Getting Humanoids to Move and Imitate

Maja J. Matarić, University of Southern California

THIS BEHAVIOR-BASED APPROACH TO STRUCTURING AND CONTROLLING COMPLEX ROBOTIC SYSTEMS USES IMITATION FOR INTERACTION AND LEARNING. THE USE OF BASIS BEHAVIORS, OR PRIMITIVES, LETS THESE HUMANOID ROBOTS DANCE THE MACARENA, DO AEROBICS, AND THROW.

AS ROBOTS INCREASINGLY BEcome part of our everyday lives, they will serve as caretakers for the elderly and disabled, assistants in surgery and rehabilitation, and educational toys. But for this to happen, programming and control must become simpler and human–robot interaction more natural. Both challenges are particularly relevant to humanoid robots, which are highly difficult to control yet most natural for interaction with people and operation in human environments.

As this article shows, we have used biologically inspired notions of behavior-based control to address these challenges at the University of Southern California's Interaction Lab, part of the USC Robotics Research Labs. By endowing robots with the ability to imitate, we can program and interact with them through human demonstration, a natural human–humanoid interface. The human ability to imitate—to observe and repeat behaviors performed by a teacher—is a poorly understood but powerful form of skill learning. Two fundamental open problems in imitation involve interpreting and understanding the observed behavior and integrating the visual perception and movement control systems to reconstruct what was observed.

Our research has a similarly twofold goal: we are developing methods for segmenting and classifying visual input for recognizing human behavior as well as methods for structuring the motor control system for general movement and imitation-learning capabilities. Our approach brings these two pursuits together much as the evolutionary process brought them together in biological systems.[1,2] We structure the motor system into a collection of movement primitives, which then serve both to generate the humanoid's movement repertoire and to provide prediction and classification capabilities for visual perception and interpretation of movement. This way, what the humanoid can do helps it understand what it sees and vice versa. The more it sees, the more it learns to do, and thus the better it gets at understanding what it sees for further learning; this is the imitation process.

Behavior-based robotics

Our work over the last 15 years has focused on developing distributed, behavior-based methods for controlling groups of mobile robots and, most recently, humanoids. Behavior-based control involves the design of control systems consisting of a collection of behaviors.[3] Behaviors are real-time processes that take inputs from sensors (such as vision, sonar, or infrared) or other behaviors and send output commands to effectors (wheels, motors, or arms) or other system behaviors. The controller then is a distributed network of such communicating, concurrently executed behaviors, typically with excellent real-time and scaling properties. The interaction of the behaviors through the environment produces the desired overall system performance.

The inspiration for behavior-based control comes from biology, where natural systems are believed to be similarly organized, from spinal reflex movements up to more complex behaviors such as flocking and foraging.[4,5] We have focused on applying these principles to high-dimensional, behavior-based systems such as those involved in controlling individual and group behavior of mobile and humanoids robots. In all problem domains, we have used basis behaviors, or *primitives*, to structure and simplify the control problem, as well as to enable adaptation and learning.

Basis behaviors and primitives. Several methods for principled behavior design and coordination are possible.[3] In 1992, we introduced the concept of *basis behaviors*, a small set of necessary and sufficient behaviors that could be composed (by sequencing or superposition) for handling controller complexity and simplifying robot programming. Basis behaviors are the primitives that serve as a substrate for control, representation, and learning in our behavior-based systems. In first demonstrating their effectiveness on groups of mobile robots, we used a basis set consisting of avoidance, following, homing, aggregation, and dispersion, which served to demonstrate higher-level group behaviors including flocking, foraging and collection, and herding.[5] We also demonstrated how, given such a basis behavior set, a learning algorithm could improve behavior selection over time.

Collections of behaviors are a natural representation for controlling collections of robots. But how can we use the same idea in the humanoid control domain, where the body's individual degrees of freedom are more coupled and constrained? For this, we have combined the notion of primitives with another line of evidence from neuroscience—*mirror neurons*—to structure humanoid motor control into a general and robust system capable of a variety of skills and learning by imitation.[6]

Humanoid control and imitation. Robot control is a complex problem, involving sensory and effector limitations and various forms of uncertainty. The more complex the system to be controlled, the more we must modularize the approach to make control viable and efficient. Humanoid agents and robots are highly complex; a human arm has seven degrees of freedom (DOF), the hand has 23, and the control of an actuated human spine is beyond current consideration. Yet humans display complex dynamic behaviors in real time and learn various motor skills throughout life, often through imitation.

