Dynamical Motion Vocabularies for Kinematic Tracking and Activity Recognition

Odest Chadwicke Jenkins, Germán González, Matthew Loper
Computer Science Department - Brown University
115 Waterman St., 4th Floor
Providence, Rhode Island
{cjenkins, gerg, matt}@cs.brown.edu

Abstract

We present a method for 3D monocular kinematic pose estimation and activity recognition through the use of dynamical human motion vocabularies. A motion vocabulary is comprised as a set of primitives that each describe the movement dynamics of an activity in a low-dimensional space. Given image observations over time, each primitive is used to infer the pose independently using its expected dynamics in the context of a particle filter. Pose estimates from a set of primitives are inferred in parallel and arbitrated to estimate the activity being performed. The approach presented is evaluated through tracking and activity recognition over extended motion trials. The results suggest robustness with respect to multi-activity movement, movement speed, and camera viewpoint.

1. Introduction

Computer vision algorithms are often faced with the problem of high-dimensional state estimation from uncertain and ambiguous information. This circumstance especially holds in the cases of monocular kinematic tracking (state represented as joint angles) and recognition of human activity (state represented as combinations of activities)[12, 2]. One approach to aid computationally tractable inference for these problems is to construct parsimonious models describing and constraining motion or activity over time. More specifically, such models: 1) constrain the state space based on an estimate of an*a priori* probability and 2) describe dynamical behavior as the probability of moving from one state to another.

The problem of *motion modeling* underlies the generation of parsimonious models for use in a variety of applications (e.g., robotics, animation, biomechanics, neuroscience). Many techniques construct motion models using human expertise and domain knowledge [2, 18]. Such

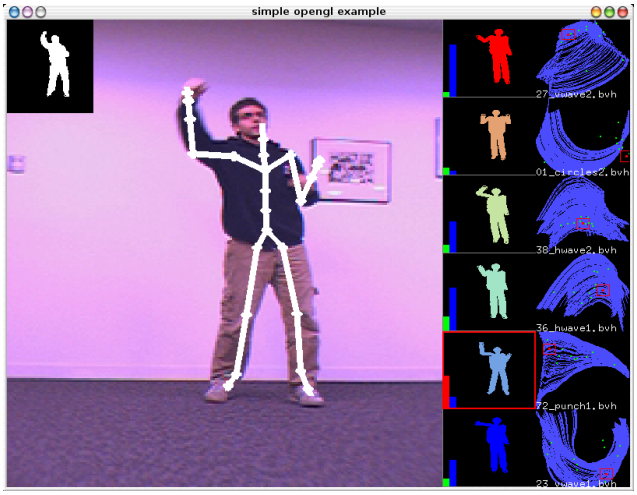


Figure 1. A snapshot of our tracking and recognition system tracking a "punch" activity. Primitives and the dynamics of their associated activities are modeled as manifold-like gradient fields in kinematic joint angle space (right). The particle filter for each primitive yields a pose estimate (middle) 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.

model-based techniques, however, are prone to errors in human judgment, tedious to construct, and often lack scalability for new or unforseen behavior. With advances in machine learning and motion capture technology, *data-driven motion modeling* [24, 10] has become an increasingly compelling approach to learning models from human performance in an application-independent manner.

In this paper, we describe our approach to monocular kinematic tracking and activity recognition using a vocabulary of dynamical motion primitives. We consider a vocabulary a repertoire of predictive primitives, where each

primitive expresses the nonlinear state dynamics for its associated activity.

We learn primitives from natural motion data through the methodology described by Jenkins and Matarić [10]. A primitive of this form is given by a set of example trajectories in joint angle space consistent with a common activity. These trajectories are generalized through interpolation [18], speculatively evaluated to form a manifold, and traversed dynamically. As shown in [10, 24], 2-3 dimensions are often sufficient for describing the space spanned by the manifold of a primitive. Given these results, we construct a low dimensional latent space for each primitive using Principal Components Analysis (PCA). These latent spaces provide a parsimonous space for expressing movement dynamics suitable for probabilistic inference [9, 23] of pose over time.

We view primitives as a discretization of activity states to bridge kinematic tracking and activity recognition, similar to [16]. A baseline activity recognition procedure is presented that estimates the current activity being performed as the maximally likely pose estimate over all primitives. In this respect, our primitives attempts to emulate hypothesized neuroscientific spinal field primitives [13] and recognition of activities as mirror neuron firing [17].

2. Related Work

Many different approaches to data-driven motion modeling have been proposed in the various fields of computer vision, animation, and robotics. The reader is referred to other papers [10, 24] for a broader coverage.

