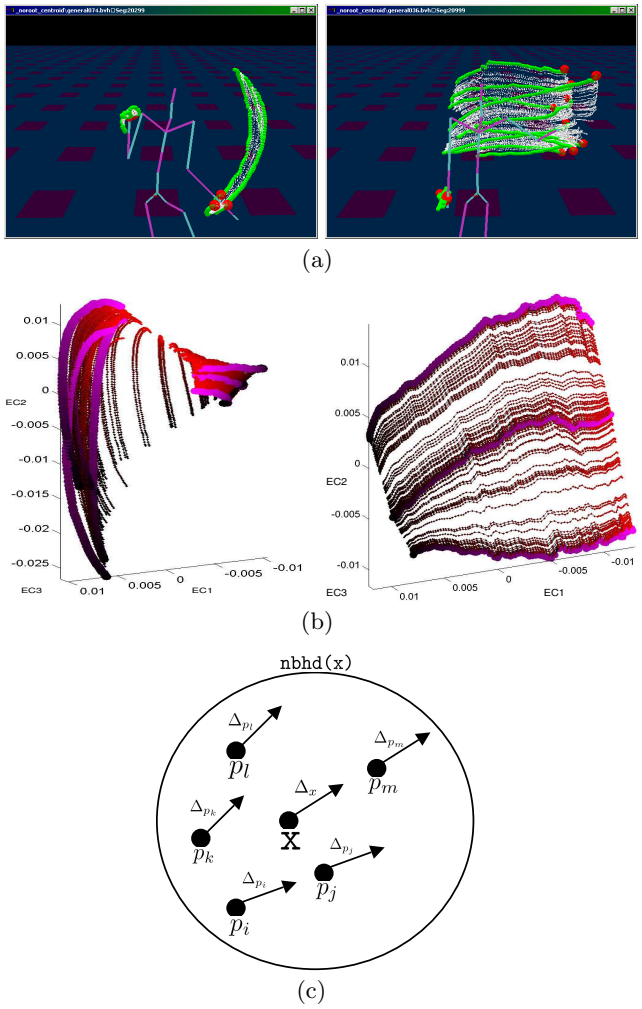

Figure 3: (a) Kinematic endpoint trajectories for learned primitive manifolds, (b) corresponding joint angle space primitive manifolds (view from first three principal components), and (c) an instantaneous prediction example (illustrated as a zoomed-in view on a primitive manifold.

X_i comprising primitive i and form the output equation of the dynamical system, such as in [6]:

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

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[4]. Although 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, 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 [9, 25] 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 [3]. A particle filter of the following form is instantiated in the latent space of each primitive

$$p(y_i[1:t] \mid z_i[1:t]) \propto p(z[t] \mid g_i^{-1}(y_i[t])) \sum_{y_i} p(y_i[t] \mid y_i[t-1]) p(y_i[1:t-1] \mid z[1:t-1]) \tag{5}$$

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

At first glance, the motion distribution $p(y_i[t] \mid y_i[t-1])$ could be given by the instantaneous "flow", as proposed by Ong et al. [18], 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 (Figure 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:

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

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.

3.3 Action Recognition

For action recognition, a probability distribution across primitives of the vocabulary is created[5]. The likelihood of the pose estimate from each primitive is normalized into a probability distribution:

