Monocular Virtual Trajectory Estimation with Dynamical Primitives

Odest Chadwicke Jenkins, Germán González, Matthew M. Loper
Department of Computer Science
Brown University
Providence, RI 02912-1910
cjenkins,gerg,matt@cs.brown.edu

Abstract

We present a method for monocular kinematic pose estimation and activity recognition from video for movement imitation. Learned vocabularies of kinematic motion primitives are used emulate the function of hypothesized neuroscientific models for spinal fields and mirror neurons in the process of imitation. For imitation, we assume the movement of a demonstrator is produced through a "virtual trajectory" specifying desired body poses over time. Each pose represents a decomposition into mirror neuron firing coefficients that specify the attractive dynamics to this configuration through a linear combination of primitives. Each primitive is a nonlinear dynamical system that predicts expected motion with respect to an underlying activity. Our aim is to invert this process by estimating a demonstrator's virtual trajectory from monocular image observations in a bottom-up fashion. At our lowest level, pose estimates are inferred in a modular fashion through the use of a particle filter with each primitive. We hypothesize the likelihood of these pose estimates over time emulate the firing of mirror neurons from the formation of the virtual trajectory. We present preliminary results our method applied to video composed of multiple activities performed at various speeds and viewpoints.

Introduction

Human movement is a rich and untapped source of information for the autonomous control of robots. Many perception and control tasks necessary for autonomous robots could be performed or enhanced by greater understanding of human movement. Efforts in several robotics domains, including humanoid robotics and human-robot interaction, use human motion to drive (teleoperation), bootstrap (learning from demonstration), or guide (human-robot interaction) robot control processes towards greater functionality. Each of these require accurate measurements of the human's state, as kinematic state (joint angle or endeffector configurations) and/or behavioral state (activities being performed). Kinematic state measurements are typically used to drive a robot's actuators directly or generalized into a control policy. Behavioral state estimates are applicable towards control policies that are instructive (learning the decision process of the human) and collaborative (learning controllers for reacting to a human).

The analysis of human motion and its underlying mechanics is fraught with uncertainty, potentially due to the interactions of latent biomechanical structures. While many factors contribute to this uncertainty, much of this difficulty can be attributed to problems in state estimation and activity discovery. Kinematic state (or pose) estimation is a vastly underdetermined problem due to the large number of degrees-of-freedom (DOFs) in human kinematics and the limitations of current sensing technologies. Robot vision and sensing typically yields very limited information about the motion process for distinguishing pose. Additionally, such partial observations can introduce large degrees of uncertainty, such as variations in lighting and occlusions for cameras. Behavioral state estimation inherits many similar problems in addition to uncertainty about what comprises behavioral state. Unlike its kinematic equivalent, it is currently unclear what activities model human behavior. Consequently, the definition of a comprehensive state space for behavior has remained a moving target.

Motion modeling, the generation of parsimonious models of human movement, is central to defining the state space for behavior. Motion modeling is centered about two primary issues: 1) what activities should be modeled and 2) how do such models express the intrinsic structure of motion related to a specific activity? While many approaches to motion modeling exist, we are particularly inspired by neuroscientific hypotheses about how biological motion is produced and perceived(Mussa-Ivaldi & Bizzi 2000; Rizzolatti *et al.* 1996). We attempt to create motion models with similar functionality such that these models are suitable for robot perception and control.

In this paper, we describe our greater approach to learning of human motion vocabularies and their application to monocular kinematic tracking and activity recognition. From our previous work, we are able to learn a repertoire (or vocabulary) of predictive primitives, each expressing the expected nonlinear state dynamics an latent activity. We now use these primitives to perform movement imitation from monocular video. This process proceeds in a bottom-up fashion. Particle filters (Isard & Blake 1998; Thrun, Burgard, & Fox 2005), one for each primitive, are executed in parallel to infer kinematic state and emulate the

