

Learning to Perform Actions Through Multimodal Interaction

Chen Yu, Xue Gu and Dana H. Ballard

Department of Computer Science
University of Rochester
Rochester, NY, 14620
{yu,xgu,dana}@cs.rochester.edu

Anthropomorphic robots are expected to collaborate with users as human-like assistants. For instance, users can control the behaviors of robots by speech. When a user sends a spoken command “pick up the apple on the desk”, a robot can perform the corresponding action and bring the apple to the user. To achieve this goal, the robot needs to acquire two basic sensorimotor skills. First, it needs to map action verbs (linguistic labels in speech) with corresponding actions. Second, it needs to know how to perform those actions. How does the robot acquire those skills? A new approach is proposed by (Weng *et al.* 2001) in which a brain-like artificial embodied system develops and learns based on real-time interactions with the environments by using multiple sensors and effectors. This work presents the first steps toward this kind of autonomous learning. Specifically, the embodied system described in this paper is able to acquire basic skills through natural interactions with users.

Learning to associate action verbs with hand motion trajectories

To ground action verbs in body movements, an embodied system needs to have sensorimotor experiences by interacting with the physical world. Our solution is to attach different kinds of sensors to a real person to share his/her sensorimotor experiences. Those sensors include a head-mounted CCD camera to capture a first-person point of view, a microphone to sense acoustic signals, an eye tracker to track the course of eye movements that indicate the agent’s attention, and position sensors attached to the head and hands of the agent to simulate proprioception in the sense of motion. The functions of those sensors are similar to human sensory systems and they allow the computational system to collect user-centric multisensory data to simulate the development of human-like perceptual capabilities. In the learning phase, the human agent performs some everyday tasks, such as making a sandwich, pouring some drinks or stapling a letter, while describing his/her actions verbally. We collect acoustic signals in concert with user-centric multisensory information from non-speech modalities, such as user’s perspective video, gaze positions, head directions and hand movements. A multimodal learning algorithm is developed that

first spots words from continuous speech and then builds the grounded semantics by associating action verbs with body movements.

It has been shown that eye and head movements are closely related to the requirements of motor tasks and almost every action in an action sequence is guided and checked by vision, with eye and head movements usually preceding motor actions. We develop a method that utilizes eye gaze and head position information to detect the performer’s focus of attention. Attention, as represented by eye fixation, is used for spotting the target object related to the action. Attention switches are calculated and used to segment the hand motion sequence into action primitives. Next, a temporal sequence of feature vectors are extracted from each action unit. Let S denote a hand motion trajectory that is a multivariate time series spanning n time steps such that $S = \{s_t \mid 1 \leq t \leq n\}$. s_t is a vector of values containing one element for the value of each of the component univariate time series at time t . Given a set of m multivariate time series of hand motion, we want to obtain in an unsupervised manner a partition of these time series into subsets such that each cluster corresponds to a qualitatively different action types. Our clustering approach is based on the combination of a Mixture hidden Markov Model and Dynamic Time Warping (DTW) (Yu & Ballard 2004).

Next, we utilize the co-occurrence of multimodal data to select meaningful semantics that associate body movements with spoken words. We take a novel view of this problem as the word correspondence problem in machine translation. For example, body movements can be looked as a “body language”. Thus, associating body movements with action verbs can be viewed as the problem of identifying word correspondences between English and “body language”. In light of this, we apply a technique from machine translation to address this problem. We model the probability of each word as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, the Expectation-Maximization (EM) algorithm is employed to find the reliable associations of spoken words and their grounded meanings which maximize the likelihood function of observing the data. Finally, the embodied learning system stores the grounded word-meaning pairs represented by phoneme sequences and hand motion types. Technical details can be found in (Yu & Ballard 2004).

Action Generation

The previous section describes how the robot recognizes and categorizes hand motion sequences (end effector trajectories) and associate them with linguistic labels (action verbs). This section presents how to re-generate actions based on observing others' end effector trajectories. This action generation process needs to transform the data from the Cartesian space to the robot joint space, where control actually occurs. Human arms have many more intrinsic degrees of freedom (DOF) than the external Cartesian space. For instance, in the action of arm reaching, the external space includes 3 dimensions, while the intrinsic control space includes 7 DOFs. This redundancy gives the human body flexibility, but greatly increases the difficulty of developing the underlying control mechanism.