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

[4] $x[t]$ and $y_i[t]$ are assumed to be homogeneous in 3

[5] We assume each primitive describes an action of interest

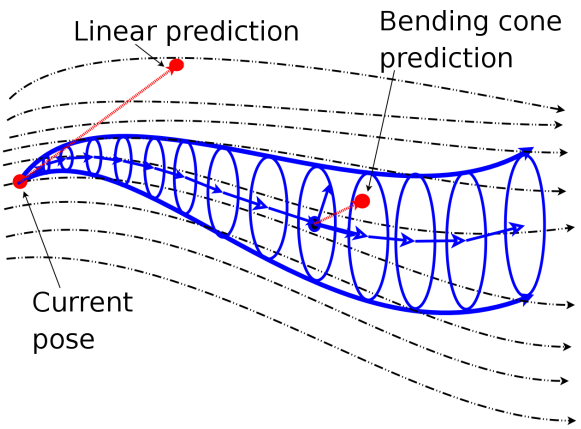

Figure 4: Illustration of the predictive bending cone distribution. The thin dashed black lines indicate the flow of a primitive's gradient field. Linear prediction from the current pose $y_i(t)$ will lead to divergence from the gradient field as the prediction magnitude increases. Instead, we use a 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.

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 is monotonically increasing. 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:

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

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 Figure 5, our system takes video input from a Fire-i webcam (15 frames per second, at a resolution of 120x160) 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 [10] was used to learn motion primitives for performing punching, hand circles, vertical hand waving, and horizontal hand waving.

We utilize a basic likelihood function, $p(z[t] \mid g_i^{-1}(y_i[t]))$, 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 distance was used:

Figure 5: Robot platform and camera used in our experiments.

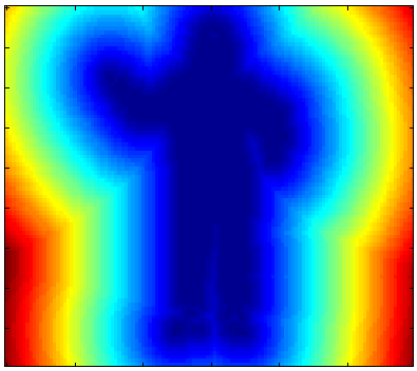

Figure 6: Hausdorff distance map for an image of standing human with one arm raised.

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

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

This measure is an intermediate between undirected and generalized Hausdorff distance 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 Figure 6.

To enable fast monocular tracking, we applied our system with sparse distributions (6 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, the actor performs three actions described by the motion primitives: hand circles, vertical hand waving and horizontal hand waving. For the purposes of evaluation, we compared the ground truth trajectories with the trajectories produced with sparse set of particles, ranging between six and two hundred. As shown in Figures 7

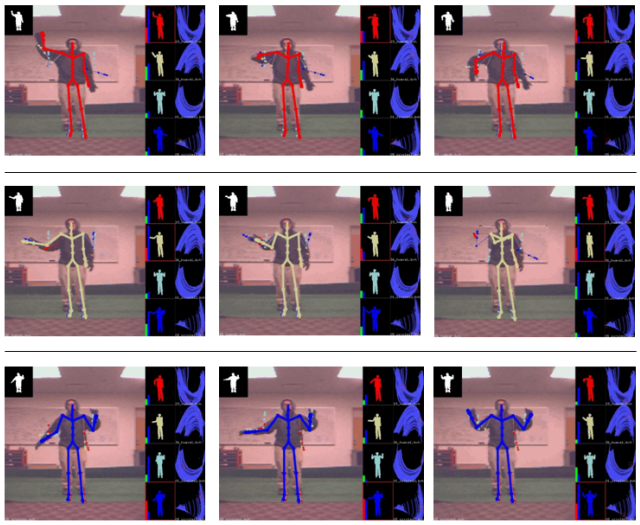

Figure 7: Tracking of motion sequence containing three distinct actions performed in sequence without stopping. Each row shows the recognition of individual actions for waving a hand top-to-bottom (top row), across the body (middle row), and bottom-to-top in a circular fashion (bottom row). The kinematic estimates are show 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.

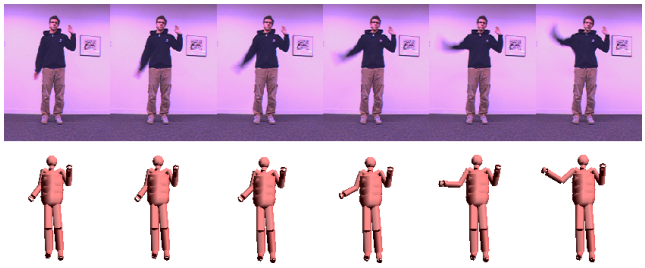

Figure 9: Tracking of a fast waving motion. Observed images (top) and pose estimates from the camera view (bottom).