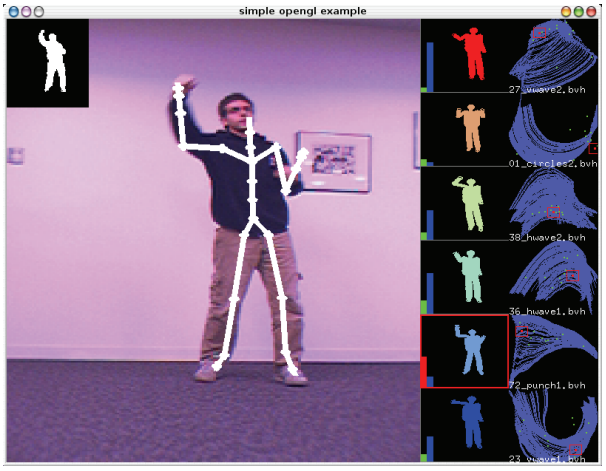


Figure 1: A snapshot of our tracking and recognition system processing a "punch" activity. Primitives and the dynamics of their associated activities are modeled as manifold-like gradient fields in kinematic joint angle space (right). The particle filter for each primitive yields a pose estimate (middle) of the sensor features (silhouette on left) for each activity. The estimate of the currently performed activity is extracted from the pose estimate likelihoods across activities. The height of the lefthand bars represent relative activity likelihood, while blue bars indicate attractor progress across the gradient field of the primitive.

firing of mirror neurons. Based on this "firing", we infer the activity performed at each time instant and form an estimate of the demonstrator's "virtual trajectory" and execute this estimate for movement imitation.

Background and Related Work

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 (Matarić 2002), 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.

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 (Mussa-Ivaldi & Bizzi 2000), 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 (Hogan 1985) 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 (Rizzolatti *et al.* 1996) 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.

While this model has desirable properties, there remain several challenges in its computational realization for autonomous robots. Namely, what are the set of primitives? How is each primitive and its parameterization expressed in computation? 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 work addresses these computational issues through the unsupervised learning of motion vocabularies and their usage with probabilistic inference. The alternative methods of Schaal et al. (Schaal *et al.* 2004; Ijspeert, Nakanishi, & Schaal 2001) 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 parsimony motion variability within each primitive and facilitates movement perception (such as monocular tracking) as well as control applications. Work proposed by Bentivegna et al. (Bentivegna & Atkeson 2001) and Grupen et al. (Grupen *et al.* 1995; Platt, Fagg, & Grupen 2004) approach robot control through sequencing and/or superposition of manually crafted behaviors.

Monocular Tracking and Data-Driven Motion Modeling

Many approaches to data-driven motion modeling have been proposed in computer vision, animation, and robotics. The reader is referred to other papers (Jenkins & Matarić 2004; Urtasun *et al.* 2005; Kovar & Gleicher 2004) for broader coverage of these methods. We pay particular attention to methods for kinematic tracking and activity recognition. Particle filtering (Isard & Blake 1998; Thrun, Burgard, & Fox 2005) is a well established means for inferring kinematic pose from image observations. Yet, particle filtering often requires additional procedures, like annealing

(Deutscher, Blake, & Reid 2000) or nonparametric belief propagation (Sigal *et al.* 2004; Sudderth *et al.* 2003), to account for the high dimensionality and local extrema of kinematic joint angle space. We explore the use of nonlinear dynamical models proposed by Jenkins and Matarić (Jenkins & Matarić 2004) to represent families of motion as manifolds in parsimonious state spaces.

The method we propose for motion modeling and kinematic tracking is similar to the work of Urtasun et al. (Urtasun *et al.* 2005) and Ong et al. (Ong, Hilton, & Micilotta 2005). Unlike these approaches geared towards a single motion and activity, our work has a broader focus in modeling and tracking motion composed of multiple activities. Although we use predictions similar to the instantaneous "flow" of Ong et al., we developed a "bending code" distribution to account for faster motion with a extended lookahead in time. Urtasun et al. use similar manifold learning techniques, but is focused on interpolating a set of motion in a probabilistic manner. In contrast, our approach to learning emphasizes clustering of motion representative of the demonstrator's underlying motor repertoire. Motion clusters produced by our method could be complimented by a probabilistic interpolation procedure.