Methods for automating robot programming are in high demand. Reinforcement learning, which lets a robot improve its behavior based on trial-and-error feedback, is very popular. However, reinforcement learning is slow, as the robot must repeatedly try various behaviors in different situations. It can also jeopardize the robot. In contrast, learning by imitation is particularly appealing because it lets the designer specify entire behaviors by demonstration, instead of using low-level programming or trial and error by the robot.

In biological systems, imitation appears to be a complex learning mechanism that involves an intricate interaction between visual perception and motor control, both of which are complex in themselves. Although various animals demonstrate simple mimicry, only a very few species, including humans, chimps, and dolphins, are capable of so-called *true imitation*, which involves the ability to learn arbitrary new skills by observation.[7] This suggests a complex mechanism, which has been implicated as the basis for gestural communication and even language evolution.[8] The origins of such a mechanism appear to lie in an evolutionarily older system that combines perception and motor control and enables mimicry—that is, the mirror system.[6]

Neuroscience inspiration

Evidence from neuroscience studies in animals points to two neural structures we find of key relevance to imitation: spinal fields and mirror neurons. Spinal fields, found in frogs and rats so far, code for complete primitive movements (or behaviors), such as reaching and wiping.[4] More interestingly, they are additive; when multiple fields are stimulated, the resulting movement is a meaningful combination. Because the spine encodes a finite number of such fields, they represent a basis set of primitives and were precisely the inspiration for our work on basis behaviors.

Investigators recently found neurons with so-called mirror properties in monkeys and humans. They appear to directly connect the visual and motor control systems by mapping observed behaviors, such as reaching and grasping, to motor structures that produce them.[6] How many such movements the mirror system directly maps is still unknown, but the basic idea serves as rich inspiration for structuring a robotic imitation system.

We combine these two lines of evidence, spinal basis fields and mirror neurons, into a more sophisticated notion of behaviors, or *perceptual-motor primitives*. These let a complex system, such as a humanoid, recognize, reproduce, and learn motor skills. The primitives serve as the basis set for generating movements, but also as a vocabulary for classifying observed movements into executable ones. Thus, primitives can classify, predict, and act.

In our approach to imitation, the vision system continually matches any observed human movements onto its own set of motor primitives. The primitive, or combination of primitives, that best approximates the observed input also provides the best predictor of what the robot expects to observe next. This expectation facilitates visual segmentation and interpretation of the observed movement. Imitation, then, is a process of matching, classification, and prediction. Learning by imitation, in turn, creates new skills as novel sequences and superpositions of the matched and classified primitives. The hierarchical structure of our imitation approach lets the robot initially observe and imitate a skill, then perfect it through repetition, so that the skill becomes a routine and itself a primitive. The set of primitives can thus adapt over time, to allow for learning arbitrary new skills—that is, for "true" imitation.

Figure 1 shows the structure of our imitation architecture, including the visual perception and attention module, the classification module, and the motor primitives.[1] The learning component allows adaptation both at the classification level, for finding a closer
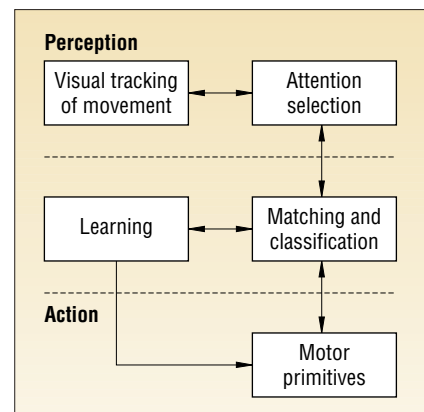

Figure 1. Our imitation architecture, structured around perceptual-motor primitives used for movement segmentation, classification, and generation.

match to the observed behavior, and the movement generation level, for optimizing and smoothing the performance.

