Statistical Imitative Learning from Perceptual Data

Tony Jebara*

Columbia University, CS
New York, NY 10027

Alex Pentland

MIT Media Lab
Cambridge, MA 02139

Abstract

Imitative learning has recently piqued the interest of various fields including neuroscience, cognitive science and robotics. In computational behavior modeling and development, it promises an accessible framework for rapidly forming behavior models without tedious supervision or reinforcement. Given the availability of low-cost wearable sensors, the robustness of real-time perception algorithms and the feasibility of archiving large amounts of audio-visual data, it is possible to unobtrusively archive the daily activities of a human teacher and his responses to external stimuli. We combine this data acquisition/representation process with statistical learning machinery (hidden Markov models) as well as discriminative estimation algorithms to form a behavioral model of a human teacher directly from the data set. The resulting system learns audio-visual interactive behavior from the human and his environment to produce an interactive autonomous agent. The agent subsequently exhibits simple audio-visual behaviors that appear coupled to real-world test stimuli.

1 Introduction

Imitative learning provides an easy approach [1] for learning agent behavior by using real examples of agents interacting in a world that can be learned from and generalized. The two components of this process, passively perceiving real world behavior and learning from it are portrayed in Figure 1. We propose a generative statistical model at the perceptual level to be able to regenerate or resynthesize virtual characters while keeping a discriminative model on the temporal learning to focus resources on the prediction task necessary for action selection.

This paper is organized as follows. We first situate imitative learning in the context of other agent learning approaches and also motivate it with background in related work. Subsequently, we describe a long-term behavioral data collection system and the audio-visual representations it utilizes to represent interactive behavior as a time series. Discriminatively trained Hidden Markov models (HMMs) are then proposed to extract and model the interactive behavior from the temporal dataset and synthesize imitative behavior. We conclude with experimental results and discussions.

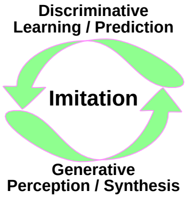

Figure 1: Imitative Learning through Discriminative Time Series Prediction and Probabilistic Perception.

2 Background

Various approaches have been proposed for learning autonomous agents in domains such as robotics and interactive graphics. These include rule-based systems where a programmer manually specifies the behavior model, supervised systems where a human labels data with appropriate output behavior, and reinforcement learning [10] systems where a human (or environment) penalizes/rewards behavior. Imitative learning[18] curiously spans both supervised and unsupervised regimes. If we collect data of real people interacting with the real world, we are shown many exemplars of appropriate reactionary behavior in response to the current context. Thus, the data is already labeled and needs no teaching effort for the supervision. This, of course, assumes that perceptual techniques can record and represent natural real-world activity automatically. In such an incarnation, imitative learning is unsupervised and only involves data collection. This makes it an attractive paradigm from an implementation point of view. Furthermore, various recent developments in other fields also motivate imitative learning as crucial to human development.

Early research in behavior and cognitive sciences ex-

[*] Corresponding author: jebara@cs.columbia.edu

[1] As Confucius says, there are 3 types of learning, "by reflection, which is noblest; second, by imitation, which is easiest; and third by experience, which is the bitterest".

hibited strong interest in the role of imitative learning. However, ground breaking works of Thorndike [19] and Piaget [15] were followed by a lull in the area of movement imitation. This was in part due to the presumption that imitation or mimicry in an entity was not necessarily the sign of higher intelligence and therefore not critical to development[18]. This prejudice slowly faded with the arrival of several studies by Meltzoff and Moore that indicated infants' ability to perform facial/manual gesture imitation from ages 12-21 days old and in some cases at an hour old [14]. Imitative learning began to be seen as an almost innate mechanism to help the development of humans and certain species [20]. More recently, through discoveries of mirror neurons, action-perception pathways and functional magnetic resonance imaging results [16] [5] [17], a neural basis for imitative learning has been recently hypothesized. Experiments indicated consistent firings in a mirror neuron either when an action was performed by a subject or when another individual was perceived performing it. These results have spurred applied efforts in imitative approaches to robotics by Mataric [12], Brooks [3], etc. where imitation has gained visibility and complemented reinforcement learning [10].