A plethora of methods have been proposed for monocular tracking, such as [1, 19]. Among them, several multiple hypothesis approaches exist, such as switching linear dynamical systems [15], piecewise mixture of Gaussians [5] and particle filtering [9, 23]. Particle filtering is a well established means for inferring kinematic pose from image observations. However, particle filtering often requires additional procedures, such as annealing [6] or nonparametric belief propagation [20, 22], to account for the high dimensionality and local extrema of the kinematic joint angle space. Such methods retain the flexibility of an unconstrained prior at a greater computational cost.

The integration of data-driven motion modeling and monocular kinematic tracking has been recently explored in [24, 14, 7, 25] to construct priors for a specific activity. Elgammal and Lee [7] use manifold learning to learn single activity latent spaces for multiple viewpoints. Similarly, Urtasun et al. [24] learn a prior through the low dimensional embedding of one motion for a single activity. Wang et al. [25] extends this method to additionally learn the dynamics of a movement. However, these dynamical models are often prone to divergence (drifting away from the manifold), which they address through an additional optimization step. We attribute this divergence tendency to the choice of an overestimated and fixed timestep. Similar to systems for rigid body dynamics [4], overextended timesteps often result in instability in the dynamics and expect movement to occur at a certain speed given the current state of the system. In contrast, we account for uncertainty in the speed and noise of an observed motion performance through a "bending cone" distribution. The bending cone integrates over small timestep predictions performed in sequence with increasing uncertainty. By making concatenated short predictions over time, our predictions avoid divergence and the need for correction through optimization. Similar to the annotation work of Ramanan and Forsyth [16], our approach attempts to fuse tracking and activity recognition. In their work, annotation is performed by searching over a motion database previously annotated with activity labels [3] to find the motion and corresponding activities that best explain the image observations. Our method differs from the annotation approach in that we model the predictive dynamics of the activities and avoid the prohibitive cost of dynamic programming and potential drift caused by "dead-ends".

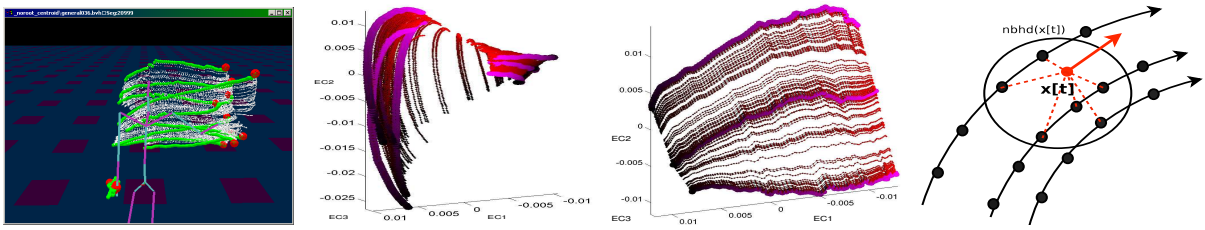
3. Kinematic Tracking

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

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 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 nonoverlapping segments representative of "atomic" movements. Spatio-temporal Isomap [11] embeds these motion trajectories into a lower dimensional space and establishes a separable clustering of movements into activities. Similar to [18], each cluster is a group of motion examples that can be interpolated to produce new motion representatives of the underlying activity. Each cluster is speculatively evaluated to produce a dense collection of examples for each uncovered activity. A primitive 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 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



(a) (b) (c)

Figure 2. (a) Kinematic endpoint trajectories for learned primitive manifolds. (b) Joint space primitive manifolds (first three dimensions) (c) a prediction example

for a location in joint space $\hat{x}[t]$ at time t[1]:

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

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 2 shows examples of learned predictive primitives.

Given results in motion latent space dimensionality [24, 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 Component Analysis (PCA) to all of the poses x_i comprising primitive i and form the dynamical system, such as in [8]:

$$y_i[t] = g_i(x[t]) = A_i x[t] \quad (2)$$

where g_i is the latent space transformation and A_i is the formulation 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:

$$\frac{g_i^{-1}(\tilde{f}_i(g_i(x[t]), u[t])) - x[t]}{\|g_i^{-1}(\tilde{f}_i(g_i(x[t]), u[t])) - x[t]\|} \approx \frac{f_i(x[t], u[t]) - x[t]}{\|f_i(x[t], u[t]) - x[t]\|} \quad (3)$$