Choosing the primitives

Movement primitives or behaviors are the unifying mechanism between visual perception and motor control in our approach, and choosing the right ones is a research challenge, driven by several constraints. On the one hand, the motor control system imposes physical bottom-up limitations, based on its kinematic and dynamic properties. It also provides top-down constraints from the type of movements the system is expected to perform, because the primitives must be sufficient for the robot's entire movement repertoire. On the other hand, the visual system's structure and inputs influence the choice of primitives for mapping the various observed movements onto its own executable repertoire.

To serve as a general and parsimonious basis set, the primitives encode groups or classes of stereotypical movements, invariant to exact position, rate of motion, size, and perspective. Thus, they represent the generic building blocks of motion that can be implemented as parametric motor controllers. Consider a primitive for reaching. Its most important parameter is the goal position of the end point—that is, the hand or held object. It might be further parametrized by a default posture for the entire arm. Such a primitive lets a robot reach toward various goals within a multitude of tasks, from grasping objects and tools, to dancing, to writing and drawing. We used just such a reaching primitive in our experiments to reconstruct the popular dance, the Macarena.[2]

This approach to motor control stands in sharp contrast to the explicit planning approach to controlling robot manipulators, which computes trajectories at runtime, whenever they are needed. While fully general, on-demand trajectory generation is computationally expensive and potentially slow. In our approach, instead of computing trajectories over again, stereotypical trajectories are built in as well as learned, then looked up and parameterized for the specific task at hand. The notion of primitives capitalizes on the fact that it is simpler to learn and reuse an approximation of the inverse kinematics for specific areas in the workspace or a specific trajectory than it is to compute them anew each time.

What constitutes a good set of primitives? We have experimented with two types: innate and learned. Innate primitives are user-selected and preprogrammed. We have demonstrated the effectiveness of a basis set consisting of three types:

- discrete straight-line movements of subsets of degrees of freedom, accounting for reaching-type motions;
- continuous oscillatory movements of subsets of DOFs, accounting for repetitive motions;[9] and
- postures, accounting for large subsets of the body's DOFs.[2]

Our approach computes the learned primitives directly from human movement data. We gather different types of such data, using the following methods: vision-based motion tracking of the human upper body (using our tracking system), magnetic markers on the arm (using the FastTrak system), and full-body joint angle data (using the Sarcos Sensuit). We first reduce the dimensionality of the movement data by employing principal components analysis, wavelet compression, and correlation across multiple DOF. Next, we use clustering techniques to extract patterns of similar movements in the data. These clusters or patterns form the basis for the primitives; the movements in the clusters are generalized and parameterized, to result in primitives for producing a variety of similar movements.

Visual classification into primitives. Visual perception is also an important constraint on the primitives and a key component of the imitation process. Because the human (and humanoid) visual attention is resource-limited, it must select the visual features that are most relevant to the given imitation task. Determining what those features are for a given demonstration is a challenging problem.

Our previous work showed that people watching videos of arm movements displayed no difference in attention whether they were just watching or intending to subsequently imitate. In both cases, they fixated at the end point—the hand or a held object.[10] However, classifying all possible end-point trajectories into a useful set of task-independent categories or primitives is impossible. Fortunately, this is not necessary, because the imitation system is targeted for mapping observed movement of bodies similar to the observer's own motor repertoire. Accordingly, the mirror system is sensitive to the biological motion of similar bodies (for example, monkeys respond to monkey and human movements). Furthermore, although human(oid) bodies are potentially capable of vast movement repertoires, the typical, everyday movement spectrum is not nearly as large.

Consequently, we can effectively bias the visual perception mechanism toward recognizing movements that it can execute, especially those movements it performs most frequently. The motor control system's structure, and its underlying set of movement primitives, provides key constraints for visual movement recognition and classification. Our primitive classifier uses the primitives' descriptions to segment a given motion based on the movement data. In the experiments described below, we used end-point data for both arms as input for the vector quantization-based classifier.[11] Again, a key issue in classification is representing the primitives such that they account for significant invariances, such as position, rotation, and scaling.

Our classification approach forms the original motion into a vector of relative end-point movements between successive frames, then smoothes and normalizes it. At the classification level, we ignore all other information about the movement, such as global position and arm configuration, enabling a small set of high-level primitive representations instead of a potentially prohibitively large set of detailed ones. Other information necessary for correct imitation serves for parameterizing the selected primitives at the level of movement reconstruction and execution.