However, these domains have predominantly focused on uncovering direct mappings between action and perception [18] [12]. It is through such a mapping that the imitation learning problem can be translated into a direct supervised learning scenario. This complex mapping is to a certain extent the Achilles' heel of imitation learning and extensive effort in humanoid robot imitation rests in resolving Meltzoff and Moore's so-called 'Active Intermodal Mapping' (AIM) problem. That is, the creation of a mapping of the visual perception of a teacher's movement to high-level representations that can then be matched to other high-level representations of the learner's action space and proprioceptive senses.

An alternative approach is to do away with the AIM problem altogether by either providing the teacher's perceptual data in terms of the action-space of the learner [21] [13] or by only considering virtual characters [9] [8] whose action space is in the perceptual space. For example, Weng [21] describes a human pushing a robot down a hallway while the robot collects images of its context. The actuators in the robot (not its cameras) measure the human's displacement and therefore to imitate the human, the displacement values need only be regurgitated (under the appropriate visual context). Hogg [9] alternatively describes a vision system which obtains perceptual measurements and needs only resynthesize behavior in the visual space to generate an action virtually. Both methods cleverly avoid a direct mapping of the perception of a teacher's activity into the learner's action-space. We shall employ a similar strategy and only consider generating virtual characters that can be resynthesized on-screen directly from previous perceptual measurements.

3 Perceptual Interaction Data

The imitation framework we will describe learns an autonomous agent that is able to interact and respond appropriately to external stimulus from the world and participants within it.[2] Given the ability to perceive real behavior in humans interacting in the world, we can collect data to learn a predictive model. To resynthesize behavior, we avoid the AIM problem by simply re-rendering the perceptual measurements via a virtual audio-visual character.

To resolve the issue of maintaining a consistent long-term perception of the teacher's behavior, we propose the use of a wearable computer system. This is a convenient way to collect a sizeable amount of data while the teacher engages in natural activity and also preserves the regular conditions and point of view necessary for a non-AIM based imitative learning framework. In Figure 2 a user (or teacher) has a head mounted microphone and a camera mounted to a boom that perceives his face. This audio-video source is the perception space as well as the action space for the system. Furthermore, a camera affixed to his glasses and a microphone that is aimed outwardly track the external world context. The wearable stores this data via a small computer that transcribes both channels of audio-video signals at roughly 100 megabytes per hour. Thus, the platform provides a large data set of consistent stimulus-response data which is appropriate for imitative learning.

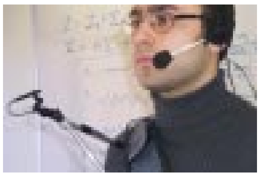

Figure 2: 2-channel A/V wearable imitation platform.

Video is stored at 7Hz from both cameras as 60 by 80 pixel RGB images. The images are illumination-normalized by histogram fitting and then filtered using a skin-color distribution model which selects only the skin-colored pixels from the images. This focuses modeling resources on the wearable user's face and focuses the external stimulus data mostly on the head and hands of people in the scene. Both audio streams are captured in real-time and represented as 200-element spectrograms (which are generated at 60Hz).

[2] This short paper is a brief overview of the imitative learning approach. More elaborate details are provided in [6].

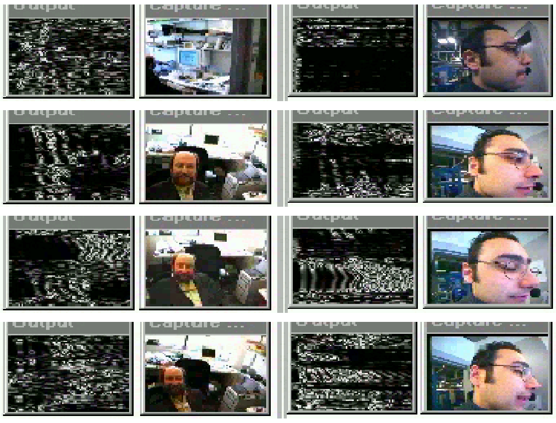

Figure 3: Wearable Interaction Data. Both the teacher and the world video are visible as well as the spectrograms for their audio signals.

