Spatio–Temporal Multimodal Developmental Learning

Yilu Zhang, *Senior Member, IEEE*, and Juyang Weng, *Fellow, IEEE*

***Abstract*—It is elusive how the skull-enclosed brain enables spatio–temporal multimodal developmental learning. By multimodal, we mean that the system has at least two sensory modalities, e.g., visual and auditory in our experiments. By spatio–temporal, we mean that the behavior from the system depends not only on the spatial pattern in the current sensory inputs, but also those of the recent past. Traditional machine learning requires humans to train every module using hand-transcribed data, using handcrafted symbols among modules, and hand-link modules internally. Such a system is limited by a static set of symbols and static module performance. A key characteristic of developmental learning is that the "brain" is "skull-closed" after birth—not directly manipulatable by the system designer—so that the system can continue to learn incrementally without the need for reprogramming. In this paper, we propose an architecture for multimodal developmental learning—parallel modality pathways all situate between a sensory end and the motor end. Motor signals are not only used as output behaviors, but also as part of input to all the related pathways. For example, the proposed developmental learning does not use silence as cut points for speech processing or motion static points as key frames for visual processing.**

***Index Terms*—Developmental architecture, multimodal development, speech recognition, visual recognition.**

I. Introduction

MUCH research has been conducted to understand the underlying mechanism that facilitates the superior human performance in generalization, variability toleration, and uncertainty handling. Along this line of research, developmental learning has been proposed as a major mechanism for learning such capabilities, since early learned skills assist the learning of later more sophisticated skills while the system conduct incremental and online learning. We feel that the task-nonspecificity of developmental learning [1] is a major characteristic of developmental learning since the developmental program (DP) must enable the system to learn from simple task contexts to more complex task contexts, but the tasks including their environments cannot be fully anticipated. As discussed in [1], a developmental program is body specific and sensory specific. Therefore, some resource parameters of the developmental program are specific to certain sensory modality, e.g., vision versus audition.

The brain is "skull-closed"—without a need for a human to manually alter internal representation or operation after the birth. In contrast, traditional machine learning uses an "skull-open" approach—manual internal development: Using transcribed video, transcribed audio, separate module-specific training, and manual intermodule linking, as illustrated in Fig. 1(a). After the intermodule linking, the resulted system is only a performer, not able to learn again without further manual internal manipulation. Such a methodology does not allow autonomous development for an open number of skills without opening the "skull" after the "birth" [1].

The motor end of a sensorimotor pathway not only generates actions, but also input actions when the effector of the pathway is supervised externally. In the new multimodal developmental architecture shown in Fig. 1(b), every sensorimotor pathway (visuomotor, auditory motor, and bimodal motor) all have access to the agent motor end, as illustrated in Fig. 1(b), since each pathway requires the information from the motor end to learn.

This architecture is consistent with the idea of behavior-based robots [2], [3], but is different from behavior-based architectures in that our architecture does not explicitly model the decomposition and coordination of different sensorimotor behaviors. For tight integration of sensorimotor behaviors, the coordination of different sensorimotor behaviors within each sensing modality is an emergent property of the pathway.

Consider a setting of visuoauditory learning. During a training session, a teacher presents a new toy to the child using her hand and rotating it continuously so the child can see it from different viewing angles. The teacher asks the question "name?" and then guides the child to produce a correct response, e.g., pick up a label card marked "doggy." The teacher also asks questions about the properties of the toy, e.g., "size?" and then guides the child to answer it correctly, e.g., pick up a label card marked "large." The teacher can test by asking the same question and observe the child's response. Several practices are needed before a child can produce a desired action reliably.

We call this process spatio–temporal multimodal developmental learning. The teacher does not directly manipulate the child's "brain" while new questions and interactions are introduced one after another through sensory and motor interactions. The DP regulates the resource needed, including sensors, effectors, and cortical resources, but it is not a holistically aware central controller (e.g., it did not specify that task-specific concepts "name" and "size" are needed). During this process, most of time except that the effector (hand) is imposed

Manuscript received February 20, 2010; revised April 18, 2010; accepted April 28, 2010. Date of publication May 27, 2010; date of current version September 10, 2010. This is an archival paper of our work on multimodal development that was finished by 2003. This work was supported in part by National Science Foundation under Grant IIS 9815191, DARPA ETO under contract DAAN02-98-C-4025, and DARPA ITO under Grant DABT63-99-1-0014.