Our unsupervised approach to activity recognition with a motion vocabulary is comparable to the semi-supervised techniques of Ramanan and Forsyth (Ramanan & Forsyth 2003). A *monolithic* semi-supervised motion library (Arikan, Forsyth, & O'Brien 2002) is created by learning decision boundaries between activities labeled through manual annotation. Monocular tracking and activity estimation is performed by synthesizing 3D motion with constraints given by the motion library and 2D image features. Our tracking and recognition approach is *modular*, allowing scalability for potentially large numbers of motion primitives.

Approach

For this paper, we focus specifically on movement imitation through virtual trajectory estimation from monocular image sequences (illustrated in Figure 2). We consider a limited instance of movement imitation where the objective is to form a trajectory in the robot's joint angle space representative of a human-demonstrated motion. We assume the essence of the demonstration is a virtual trajectory formed in the brain of the human instructor. The motion observed by the robot (via monocular camera input) is an execution of the virtual trajectory by equilibrium point control with a latent set of biological motor primitives.

We attempt to invert this process in an artificial robotic agent using learned dynamical primitives and probabilistic inference. Our previous work (Jenkins & Matarić 2004) discussed methods for uncover dynamical vocabularies from a large repository of motion data representative of natural human performance. Each primitive B_i is a gradient field expressing the expected kinematic behavior over time of the i^{th} activity. In the context of dynamical systems, this gradient field $B_i(x)$ defines the predicted direction of displacement for a location in joint space $\hat{x}[t]$ (a $1 \times F$ vector) at time t[1]:

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

where $u[t]$ is the expected displacement magnitude, Δ_y is the gradient of pose y a motion example of primitive i ,

For virtual trajectory estimation, each primitive is responsible for identifying when observed movement accords to the dynamics of specific activity. Towards this end, we instantiate a particle filter for each primitive to perform pose estimation with respect to a specific activity. The particle filter allows for estimation of the demonstrator's pose given the activity and image observations. Each primitive constrains the priors and motion models of its filter to consider relevant subspaces and avoid computational intractability. These constraints are expressed through estimates of the intrinsic dimensionality and dynamic predictions of each primitive. By taking this approach we avoid computational issues incurred by standard methods for probabilistic inference in this manner (Deutscher, Blake, & Reid 2000).

We analogize the likelihood of each primitive's pose estimate to the firing of an idealized mirror neuron. We hypothesize that the pose likelihoods across all primitives is proportional the primitive combination coefficients $u[t]$, expressing the control dynamics at time t. These likelihoods are used by an activity recognition system to determine the superposition coefficients over time to form an estimate of the demonstrator's virtual trajectory. The estimated trajectory is then actuated by the robot to execute the imitation. Ideally, the trajectory estimate can be bypassed and the superposition coefficients from fusion could be directly executed by the robot's primitives to perform imitation. However, given that primitives are learned sequentially, we focus on activity recognition as an arbitration process and present positive results for this limited implementation. We assume a transform between the kinematics of the demonstrator and the robot is known.

Monocular Kinematic Pose Estimation

Kinematic tracking from silhouettes is performed via the following procedure. The global root of the human are determined using the best estimate of the kinematic pose in the previous time step for each primitive. Kinematic pose is then inferred using a particle filter in the latent space of each primitive. The demonstrator's activity and virtual trajectory is estimated based on arbitrating the most likely pose estimate at a given time and concatenating over time.

For each primitive, an estimate over the subject's global position (x,y,z and θ, rotation about the vertical axis) is maintained as a unimodal distribution. These estimates are initialized by a random distribution about the silhouette center of mass. These estimates are refined over time: Gaussian

[1] `nbhd()` identifies the k-nearest neighbors in an arbitrary coordinate space, used both in joint angle space and the space of motion segments.