Figure 3 portrays several frames of the user as he walks down a hallway and approaches an individual to begin conversation. The frames show both the user's face and his own eye's view (from the camera attached to his glasses). The spectrograms are shown adjacent to the corresponding video images.

We now discuss our representation of the above real-time audio-visual perceptual data. Effectively, the facial images (each consists of thousands of pixels) and the audio spectrograms (each consists of hundreds of frequencies) will be described by compact 20-dimensional vectors. Traditional ways to learn from and summarize complicated multi-variate data include principal components analysis (PCA), factor analysis and multidimensional scaling. There are indeed many shortcomings with such approaches, particularly that they do not take advantage of the underlying structure of the signals they deal with. Instead, data is rasterized directly into a vector form in a high-dimensional Euclidean space. However, images, audio, and time series are not vectors. A single color image is not a single big vector but rather a *collection* of small vectors or tuples (pixels). Each of these tuples is vector of 5 values, (X, Y, R, G, B) which specify XY location and RGB color. Similarly, an audio spectrogram is not a vector but rather a collection of 2-tuples (frequency and amplitude). This permits us to generate an interesting variant of PCA where each datum in our data set (images, audio) is not a vector but rather a *collection* of vectors. This modification of the model is depicted in Figure 4 as a Bayesian network.

Details of the implementation of the above model can be seen in [6]. Effectively, we interleave assignment matrix computations that solve a correspondence problem (between pixels tuples and eigenvector components) using the invisible hand algorithm [11] while we recompute PCA in an axis-parallel manner. These computations are iterated until we reach convergence. The solution typically improves on regular PCA yet does so non-monotonically and is plagued by local minima. Finding a cleaner global model/algorithm is important current research. Figure 4 depicts the model's ability to reconstruct facial images of the teacher (after skin-color based segmentation) from a 20 dimensional representation (compare it with the reconstruction of PCA directly on these images). The variant effectively has a squared reconstruction error that is up to 2.5 orders of magnitude better than PCA for the same level of dimensionality reduction. Effectively, the variant captures much more of the image structure by permitting pixel permutations.

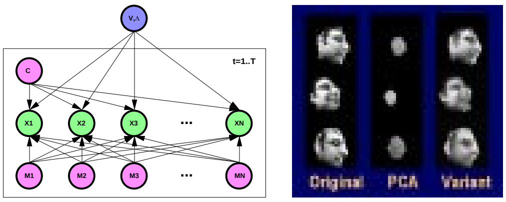

Figure 4: Bayesian Network Representing a PCA Variant for Collections of Vectors and Face Image Reconstruction

All the facial and external world images in the data set are processed as above and stored as 20-dimensional vectors each. Furthermore, this variant of PCA is applied to the agent and the external world's spectrograms to compress them each into 21-dimensional vectors as well (accuracy is again superior to PCA). The above representations generate a 20 or 21 dimensional vector for each frame of agent audio, agent video, world audio and world video. By processing the whole data set, we obtain four multidimensional time series of these coefficients. These are then time-aligned and aggregated into a large 82-dimensional time series.

4 Conditional Hidden Markov Models

Given this large time series data set, our task will be to train an imitative agent by learning a good predictive model of what the teacher will do given the external stimulus. Figure 5 depicts how the data is to be partitioned for imitative learning. The data set which spans several hours and hundreds of megabytes of images/spectrograms is initially split temporally into a training portion as well as a testing portion. We also split the 82 dimensions in half and denote the external world measurements as x (audio and video) and the agent's representation as y (audio and video). Therefore, we have $x = world$ and $y = agent$.

We now have a standard regression formulation where we need to obtain y from x by learning from the training data.

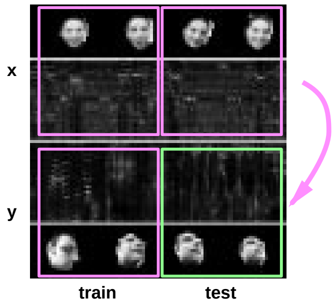

Figure 5: The Imitation Task. Using the training data, we learn a model for mapping outside world measurements x to the measurements y of the 'teacher'. The remaining unseen test data is used to evaluate the learned mapping.