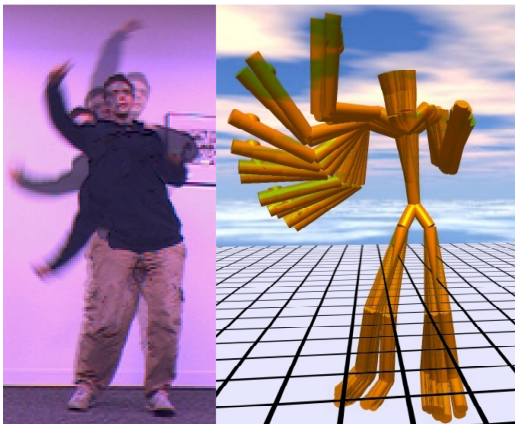

Figure 10: Illustrations of a demonstrated fast moving "punch" movement (left) and the estimated virtual trajectory (right) as traversed by our physically simulated humanoid simulation.

and 8, reasonable tracking estimates can be generated from as few as six particles. However, some tracking artifacts can be seen in Figure 8 due to resolution issues in the likelihood function. 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.

In trial two, we analyzed the temporal robustness of the tracking system. The same action is performed at different speeds, ranging from slow (hand moving at ≈ 3 cm/s) to fast motion (hand moving at ≈ 6 m/s). The fast motion is accurately predicted as seen in Figure 9. Additionally, we were able to track a fast moving punching motion (Figure 10) and successfully execute the motion with our physics-based humanoid simulation. Our simulation system is described in [27].

Viewpoint invariance was tested with video from a trial with an overhead camera, shown in Figure 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 all of the primitives into a probability distribution, as shown in Figure 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 Figure 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 [17]. 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

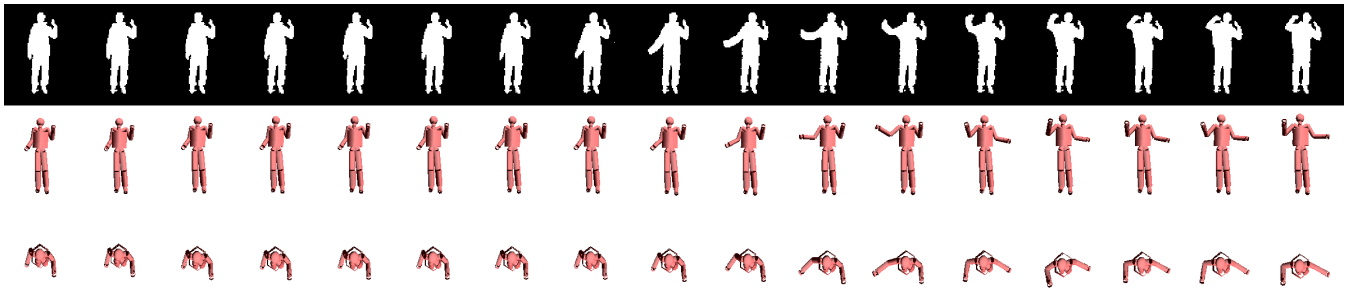

Figure 8: A sequence of predicted pose estimates from a multi-action motion. Observed silhouettes (top) and pose estimates (middle) with six particles per primitive. Pose estimates from an overhead view (bottom).