This work extends the computational model proposed by (Torres & Zipser 2002). We define two objective functions that need to be minimized during action generation. The first function provides an inverse kinematics solution in the joint space given a trajectory in the Cartesian space. The second objective function is used to solve the redundancy problem, which picks up a biologically reasonable solution among infinite number of options. Gradient descent is applied to those two objective functions to generate arm motions along the trajectory.

The first function is defined as the Euclidean distance from the current end effector position to the next step trajectory point.

$$r(\theta(t), x(t+1)) = \sqrt{\sum_{i=1}^3 (x_i(t+1) - f_i(\theta(t)))^2} \quad (1)$$

where $x(t)$ represents the end effector trajectory and $\theta(t)$ are joint space configurations. f is the function that can uniquely map the joint configuration to an end effector point in the Cartesian space. The gradient of equation (1) is as follows:

$$d\theta(t) = -\frac{(x(t+1) - f(\theta(t))) \times J(f(\theta(t)))}{\sqrt{\sum_{i=1}^3 (x_i(t+1) - f_i(\theta(t)))^2}} \quad (2)$$

where J is a Jacobian matrix. Equation (2) specifies both the direction and the speed of the movement along the solution path.

In order to solve the redundancy problem, we introduce the second objective function to simulate human's tendency to save energy by picking up certain postures during action generation. The energy is described in term of the potential energy defined as follows:

$$E(\theta(t)) = m_1gh_1(\theta(t)) + m_2gh_2(\theta(t)) + m_3gh_3(\theta(t))$$

where h_1 , h_2 and h_3 are the heights of all three centers of mass of upper arm, forearm and hand respectively, and m_1 , m_2 and m_3 are the masses of three arm links.

Now that we have two objective functions that need to be minimized at the same time. The movement can be described as to bring the hand along the action trajectory, while at the same time to save energy by favoring certain postures. Our control strategy is to combine the gradient of those two objective functions as follows:

$$\frac{d(r + E)}{d\theta} = \alpha \times \frac{\partial r}{\partial \theta} + (1 - \alpha) \times \frac{\partial E}{\partial \theta} \quad (3)$$

Here, α is the weight factor.

Demonstration Using A Virtual Human

Since we are interested in the learning and control problems in humanoid robots, we developed a virtual human that is able to demonstrate some human-like behaviors. The virtual human is provided by Boston Dynamic's Di-Guy and graphic rendering is achieved by SGI OpenPerformer libraries so that the virtual human and a user can interact in a virtual environment. In the speech dictation mode as shown in Figure 1, the user sends spoken commands, such as "pick up the apple". The virtual agent matches the spoken words with the phoneme strings of words that he has learned before. A simple grammar allows the phrase "action + object". In this way, the virtual human recognizes the corresponding action verbs and object names. In response to speech, the virtual human then finds the corresponding hand motion trajectories according to the recognized action verbs. Next, he performs the actions based on the method described in the section of action generation.



Figure 1: The user (in upper left) sends spoken commands, and the virtual agent performs the actions accordingly in a virtual environment.

Conclusion

We present a multimodal learning system that is trained in an unsupervised mode in which users perform everyday tasks while providing natural language descriptions of their behaviors. In addition to recognize hand motion types and associate them with action verbs, the system also learns how to re-generate actions based on inverse kinematics. A real-time speech dictation system shows that the virtual human can interact with users and perform actions according to spoken commands.

References

- Torres, E., and Zipser, D. 2002. Reaching to grasp with a multi-jointed arm. i. computational model. *Journal of Neurophysiology* 88:2355–2367.
- Weng, J.; McClelland, J.; Pentland, A.; Sporns, O.; Stockman, I.; Sur, M.; and Thelen, E. 2001. Artificial intelligence: Autonomous mental development by robots and animals. *Science* 291(5504):599–600.
- Yu, C., and Ballard, D. H. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception* 1.