It is widely known that a hidden Markov model is well-suited to time series data and can effectively model time warpings and sequential variations. Since we are interested in predicting the component of the time series (audio and video) that the agent would generate, we now have a discriminative prediction task, namely to predict y *from* x. Therefore we will employ a maximum conditional likelihood criterion to learn an HMM that specifically produces a good model of the mapping from x to y. More specifically, we have an input-output HMM structure [1] since the inputs x are related to outputs y through a hidden state s which evolves with Markovian dynamics. This HMM is trained discriminatively (or conditionally) using the CEM algorithm [6] [7]. We train one HMM to predict the agent's audio and one HMM to predict his video signals (both HMMs use the external world's A/V signal as input). Each HMM has 30 hidden states and we assume Gaussian emission models with diagonal covariance matrices.

Since we estimate the HMMs using conditional likelihood, our model's resources focus specifically on predicting the agent's behavior *from* external world stimulus. The objective function to maximize is more specifically the joint log-likelihood of an HMM over both x and y components of the time series (i.e. both inputs and outputs) *minus* the marginal log-likelihood of the HMM over the x component of the time series (input only). This process is depicted in Figure 6. This permits us to focus the learning on salient stimuli such as conversations since they result in a significant reaction from the agent. Meanwhile, episodes of the data where the agent is not expressing any interesting behavior (i.e. walking alone in the hallway, etc.) will be ignored since the external-world stimuli there do not help predict the agent's activity. In a standard maximum likelihood scenario, these irrelevant external world stimuli would get modeled undiscriminatively and quickly waste the HMM's modeling resources. Thus, we obtain more robustness and superior learning through a maximum conditional likelihood criterion [6].

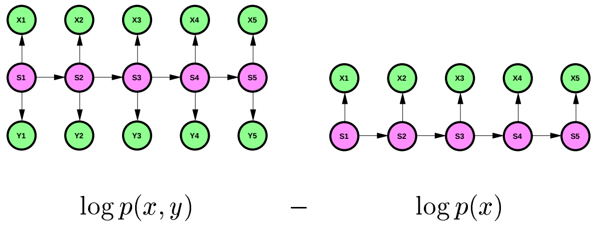

Figure 6: Conditional HMM Estimation. Effectively, we maximize the joint $p(x, y)$ log-likelihood of the HMM over both x and y (inputs and outputs) minus the marginal log-likelihood of $p(x)$ from an HMM over the x input space alone.

5 Experiments

Using the training data, the hidden Markov models were estimated with the CEM algorithm which maximizes conditional likelihood to discriminatively predict the agent's response *from* the external stimulus. For comparison, we also also trained the HMMs using the traditional EM algorithm. EM maximizes joint likelihood which is inappropriate here since the world signals x will always be measurable and need not be predicted or modeled on their own. We only care about the conditional mapping from x to y, so conditional likelihood is the more principled training and testing criterion. Table 5 depicts the conditional log-likelihoods for the HMMs on the agent's audio and video prediction tasks using novel testing data. As shown, CEM outperforms EM (note the logarithmic scale).

	EM	CEM
Audio Prediction HMM	99.61	100.58
Video Prediction HMM	-122.46	-121.26

Table 1: Conditional Log-Likelihoods on Test Data.

Following this quantitative evaluation of the models, we used the HMMs to resynthesize the learned agent behavior to verify if it qualitatively matches our expectations. Given a generative statistical model on the temporal data (i.e. our HMMs) as well as a generative model of the perceptual input (i.e. the structured variant of PCA), it is straightforward to reconstruct the audio and video that correspond to the agent and visualize the reactions and behavior the model is predicting. We used the HMMs to synthesize agent reactions to test data where only the world stimulus was measurable. This is done by first solving for the state distributions using x alone (with a marginalized HMM). Given the hidden states therein, it is straightforward to compute the expected value of the output vectors y_t at each time point by averaging the means of the Gaussian emissions weighted by their corresponding state assignment probability [2] [4].