[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

[4] $x[t]$ and $y_i[t]$ are assumed to be homogeneous in equation 2

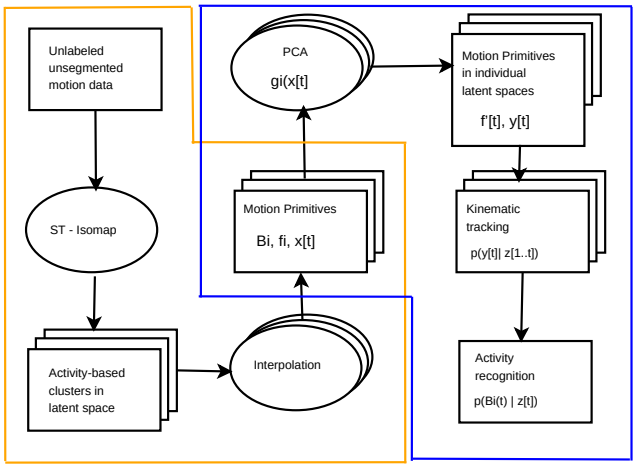


Figure 3. Creation and usage of the motion primitives. The orange box indicates the work of Jenkins and Matarić [10] while the blue box indicates the contributions of this paper.

.

3.2. Kinematic Pose Estimation

Kinematic tracking is performed by particle filtering [9, 23] in the individual latent spaces created for each primitive in a motion vocabulary. We infer the kinematic pose with each primitive individually and in parallel to avoid high-dimensional state spaces, encountered in e.g. [6]. A particle filter of the following form is instantiated in the latent space of each primitive.

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

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. It can be based in distance transform of edges [14],

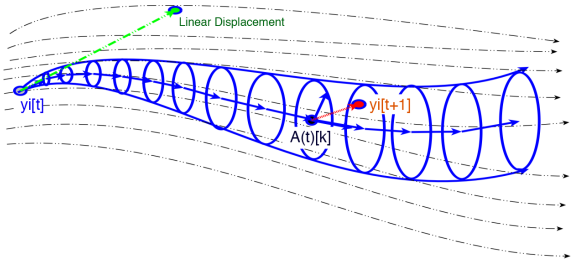


Figure 4. Illustration of the bending cone distribution. The blue dot represents $y_i(t)$ a pose hypothesis. The green line and dot show the divergence that results from prediction by extending linear displacement. Instead, we use a bending cone (in blue) to provide an extended hypothesis horizon. We 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.

silhouettes, image contours, optical flow or a combination of these [21]. 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 instanteous "flow", as proposed by Ong et al. [14], 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 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 has 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 the local axis $a - b$ and normally distributed variance σ about this axis. The resulting parametric distribution:

$$p(y_i[t] \mid y_i[t-1]) = \sum_{\hat{y}_i[t]}^{k} \mathcal{C}(\hat{y}_i[k+1], \hat{y}_i[k], f(k)) \quad (5)$$

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. Activity Recognition

For activity 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:

$$p(B_i \mid z[t]) = \frac{p(z[t] \mid \bar{x}_i[t])}{\sum_{B_i} p(z[t] \mid \bar{x}_i[t])} \quad (6)$$

where $\bar{x}_i[t]$ is the pose estimate for primitive Bi. The primitive with the maximum probability is estimated as the activity 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. A better estimator of activity would thus use the monotonic progress consistent with the dynamics of the trajectories in the manifold as a factor in the likelihood of the activity. We call this property *attractor progress*. For an activity being performed, 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:

$$p(B_i \mid z[t]) = \frac{p(z[t] \mid \bar{x}_i[t], w_i[1:t-1])}{\sum_{B_i} p(z[t] \mid \bar{x}_i[t], w_i[1:t-1])} \quad (7)$$

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

4. Experimental Results

In our experimentation, a system in C++ that used a vocabulary of learned motion primitives was used to track kinematic motion from monocular silhouettes. The motion primitives employed describe the activities of punching, hand circles, vertical and horizontal hand waving. Silhouettes were computed with standard techniques[6] from color images. Image sequences were obtained from a Fire-i webcam at 15 frames per second, at a resolution of 120x160.

The likelihood $p(z[t] \mid g_i^{-1}(y_i[t]))$ is calculated for each particle by measuring the similarity of its hypothesized silhouette with an observed silhouette. Silhouette hypotheses were rendered from a cylindrical 3D body model. We did not adapt the body model to either of the subjects used for the videos. To measure similarity $R(A, B)$ between two such silhouettes A and B, we used a metric closely related to the inverse of the generalized Hausdorff distance:

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

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

,

where ϵ is present only to avoid divide-by-zero errors.