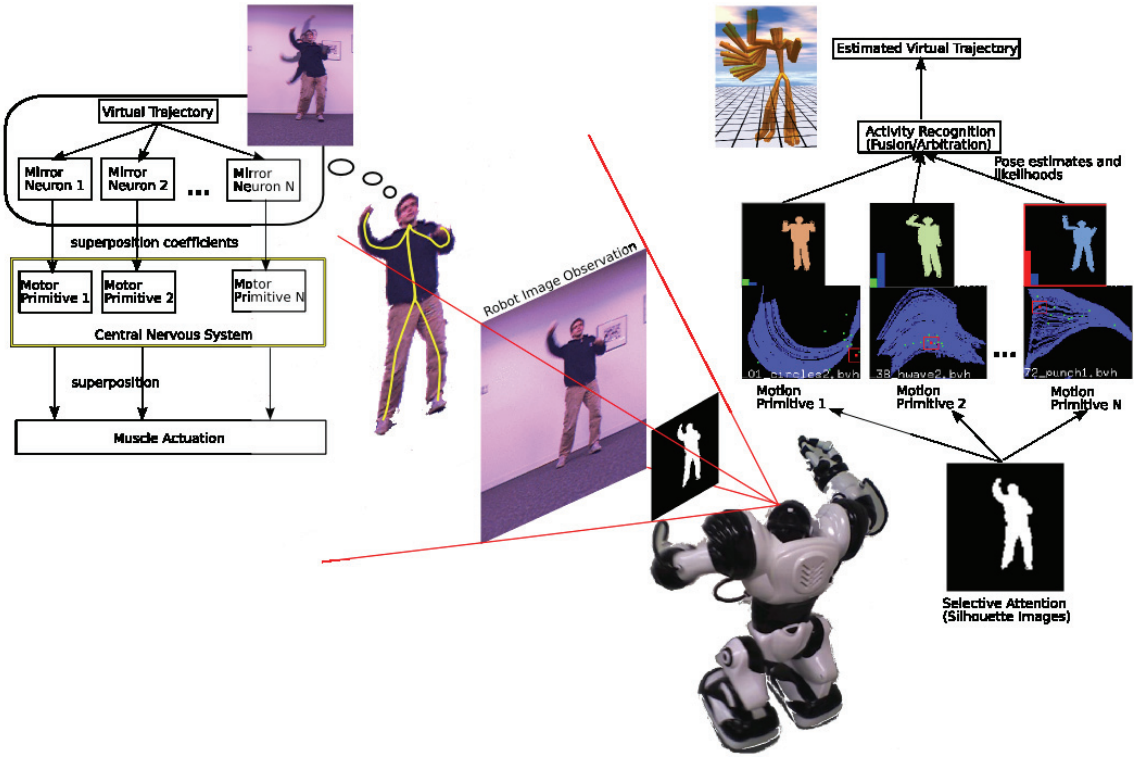


Figure 2: A "toy" example our approach to movement imitation. The movement of a human demonstrator assumed to be generated by virtual trajectory executed as a weighted superposition of motor primitives. We estimate these motor primitives from human motion data as a vocabulary of predictive low-dimensional 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.

noise is added to the current position, and if the pose likelihood is calculated to be higher, the new position is accepted.

Kinematic Pose Estimation

Kinematic tracking is performed by particle filtering (Isard & Blake 1998; Thrun, Burgard, & Fox 2005) in the individual latent spaces created for each primitive in a motion vocabulary. The primitives infer pose individually and in parallel to avoid high-dimensional state spaces, encountered in (Deutscher, Blake, & Reid 2000). Given results in motion latent space dimensionality (Urtasun *et al.* 2005; Jenkins & Matarić 2004), we construct a low dimensional latent space to provide parsimonious observables y_i of the joint angle space for primitive i. These observables are expressed in the output equation of the dynamical system of each primitive, such as in (Howe, Leventon, & Freeman 2000):