The hidden Markov models thus provide us with a time series of predicted vectors of audio $\hat{y}_a$ and video $\hat{y}_v$ coefficients for the agent for each time step. Using these we reconstruct the original signals (images and spectrograms) by multiplying the coefficients with the eigenvectors produced by the PCA variant. Ultimately we obtain a collection of spectrogram tuples or collections of pixels. For easier visualization, we render the image in the training data that is closest to a given synthesized collection of pixels (using Euclidean distance). Spectrograms too can be inverted to play back sound, however noise arises since the phase information is lost.

We first verified the resynthesis on training data (which the HMM was originally estimated with) as in Figure 7. We see the synthetic agent initiating a conversation as he approaches an individual in the external world. The agent says “Hi” followed by the human saying “Hi”. Then the agent says “how are you” to which the human replies “fine and you”. Most of the other agent articulations are mumbles that are difficult to decipher and generally sound like “I see”, “hmm” and “yeah” except these are interleaved appropriately into the acoustics of the real human’s speech in the external world stream. Furthermore, the system animates the agent such that its head movement and visual cues coincides appropriately with the external stimuli from the real world. Unfortunately, the agent also says what sounds like “hi” and “how are you” at semantically inappropriate places in the conversation.

To truly test the imitative learning and the HMMs, we maintained the latter portion of the time series samples fully hidden from the training algorithms. Figure 8 depicts the resulting synthesized agent just as in the previous format. As the agent initially approaches the human, it remains quiet. Once they are within range and conversing, it interleaves rather mumbled “Hi”, “I see”, “How are you” and “I’m not sure” expressions with the audio of the human in the external world channel. Once again, there is some facial motion which appears to be most active when the agent is producing audio. Although most of the interactions are semantically meaningless, it is interesting that the model recovers some of the timing issues in the interactions and interjects audio that integrates smoothly into the conversation flow with “Umms”, “Yeahs” and so forth. Such prosodic and textural interaction is difficult to design into structured synthetic conversational agents and speech recognition systems due to its behavioral as opposed to syntactical/semantic nature.

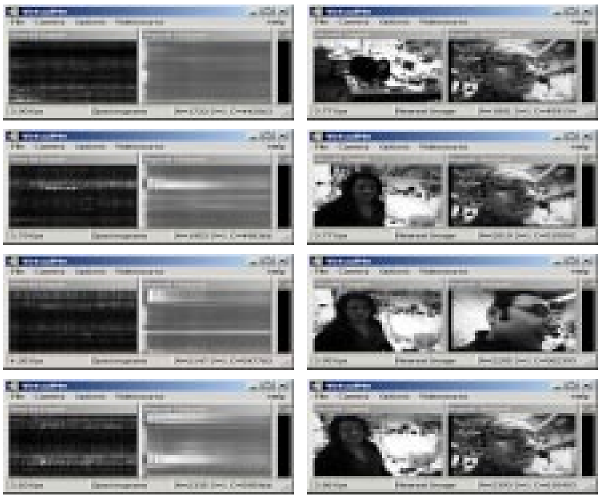

Figure 7: HMM Resynthesis on Training Data. Time is increasing from top to bottom at 3 seconds per frame. Left to right: world spectrograms, synthesized agent spectrograms, world image data and synthesized agent image data.

6 Discussion and Ongoing Work

We have discussed a statistical approach to imitative learning where we predict an agent’s behavior given the external world via the distribution $p(agent|external\, world)$. This distribution is modeled as an input-output hidden Markov model that is discriminatively learned from data using CEM. The system quantitatively performed better than EM and qualitatively generates interesting yet simple audio-visual responses to the external real-world channel triggers. This behavioral model was estimated automatically and without supervision from a large perceptual dataset of human interactions acquired with a wearable computer. The perceptual data was represented accurately and compactly in vector space using a variant of PCA that takes advantage of the inherent structure in acoustic spectrograms and in image pixel data.