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

[6] Background subtraction and median filtering

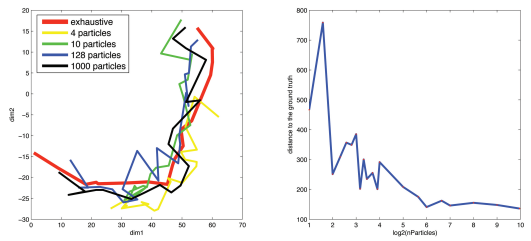


Figure 5. a) Comparison of trajectories in the latent space. b) A plot of the Euclidean distance between trajectories of particles in the latent space, as a function of the number of particles in each primitive.

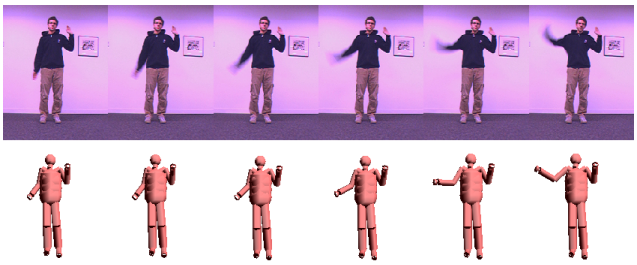


Figure 6. Tracking of a fast waving motion. Observed images can be compared with our best pose estimate from the camera view (bottom).

4.1. Dynamical Tracking with Sparse Particles

The system was applied to three trial silhouette sequences with sparse distributions (6 particles per primitive). 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", generated by exhaustive search over the points of the manifod, with the trajectories produced with a sparse set of particles. The results can be seen in Figure 5. As expected, the Euclidean distance between the estimates and the ground truth decreases with the number of particles used in the simulation.

In the second test sequence the temporal robustness of the tracking system is analyzed. The same action is performed at different speeds, ranging from slow (hand moving at $\approx$ 3 cm/s) to fast motion (hand moving at $\approx$ 6 m/s). The fast motion is accuratelly predicted as seen in Figure 6. To test the viewpoint invariance of our system, an overhead trial was conducted, resulting in the pose estimates in Figure 7. Even given limited cues from the silhouette, the system is able to infer the horizontal waving of an arm. Notice that the arm motion is smooth throughout most of the sequence.

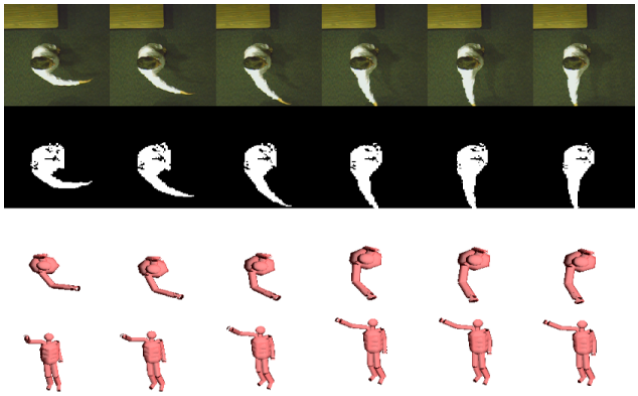


Figure 7. A sequence of pose estimates for tracking a reaching motion. Observed silhouettes (second from top) can be compared with our best pose estimate from the camera view (second from bottom) and from a frontal view (bottom).

4.2. Towards Dynamical Activity Recognition

Using the same silhouette trials, we measured the ability of our system to recognize performed activities. In the current system, activity is recognized as the pose estimate likelihoods normalized over all of the primitives into a probability distribution, as shown in Figure 8d. Preliminary results of the progress of the most likely particle in each motion primitive are shown in Figure 8a and b.

5. Conclusion

This paper presented a modular methodology for monocular tracking and activity recognition using a vocabulary of nonlinear latent space dynamics derived from human performance. A probability density function in the form of a "bending cone" is proposed in order to deal with the nonlinearities inherent in the models of human performance. The use of this probability density function allows tracking of movements whose speed of performance is different from the movements described by the motion primitives, ranging from slow to fast motions. Pose information inferred by the tracking system and its temporal evolution along the primitives is used to estimate the activity the subject is performing. The robustness of the algorithm is tested with respect to the viewpoint of the camera and the speed of the activity. The algorithm presented runs at a rate of a frame per second in a modern AMD64 3Ghz computer.