To simplify matching, our approach describes primitives themselves in the same normalized form. For each time step of the observed motion, we compare a fixed-

WE CAN EFFECTIVELY BIAS THE VISUAL PERCEPTION MECHANISM TOWARD RECOGNIZING MOVEMENTS THAT IT CAN EXECUTE, ESPECIALLY THOSE MOVEMENTS IT PERFORMS MOST FREQUENTLY.

horizon window to every primitive and select the one that best matches the input. Adjacent windows with identical classifications connect to form continuous segments. For any segments that fail to match the given primitives, our approach uses the reaching primitive to move the end point frame by frame. Because the horizon window is of fixed size, the perception of a distinct match of a primitive applies only for the given timescale. We are currently working on addressing classification at multiple timescales.

To validate our approach, we implemented various examples of imitation, including reaching, ball throwing, aerobics moves, and dance, all on humanoid testbeds, taking human demonstrations as input.[12,13]

Using human movement data. We used several types of human movement data as input to test our approach to imitation, including visual motion tracking, Cartesian magnetic markers, and joint angle data. For vision-based motion tracking, we developed a motion-tracking system that extracts features from a video stream. The system is simple, since it can easily be replaced by a more sophisticated commercial version. It relies on a constrained (unoccluded and unambiguous) initial position and kinematic model of the object or body being imitated. This approach greatly simplifies finding the initial match between the features in the visual image and the observed body. The match enables tracking of the body over time, allowing for fast computation and updating of current limb position, as well as simple prediction of future position, used in turn to speed up recognition. Our ongoing work will address how the primitives themselves also provide further predictive capability for the visual tracking. Figure 2 shows a snapshot of the output of the vision-based motion-tracking system.

We also used 3D magnetic marker data from the human arm, gathered from subjects imitating videos of arm movements while wearing FastTrak markers for position recording. (These data were gathered at the National Institutes of Health Resource for the Study of Neural Models of Behavior at the University of Rochester.) We used four markers: near the shoulder, the elbow, the wrist, and the start of the middle finger. The movement data resulting from this experiment serve as input into our imitation system, as well as for automatically learning the primitives.

Finally, we used full-body joint angle data gathered with the Sarcos Sensuit, a wearable exoskeleton that simultaneously records the joint positions of 35 DOF: the shoulders, elbows, wrists, hips, knees, ankles, and waist. (These data are obtained through a collaboration with the ATR Dynamic Brain Project at the Human Information Processing Labs in Kyoto, Japan.) We are currently focusing on reproducing the upper-body movements from those data on our testbeds, described next.

Evaluation testbeds

To properly validate our approach to humanoid motor control and imitation, we use different experimental testbeds. Most of our work so far has been done on Adonis (developed at the Georgia Institute of Technology Animation Lab), a 3D rigid-body simulation of a human with full dynamics (Figure 3). Mass and moment-of-inertia information comes from the graphical body and human density estimates, and the motion equations are calculated using a commercial solver, SD/Fast. The simulation acts under gravity and accepts external forces from the environment. The static ground contact removes the need for explicit balance con-

trol. We have added joint limits and self-collision detection. Adonis consists of eight rigid links connected with revolute joints of one DOF and three DOF, totaling 20 DOF,