Ultimately, however, the problems of imitative learning and development both still loom large and our implementation merely scratches the surface of what is possible with a statistical learning paradigm when it is applied to a large data set of perceptual measurements from human activity. As future work, we recognize the acquisition of “behavior” as described in this paper is quite constrained. With more sophisticated perceptual mechanisms (speech recognition, natural language processing, specialized computer vision) and more structured temporal learning models (i.e.

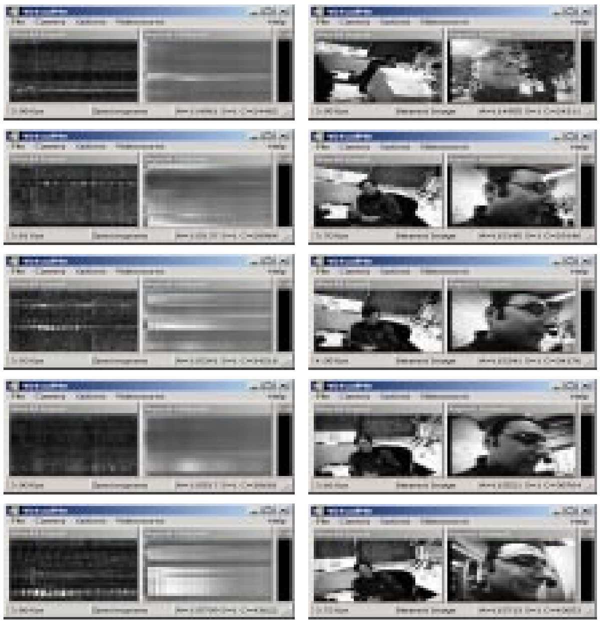

Figure 8: HMM Resynthesis on Test Data.

hierarchical HMMs), it would be possible to naturally extend the imitative learning framework herein to produce increasingly more realistic and more compelling behavioral models.

References

[1] Y. Bengio and P. Frasconi. Input-output HMM's for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249, September 1996.

[2] M. Brand and A. Hertzmann. Style machines. In K. Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[3] R. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. Williamson. *The Cog Project: Building a Humanoid Robot*. Lecture Notes in Computer Science: Springer, (in press).

[4] D. Hogg. Synthetic interaction. http://www.comp.leeds.ac.uk/dch/interaction.htm, 2000. Web Page.

[5] M. Iacobini, R. Woods, M. Brass, H. Bekkering, J. Mazziotta, and G. Rizzolatti. Cortical mechanisms of human imitation. *Science*, 286, 1999.

[6] T. Jebara. *Discriminative, Generative and Imitative learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

[7] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the cem algorithm. In *Advances in Neural Information Processing Systems 11*, 1998.

[8] T. Jebara and A. Pentland. Action reaction learning: Automatic visual analysis and synthesis of interactive behaviour. In *International Conference on Vision Systems*, 1999.

[9] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1998.

[10] L. Kaelbling and M. Littman. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4, 1996.

[11] J. Kosowsky and A. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural Networks*, 7:477–490, 1994.

[12] M. Mataric. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. In C. Nehaniv and K. Dautenhahn, editors, *Imitation in Animals and Artifacts*. MIT Press, 1999.

[13] M. Mataric and M. Pomplun. Fixation behavior in observation and imitation of human movement. *Brain Res. Cogn. Brain Res.*, 7:191–202, 1998.

[14] A. Meltzoff and M. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:74–78, 1977.

[15] J. Piaget. *Play, dreams, and imitation in childhood*. Norton, New York, 1951.

[16] V. Ramachandran. Mirror neurons and imitation learning as the driving force behind the ”great leap forward” in human evolution. Essay, 2000.

[17] G. Rizzolatti, L. Fadiga, B. Galles, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3, 1997.

[18] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3:233–242, 1999.

[19] E. Thorndike. Animal intelligence. an experimental study of the associative process in animals. *Psychological Review, Monograph Supplements*, 2(4):109, 1898.

[20] M. Tomasello, S. Savage-Rumbaugh, and A. Kruger. Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Dev.*, 64:1688–1705, 1993.

[21] J. Weng, W. Hwang, Y. Zhang, and C. Evans. Developmental robots: Theory, method and experimental results. In *Proc. of the International Symposium on Humanoid Robots*, 1999.