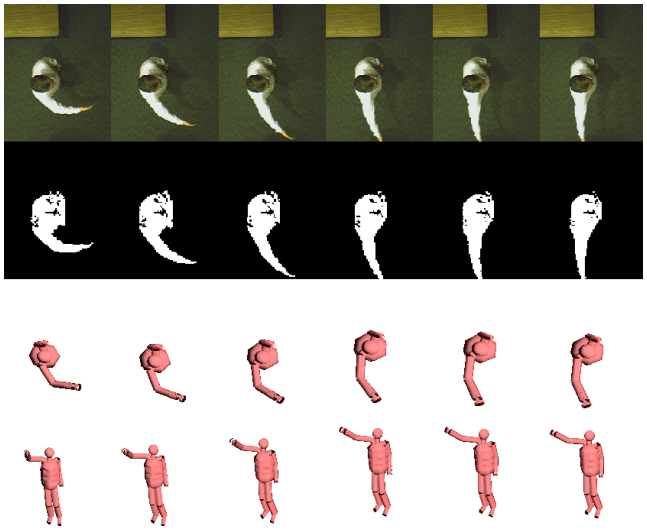

Figure 11: A sequence of pose estimates for a reaching motion. Observed silhouettes (second from top) can be compared with our pose estimates from the camera view (second from bottom) and from overhead (bottom).

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 [17].

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.

6. REFERENCES

[1] D. C. Bentivegna and C. G. Atkeson. Learning from observation using primitives. In *IEEE International Conference on Robotics and Automation*, pages 1988–1993, Seoul, Korea, May 2001.

[2] T. Darrell and A. Pentland. Active gesture recognition using learned visual attention. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 858–864. The MIT Press, 1996.

[3] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133, Hilton Head, SC, USA, 2000.

[4] R. A. Grupen, M. Huber, J. A. Coehlo Jr., and K. Souccar. A basis for distributed control of manipulation tasks. *IEEE Expert*, 10(2):9–14, 1995.

[5] N. Hogan. The mechanics of posture and movement. *Biol. Cybernet.*, 52:315–331, 1985.

[6] 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*, 12, 2000.

[7] E. Huber and D. Kortenkamp. A behavior-based approach to active stereo vision for mobile robots. *Engineering Applications of Artificial Intelligence*, 11:229–243, 1998.

[8] A. J. Ijspeert, J. Nakanishi, and S. Schaal. Trajectory formation for imitation with nonlinear dynamical systems. In *IEEE Intelligent Robots and Systems (IROS 2001)*, pages 752–757, Maui, Hawaii, USA, 2001.

[9] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[10] O. C. Jenkins and M. J. Matarić. Performance-derived behavior vocabularies: Data-driven acqusition of skills from motion. *International Journal of Humanoid Robotics*, 1(2):237–288, Jun 2004.

[11] O. C. Jenkins and M. J. Matarić. A spatio-temporal extension to isomap nonlinear dimension reduction. In *The International Conference on Machine Learning (ICML 2004)*, pages 441–448, Banff, Alberta, Canada, Jul 2004.