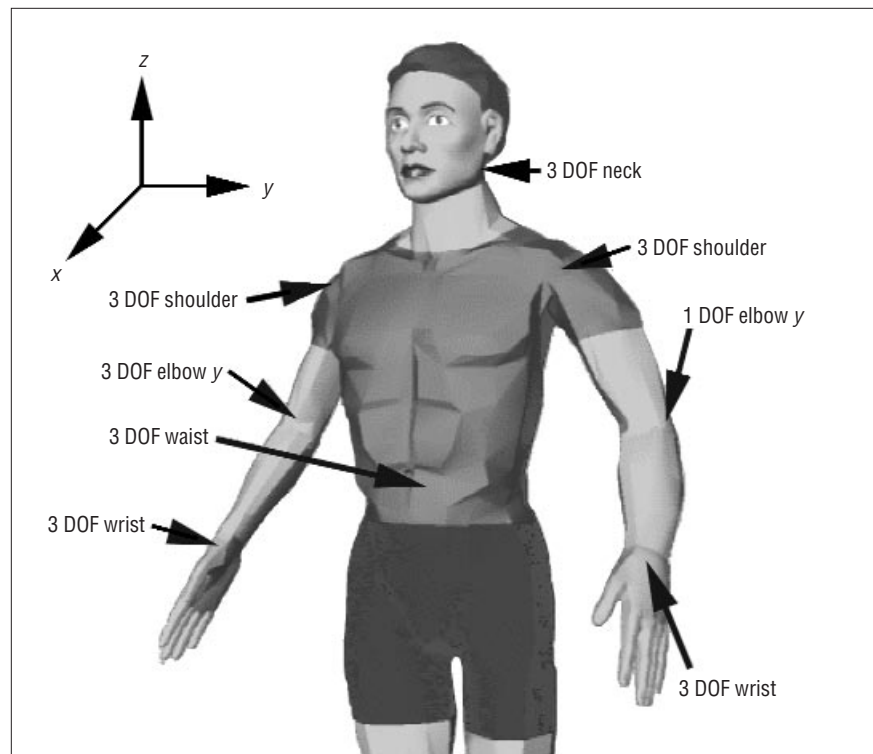Figure 2. A snapshot of the output of our vision-based motion-tracking system.

Figure 3. The Adonis 20-degree-of-freedom full-body dynamic simulation.

but most of our experiments use the 13 DOF of Adonis' upper body and currently do not address locomotion.

We also use a pair of Cosimir 3D humanoid avatars (developed at the University of Dortmund), which provide graphical simulations with 65 DOF each (Figure 4). Unlike Adonis, these avatars have actuated fingers, allowing experiments with fine motion control. However, the avatars are not directly equipped with dynamics, so we can experiment with different models of mechanics, muscle models, and so forth. As with Adonis, we do not address balance with these avatars, but anchor them at the hips. The avatars serve to imitate each other or an external demonstrator.

As the project progresses, we plan to use physical humanoid robots as the ultimate testbeds for evaluating our imitation architecture. The NASA Robonaut is one candidate, through collaboration with the Johnson Space Center. The Sarcos full-body humanoid robot is another, through collaboration with the ATR Dynamic Brain Project. Both robots are highly complex, built to approximate human body structure as faithfully as practically possible, and feature binocular cameras for embodied visual perception critical for imitation.

had limitations for particular types of movements. This led us to propose and explore a combination approach, where multiple types of primitives can be sequenced and combined. For example, we constructed a basis behavior set consisting of three types of primitives:

- discrete straight-line movements using impedance control;
- continuous oscillatory movements using coupled oscillators (or a collection of piece-wise linear segments using impedance control); and
- postures, using PD-servos to directly control the joints.

We also added a forth type of primitive, for avoidance, implemented as a repulsive vector field. The continuously active fourth primitive combined with whatever other primitive was executing to prevent any collisions
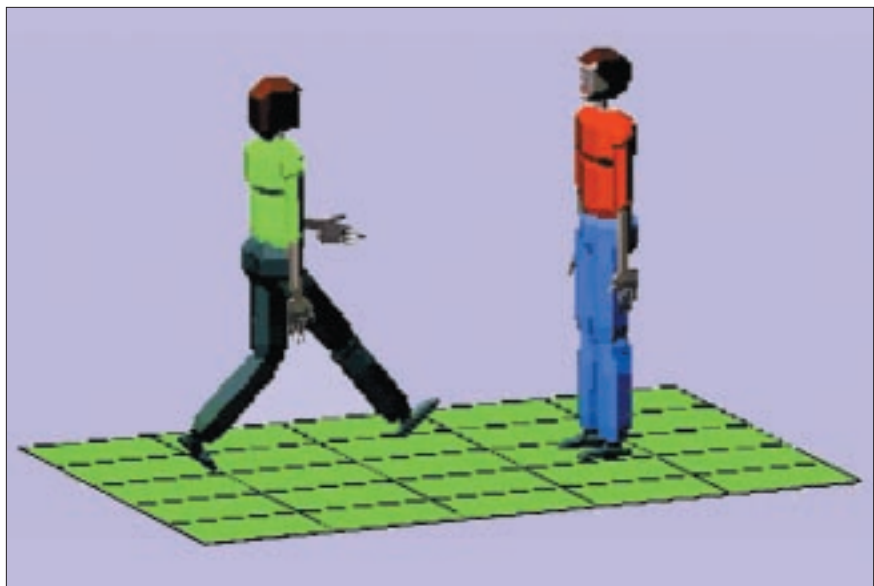

Figure 4. The Cosimir 65-degree-of-freedom full-body avatars.