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

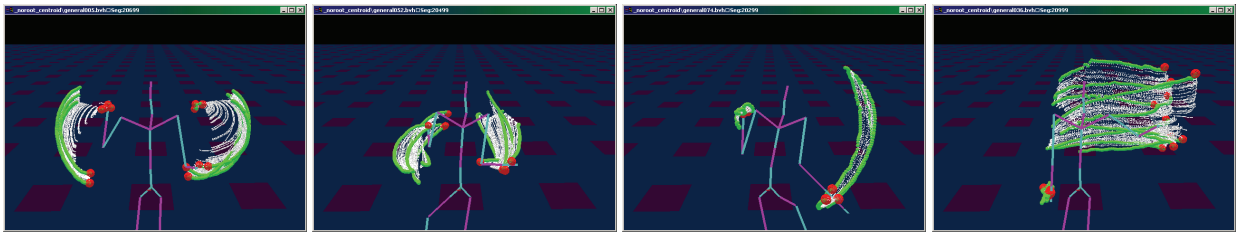
where g_i is the latent space transformation and A_i is the expression of g_i as an affine transformation into the principal component space of primitive i[2]. We chose a linear transform for g_i for inversion simplicity and evaluation speed. A particle filter of the following form is instantiated in the latent space of each primitive :

$$p(y_i[1:t] \mid z_i[1:t]) \propto p(z[t] \mid g_i^{-1}(y_i[t])) \sum_{y_i} p(y_i[t] \mid y_i[t-1]) p(y_i[1:t-1] \mid z[1:t-1]) \tag{4}$$

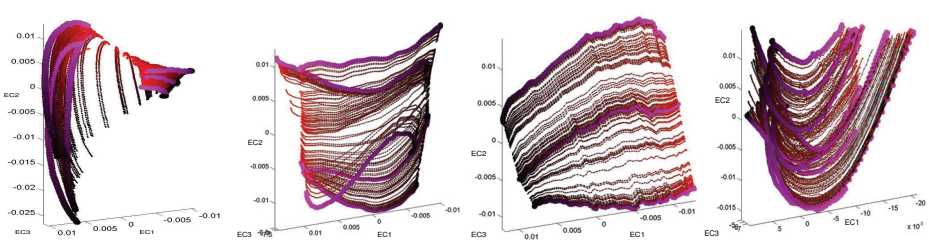
where $z_i[t]$ are the observed sensory features at time t and g_i^{-1} is the transformation into joint angle space from the latent space of primitive i. The likelihood function $p(z[t] \mid g_i^{-1}(y_i[t]))$ can be any reasonable choice for comparing the hypothesized observations from a latent space particle and the sensor observations. Ideally, this function will be monotonic with discrepancy in the joint angle space. With linear transformation $g_i(x)$, the posterior from the previous time step $p(y_i[1:t-1] \mid z_i[1:t-1])$ is a d-dimensional ellipsoid in joint angle space. For faster computation, we use latent space motion dynamics in formulation of the motion model distribution $p(y_i[t] \mid y_i[t-1])$. We assume latent space dynamics, governed by $\tilde{f}_i$[3] , are a reasonable direc-

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

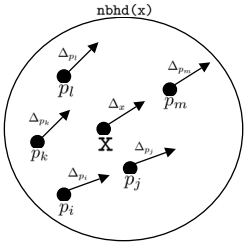
[3] $\tilde{f}_i$ in a latent space is computed in the same manner as f in



(a) Kinematic endpoint trajectories for learned primitive manifolds: (left to right) hand circles, dancing "the twist", punching, arm waving



(b) Joint space primitive manifolds view from their first 3 principal components



(c) Prediction example

Figure 3: Examples of manifold structures (a,b) learned from unlabeled kinematic time-series and their predictive dynamics (c) within a neighborhood of a joint space point x.

tional approximation of joint space dynamics:

$$\frac{g_i^{-1}(\tilde{f}_i(g_i(x[t]),u[t])) - x[t]}{\|g_i^{-1}(\tilde{f}_i(g_i(x[t]),u[t])) - x[t]\|} \approx \frac{f_i(x[t],u[t]) - x[t]}{\|f_i(x[t],u[t]) - x[t]\|} \tag{5}$$

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. (Ong, Hilton, & Micilotta 2005), where a locally linear displacement with some noise is expected. However, such an assumption would require temporal coherence in the sensory observations of the motion. 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, we use a "bending cone" distribution (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 about this 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.

joint space.