References

[1] A. Agarwal and B. Triggs. Learning to track 3d human motion from silhouettes. In *International Conference on Machine Learning*, page 2, 2004. 2

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999. 1

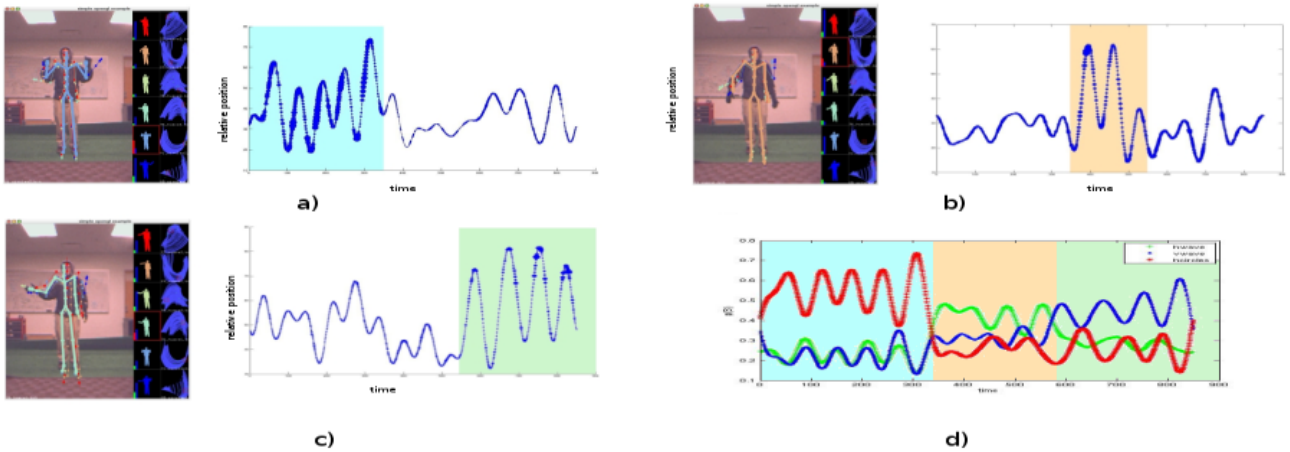


Figure 8. An evaluation of our activity recognition system over time. Attractor progress for (a) hand circles actvity, (b) vertical hand-waving motion, (c) horizontal hand waving activity exhibits a sinusoidal pattern during their performance. (d) Evolution of the likelihood for these three primitives appropriately classifies the movement trial into activities, idealizing the firing of a mirror neuron.

[3] O. Arikan, D. A. Forsyth, and J. F. O'Brien. Motion synthesis from annotations. *ACM Transactions on Graphics*, 22(3):402–408, 2002. 2

[4] D. Baraff and A. Witkin. Physically-based modeling, principles and practice. Technical Report Course 34, SIGGRAPH 1997 Course Notes, ACM, Aug 1997. 2

[5] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition*, pages 2239–2245, 1999. 2

[6] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Computer Vision and Pattern Recognition*, pages 126–133, 2000. 2, 3

[7] A. M. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Computer Vision and Pattern Recognition*, pages 681–688, 2004. 2

[8] 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. 3

[9] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998. 2, 3

[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. 1, 2, 3

[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. 2

[12] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001. 1

[13] F. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Phil. Trans. R. Soc. Lond. B*, 355:1755–1769, 2000. 2

[14] E. Ong, A. Hilton, and A. Micilotta. Viewpoint invariant exemplar-based 3d human tracking. In *ICCV Modeling People and Human Interaction Workshop*, 2005. 2, 3, 4

[15] V. Pavlovic, J. M. Rehg, T.-J. Cham, and K. P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *ICCV*, pages 94–101, 1999. 2

[16] D. Ramanan and D. A. Forsyth. Automatic annotation of everyday movements. In *Neural Info. Proc. Systems*, Vancouver, Canada, 2003. 2

[17] G. Rizzolatti, L. Gadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3:131–141, 1996. 2

[18] C. Rose, M. F. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics & Applications*, 18(5):32–40, 1998. 1, 2

[19] H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV (1)*, pages 784–800, 2002. 2

[20] 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. 2

[21] C. Sminchisescu and B. Triggs. Building roadmaps of minima and transitions in visual models. *International Journal of Computer Vision*, 61(1):81–101, 2005. 4

[22] 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. 2

[23] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005. 2, 3

[24] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking from small training sets. In *International Conference in Computer Vision*, October 2005. 1, 2, 3

[25] J. Wang, F. David, and H. Aaron. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems 18*, pages 1441–1448. 2006. 2