OUR APPROACH TO HUMANOID motor control and imitation relies on the use of a set of movement primitives. We have experimented with different types of such primitives on different humanoid simulation testbeds. Specifically, we have implemented two versions of the spinal fields found in animals. One closely modeled the frog data, and used a joint-space representation—it controlled individual joints of Adonis's arms.[2] The other used another biologically inspired approach, impedance control,[14] which operates in the external coordinate frame—in our case, each of Adonis's hands. Our impedance motor controller applied forces at the hands and dragged the rest of the arm along. We also used a default posture for the arm, which provided natural-appearing, whole-arm movements that reached the desired hand destination.

We tested both types of primitives on a sequential motor task, dancing the Macarena. Both proved effective, but each
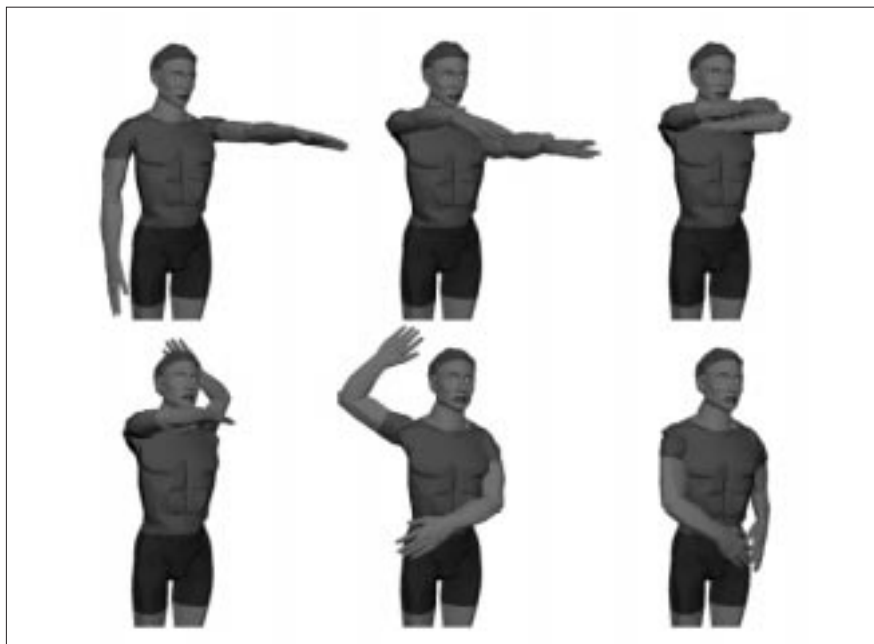

Figure 5. Snapshots of Adonis dancing the Macarena.

between body parts. In the Macarena, for example, this is necessary for arm movements around and behind the head (Figure 5).

These results validated the notion of primitives and served as a basis for our imitation architecture. Primitives provide higher-level movement descriptions; metric information serves to parameterize those descriptions for generating executable movement. This allows a small number of general primitives to represent a large class of different movements, such as reaches to various destinations on and around the body.

Our goal is not to achieve perfect, completely precise, high-fidelity imitation. While that might be possible, through the use of exact quantitative measurements of the observed movement using signal-processing techniques, it is not what happens in imitation in nature and it is neither necessary nor helpful for our main goals: natural interaction and programming of robots. For those purposes, we aim for an approximation of the observed behavior, one that allows any necessary freedom of interpretation by the humanoid robot, but achieves the task and effectively communicates and interacts with the human.