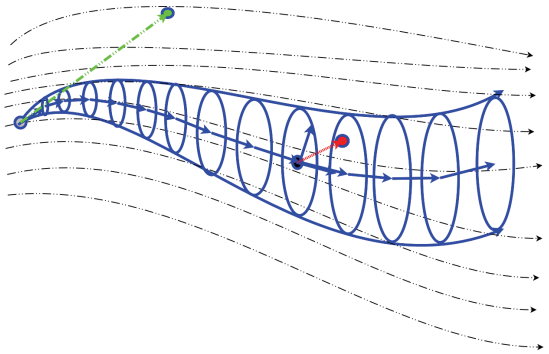


Figure 4: Illustration of the bending cone distribution. The blue dot represents $y_i(t)$ a pose hypothesis. The green line and dot show the divergence that results from prediction by extending linear displacement. Instead, we use a bending cone (in blue) to provide an extended hypothesis horizon. We sample the bending cone for a motion update $y_i(t+1)$ (red dot) by selecting a cross-section A(t)[k] (black dot) and adding cylindrical noise.

We hypothesize that observed motion can be tracked if: 1) the primitive is representative of the motion at that interval of time, 2) the bending cone is well sampled, and 3) K is a far enough look-ahead.

Activity Recognition

For activity recognition, we create a probability distribution across primitives of the vocabulary, assuming each primitive is an activity of interest. We take the likelihood of the pose estimate from each primitive and normalize them 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 activity currently being performed. As we describe in Section , this technique for activity recognition has the potential to be improved by considering "attractor progress".

Experimental Results

For our experiments, we developed a system in C++ that used a vocabulary of learned motion primitives to track kinematic motion and estimate activities performed from monocular silhouettes. The motion vocabulary was provided by the authors of (Jenkins & Matarić 2004) from their previous work. We used a subset of their motion primitives for performing punching, hand circles, vertical hand waving, and horizontal hand waving. Observations, as silhouettes, were computed with standard background subtraction and median filtering techniques for color images. Image sequences were obtained from a Fire-i webcam at 15 frames per second, at a resolution of 120x160.

We use a likelihood function, $p(z[t] \mid g_i^{-1}(y_i[t]))$, returns the similarity of a particle's hypothesized silhouette with the observed silhouette image. Silhouette hypotheses were rendered from a cylindrical 3D body model to an image buffer using OpenGL. We did not specifically adapt the body model to either of the subjects used for the videos. A similarity metric, $R(A, B)$ for two silhouettes A and B, closely related to the inverse of the generalized Hausdorff distance was used:

$$R(A, B) = \frac{1}{r(A, B) + r(B, A) + \epsilon} \tag{8}$$

$$r(A, B) = \sum_{a \in A} \left(\min_{b \in B} \|a - b\| \right)^2 \tag{9}$$

This measure is an intermediate between undirected and generalized Hausdorff distance and generalized Hausdorff distance ϵ is use only to avoid divide-by-zero errors.

Dynamical Tracking with Sparse Particles

To evaluate the practical aspects of our monocular tracking, we applied our system with sparse distributions (6 particles per primitive) to three trial silhouette sequences. Each of the trials is designed to provide insight into different aspects of the performance of our tracking system. The result of each imitation trial was output to a Biovision format motion capture file and executed on a dynamically simulated humanoid using the Open Dynamics Engine.

Figure 5: Illustrations of a demonstrated fast moving "punch" movement (left) and the estimated virtual trajectory (right) as traversed by our humanoid simulation.

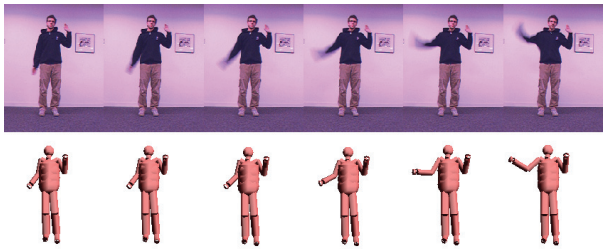


Figure 6: Tracking of a fast waving motion. Observed images (top) and pose estimates from the camera view (bottom).