[12] S. Knoop, S. Vacek, and R. Dillmann. Sensor fusion

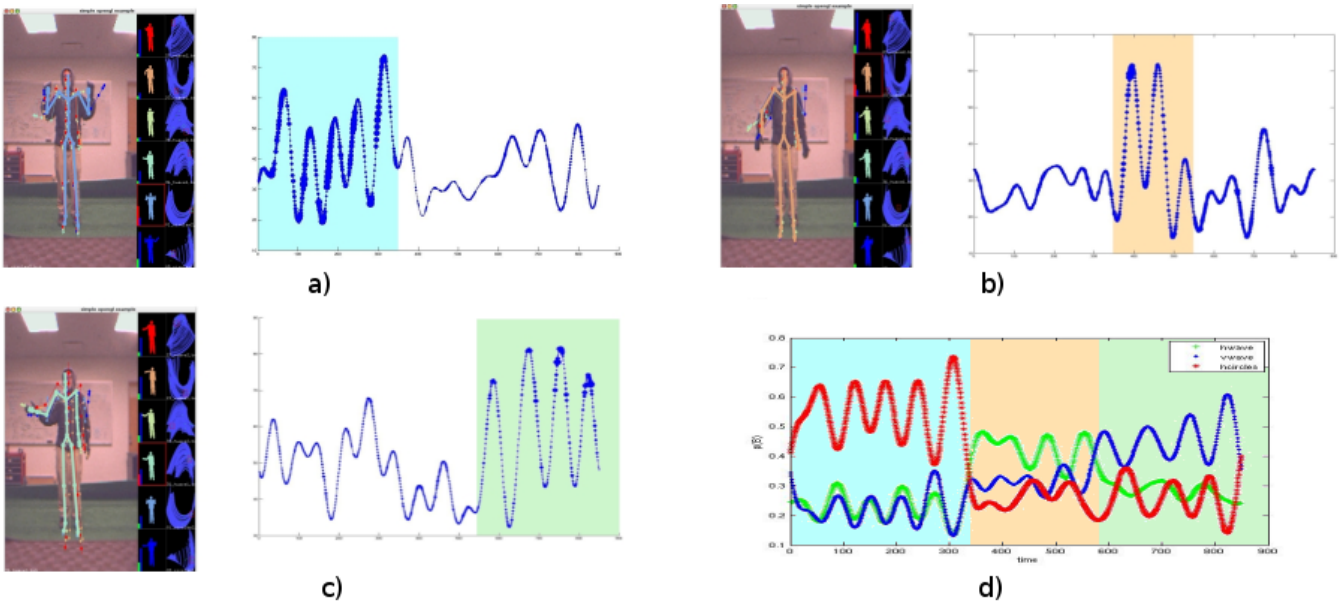

Figure 12: An evaluation of our action recognition system over time with a 3-action motion performing (a) "hand circles", (b) horizontal waving, and (c) vertical waving in sequence. Each plot shows time on the x-axis, attractor progress on the y-axis, and the width of the plot marker indicates the likelihood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

for 3D human body tracking with an articulated 3D body model. In *IEEE Intl. Conference on Robotics and Automation*, 2006.

[13] N. Kojo, T. Inamura, K. Okada, and M. Inaba. Gesture recognition for humanoids using proto-symbol space. In *IEEE International Conference on Humanoid Robots (Humanoids 2006)*, 2006.

[14] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, 2004.

[15] M. J. Matarić. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. In C. Nehaniv and K. Dautenhahn, editors, *Imitation in Animals and Artifacts*, pages 392–422. MIT Press, 2002.

[16] F. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Phil. Trans. R. Soc. Lond. B*, 355:1755–1769, 2000.

[17] M. Nicolescu, O. Jenkins, and A. Olenderski. Learning behavior fusion estimation from demonstration. In *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, Hatfield, United Kingdom, Sep 2006.

[18] E. Ong, A. Hilton, and A. Micilotta. Viewpoint invariant exemplar-based 3D human tracking. In *ICCV Modeling People and Human Interaction Workshop*, October 2005.

[19] R. Platt, A. H. Fagg, and R. R. Grupen. Manipulation gaits: Sequences of grasp control tasks. In *IEEE Conference on Robotics and Automation*, pages 801–806, New Orleans, LA, USA, 2004.

[20] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *Neural Info. Proc. Systems*, Vancouver, Canada, 2003.

[21] G. Rizzolatti, L. Gadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3:131–141, 1996.

[22] C. Rose, M. F. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics & Applications*, 18(5):32–40, Sep-Oct 1998. ISSN 0272-1716.

[23] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Computer Vision and Pattern Recognition*, pages 421–428, 2004.

[24] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *CVPR (1)*, pages 605–612. IEEE Computer Society, 2003.

[25] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[26] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *International Conference in Computer Vision*, Beijing, China, October 2005.

[27] P. Wrotek, O. Jenkins, and M. McGuire. Dynamo: Dynamic data-driven character control with adjustable balance. In *ACM SIGGRAPH Video Game Symposium*, Boston, MA, USA, Jul 2006.