Our goal is also distinct from *task-level imitation*, which only achieves the demonstration's goal, but does not imitate the behaviors involved. This problem has been studied in assembly robotics, where investigators used a robotic arm to record, segment, interpret, and then repeat a series of visual images of a human performing an object-stacking task.[15] While that work focused on visual segmentation into subgoals, we aim for a more biologically motivated model of imitation: a system capable of imitating the observed behavior—that is, the process that brings the subgoals about, with sufficient, but not perfect, accuracy.[12]

Because our current movement primitives represent whole-arm movements, our visual system either can use only the end-point (hand) location for imitation or can gather higher-fidelity data that include the position of the elbow and other markers. In the experiments described here, the position of the hand over time sufficed for effective imitation. The system segments hand trajectory over time using the classifier, which, at each point in time, matches the expected output of each of the primitives with the observed input and selects the best match. Consecutive matches of the same primitive indicate a higher confidence in the match. The classification output is a sequence of primitives and their associated parameters. These then go to the motor control system and activate the primitives in turn (and avoidance in parallel) to reconstruct the observed behavior.[12,13]

In validating the imitation architecture, we selected tasks from athletics, dancing, and aerobics; six tasks were performed by a human demonstrator in front of the vision system and presented to our imitation system. Figure 6 shows an example of the system performance on three human-demonstrated movements: a throw, a twist, and a raising of the arms.[12,13] We are currently pursuing methods for evaluating the quality of the resulting imitation, both in terms of the produced end-point trajectories and the overall qualitative appearance. Additional information about this research program and videos demonstrating the performance of the imitation system are found at robotics.usc.edu/~agents/imitation.html. ◼
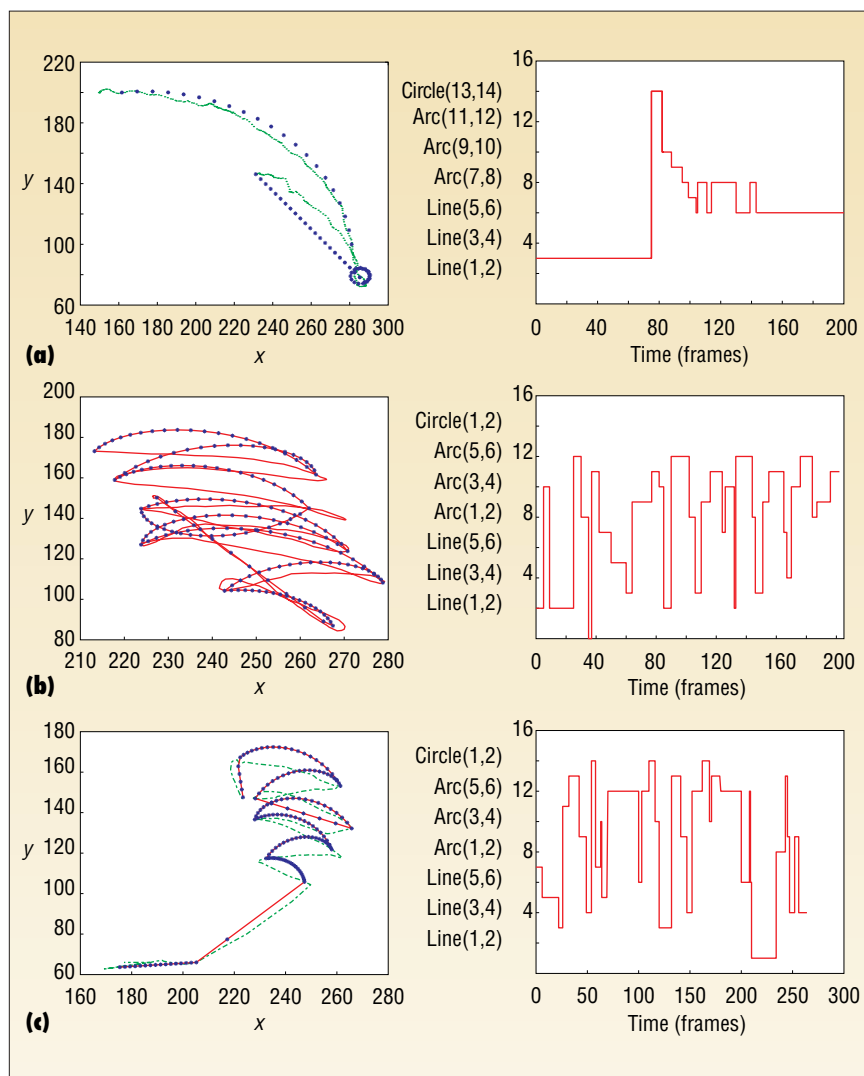

Figure 6. A selection of imitation results showing (a) a throw, (b) twists, and (c) arm raises. On the left, plots of the human-demonstrated movements and reconstructed motions; on the right, the classification results for each.