In trial one, the actor performs three activities 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 Figure **??**, 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.

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 $\approx$ 3 cm/s) to fast motion (hand moving at $\approx$ six m/s). The fast motion is accurately predicted as seen in Figure 6. Additionally, we were able to track a fast moving punching motion (Figure 5).

Viewpoint invariance is critical aspect of our system and

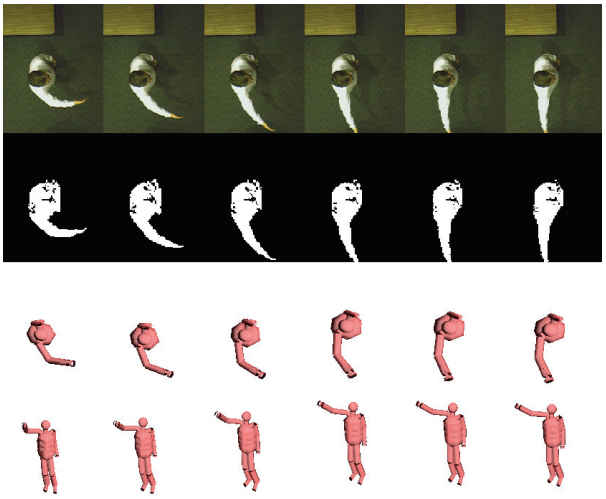


Figure 7: A sequence of pose estimates for a reaching motion. Observed silhouettes (second from top) can be compared with our pose estimates from the camera view (second from bottom) and from overhead (bottom).

was tested with video from a trial with an overhead camera, shown in Figure 7. 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.

Activity Recognition and Movement Imitation

Using the above test trials, we measured the ability of our system to recognize performed activities to provide responses similar to mirror neurons. In our current system, activity is recognized as the pose estimate likelihoods normalized over all of the primitives into a probability distribution, as shown in Figure 8. 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 activity would consider monotonic progress consistent with the dynamics of the trajectories in the manifold as a factor in the likelihood of the activity. We consider this property to be *attractor progress*. We have analyzed preliminary results from observing attractor progress in our trials, as shown in Figure 8. For an activity being performed, its attractor progress is monotonically increasing. If the activity is performed repeatedly, we can see a periodic signal emerge, as opposed to the noisier signals of the activities not being performed. These results indicate that we can use attractor progress as a feedback signal into the particle filter estimating pose for a primitive i in a form such as:

$$p(B_i[t] \mid z[t]) = \frac{p(z[t] \mid \bar{x}_i[t], w_i[1:t-1])}{\sum_B p(z[t] \mid \bar{x}_i[t], w_i[1:t-1])} \quad (10)$$

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

Because of their attractor progress properties, we believe that we can analogize these activity patterns into the firing of an idealized mirror neurons. The firing of our artificial mirror neurons provide superposition coefficients. 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 activity 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. In particular, relatively achievable improvements in speed and robustness to occlusion could be obtained with better selective attention techniques, such as approximations of the Haussdorf distance for speed.

Conclusion

We have presented a neuroinspired method for monocular tracking and activity recognition for movement imitation. Our approach combines learning vocabularies of kinematic motion offline to perform 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. Preliminary results suggest our method can perform tracking and recognition from partial observations at interactive rates. Our current system demonstrate robustness with respect to the viewpoint of the camera, the speed of performance of the action, and recovery from ambiguous situations.

References

Arikan, O.; Forsyth, D. A.; and O'Brien, J. F. 2002. Motion synthesis from annotations. *ACM Transactions on Graphics* 22(3):402–408.

Bentivegna, D. C., and Atkeson, C. G. 2001. Learning from observation using primitives. In *IEEE International Conference on Robotics and Automation*, 1988–1993.

Deutscher, J.; Blake, A.; and Reid, I. 2000. Articulated body motion capture by annealed particle filtering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 126–133.