References

1. M.J. Matarić, "Sensory-Motor Primitives as a Basis for Imitation: Linking Perception to Action and Biology to Robotics, *Imitation in Animals and Artifacts*, C. Nehaniv and K.

Dautenhahn, eds., MIT Press, Cambridge, Mass., 2000.

2. M.J. Matarić, V.B. Zordan, and M. Williamson, "Making Complex Articulated Agents Dance: An Analysis of Control Methods Drawn from Robotics, Animation, and Biology," *Autonomous Agents and Multiagent Systems,* Vol. 2, No. 1, Mar. 1999, pp. 23–44.

3. R.C. Arkin, *Behavior-Based Robotics*, MIT Press, Cambridge, Mass., 1998.

4. E. Bizzi, F.A. Mussa-Ivaldi, and S. Giszter, "Computations Underlying the Execution of Movement: A Biological Perspective," *Science*, Vol. 253, July 1991, pp. 287–291.

5. M.J. Matarić, "Designing and Understanding Adaptive Group Behavior," *Adaptive Behavior,* Vol. 4, No. 1, Dec. 1995, pp. 50–81.

6. G. Rizzolatti et al., "Premotor Cortex and the Recognition of Motor Actions," *Cognitive Brain Research,* Vol. 3, 1996, pp. 131–141.

7. R.W. Byrne and A.E. Russon, "Learning by Imitation: A Hierarchical Approach," *J. Behavioral and Brain Sciences,* Vol. 6, No. 3, 1998.

8. M. Arbib, "The Mirror System, Imitation, and the Evolution of Language," *Imitation in Natural and Artificial Systems*, K. Dautenhahn and C. Nehaniv, eds., MIT Press, Cambridge, Mass., 2000.

9. A. Billard and M.J. Matarić, "Learning Motor Skills by Imitation: A Biologically Inspired Robotic Model," *Proc. Autonomous Agents 2000,* ACM Press, New York, 2000, pp. 373–380.

10. M.J. Matarić and M. Pomplun, "Fixation Behavior in Observation and Imitation of Human Movement," *Cognitive Brain Research*, Vol. 7, No. 2, 1998, pp. 191–202.

11. S. Arya and D.M. Mount, "Algorithms for Fast Vector Quantization," *Proc. Data Compression Conf.,* IEEE Press, Piscataway, N.J., 1993, pp. 381–390.

12. S. Weber, M.J. Matarić, and O.C. Jenkins, "Experiments in Imitation Using Perceptuo-Motor Primitives," *Autonomous Agents,* ACM Press, New York, 2000, pp. 136-137.

13. O.C. Jenkins and M.J. Matarić, *Primitive-Based Movement Classification for Humanoid Imitation*, Tech. Report IRIS-00-385, Univ. of Southern California, Inst. for Robotics and Intelligent Systems, 2000.

14. N. Hogan, "Impedance Control: An Approach to Manipulation," *J. Dynamic Systems, Measurement, and Control,* Vol. 107, Mar. 1985, pp. 1–24.

15. Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by Watching, Extracting Reusable Task Knowledge from Visual Observation of Human Performance," *IEEE Trans. Robotics and Automation,* Vol. 10, No. 6, 1994, pp. 799–822.

Maja J. Matarić is an assistant professor in the Computer Science Department and the Neuroscience Program at the University of Southern California, the director of the USC Robotics Research labs, and an associate director of the Institute for Robotics and Intelligent Systems. Her group performs research into behavior-based control and learning in multirobot systems and humanoids. She received her BS in computer science from the University of Kansas, her MS in computer science from MIT, and her PhD in computer science and artificial intelligence, also from MIT. She has received the NSF Career Award, the MIT Technology Review TR100 Innovation Award, and the IEEE Early Academic Career Award in Robotics and Automation. Contact her at the Computer Science Dept., Univ. of Southern California, 941 W. 37th Place, SAL 228, Mailcode 0781, Los Angeles, CA 90089-0781; mataric@cs.usc.edu; http://robotics.usc.edu/~maja.