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

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

Howe, N. R.; Leventon, M. E.; and Freeman, W. T. 2000. Bayesian reconstruction of 3d human motion from single-camera video. *Advances In Neural Information Processing Systems* 12.

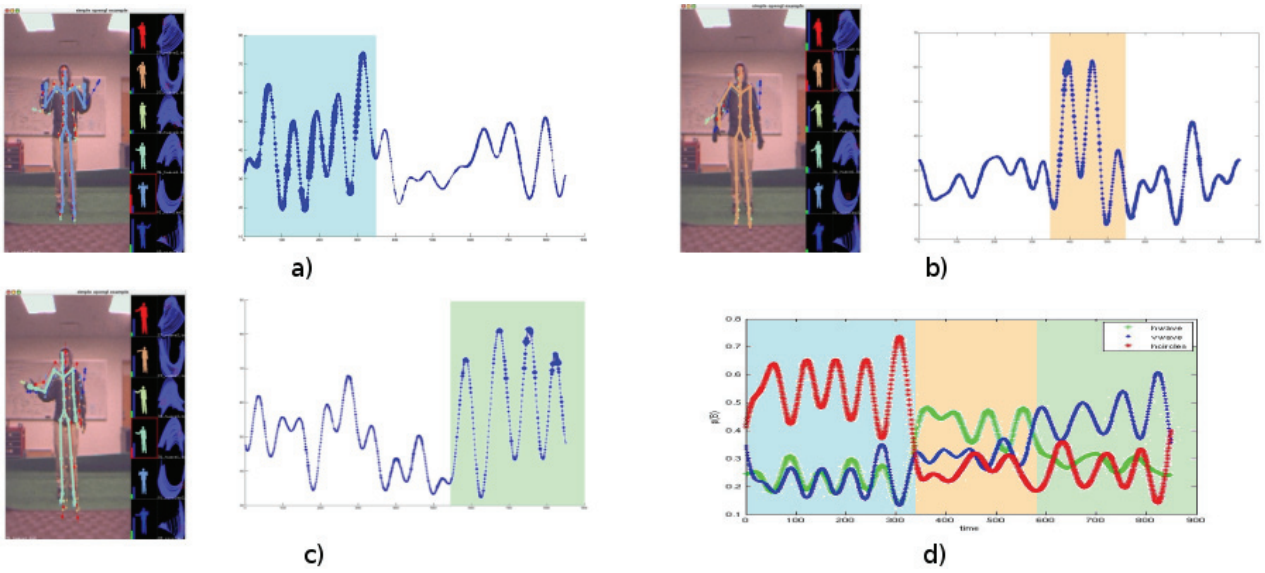


Figure 8: An evaluation of our activity recognition system over time with a 3-activity motion performing (a) "hand circles", (b) horizontal waving, and (c) vertical waving in sequence. Each plot shows time on the x-axis, attractor progress on the y-axis, and the width of the plot marker indicates the 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 activity.

Ijspeert, A. J.; Nakanishi, J.; and Schaal, S. 2001. Trajectory formation for imitation with nonlinear dynamical systems. In *IEEE Intelligent Robots and Systems (IROS 2001)*, 752–757.

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

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

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

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

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

Ong, E.; Hilton, A.; and Micilotta, A. 2005. Viewpoint invariant exemplar-based 3d human tracking. In *ICCV Modeling People and Human Interaction Workshop*.

Platt, R.; Fagg, A. H.; and Grupen, R. 2004. Manipulation gaits: Sequences of grasp control tasks. In *IEEE Conference on Robotics and Automation*, 801–806.

Ramanan, D., and Forsyth, D. A. 2003. Automatic annotation of everyday movements. In *Neural Info. Proc. Systems*.

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

Schaal, S.; Peters, J.; Nakanishi, J.; and Ijspeert, A. 2004. Learning movement primitives. In *International Symposium on Robotics Research*.

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

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

Thrun, S.; Burgard, W.; and Fox, D. 2005. *Probabilistic Robotics*. MIT Press.

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