Y. Zhang is with General Motors Global R&D, Warren, MI 48090 USA (e-mail: yilu.zhang@gm.com).

J. Weng is with Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: weng@cse.msu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2010.2051437

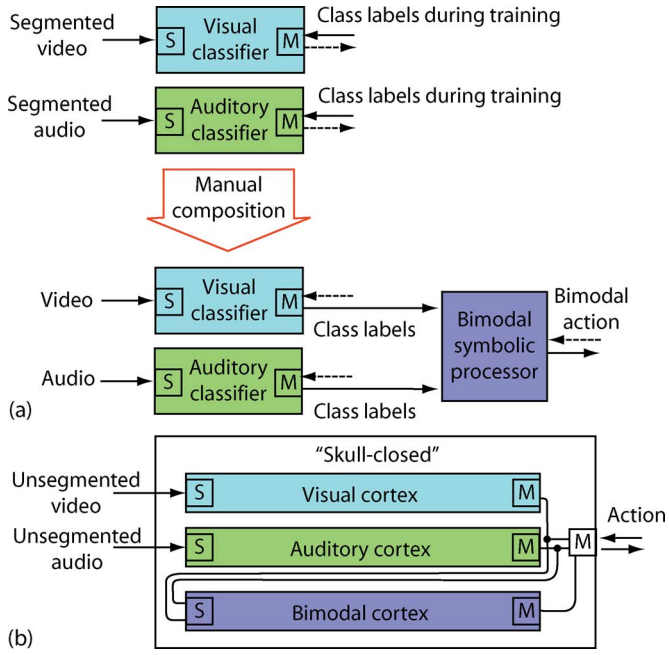

Fig. 1. Traditional multimodal machine learning versus the new multimodal developmental learning. (a) Traditional methodology—"skull-open" approach where the human designer is the external holistically aware central controller, since he understands the task and holistically selects task-specific symbols (e.g., the class labels between modules). (b) The new architecture for autonomous development with multiple sensory modalities: Every sensorimotor pathway (visuomotor, auditory motor, and bimodal motor) needs access to the system motor end. Because of the "skull-closed" nature of autonomous development, teaching for every sensorimotor pathway (e.g., visuomotor and auditory-motor) must have access to the system motor M as part of its input to the pathway. Untranscribed video and audio streams flow into the "brain." Humans can teach the "brain" through sensory and motor interactions. The space and time relationships among multiple sensory inputs and motor signals enable the agent to associate and distinguish. "S" and "M" denote the sensory end and the motor end of a module, respectively. Each connection without an arrow means two one-way connections in opposite directions, bottom–up and top–down, respectively.

(teacher's hand guides the child's hand) with desired actions (picking up the appropriate label card) during some periods. Such imposition of actions through guidance is sparse in time, while all other effectors (e.g., pan-tilt and attention effectors) are autonomous during the guidance. (Other learning modes, such as reinforcement multimodal developmental learning, are future research topics.) This process of developmental learning has the following characteristics.

1) Task-nonspecific learning: The internal self-organization is autonomous as the DP is not imbedded with information about the nature of the interactions other than the general resource needed. Autonomous learning here means autonomous internal self-organization with body-specific information but without task-specific information. However, the agent still needs external teachers. The learner is not constrained to a specific task, such as to distinguish the identity or the size of the object. The internal representation, such as the discriminant feature, is epigenetically generated through the encountered experience.
2) Online learning: The weak performance of the learner is identified right on the spot and the teacher adaptively present more training for the weak areas.
3) Multimodal learning: Vision, audition, and multimodal pathways (modules) are ready to learn simultaneously. There is evidence showing that if visual, auditory, and tactile inputs never have the chance to occur together, there is no opportunity to develop an integrated linkage between what is seen, heard, and felt [4]. While a well-known supporting experiment was done on cats [5], similar results on human babies were also reported [6].
4) Open-ended sensory stream: It has a beginning, but not an end. The stream continuously presents interesting and uninteresting events and is not precut into semantically tailored segments through a manual process called transcription.
5) Sparsely labeled sensory stream: The association of streams with the desired outputs (labels) is provided by occasional online action imposition at certain time instances, which account for only a small percentage of applicable frames, e.g., 2%.

In this paper, we present a general architecture for multimodal developmental learning, and the corresponding algorithms. Interactive verbal questions provide the nature and the time of the visual context to be used by the machine to generate desired responses. Our previous studies have shown the feasibility of real-time visual learning [7] and real-time speech learning [8]. However, facing the combination of vision and audition, new challenges emerge as we will discuss in Section II. For example, the timing of auditory is not precisely synchronized with particular video frames—a great challenge for multimodal developmental learning.

In the current stage of research, we concentrate on the aspects of multimodality and the visual orientation invariance. Although it indicates that a particular type of invariance can be learned from continuous variation of the same object, some other visual issues are not addressed directly. In the experiments, we used a fixed uniform background and each object was rotating at a fixed distance from the camera. The extension of the current system to active visual attention can potentially address other issues such as size and translation invariance and occlusions, but this is beyond the scope of this work.

For multimodal developmental machines that autonomously self-organize internal representations, we must address how their different modules autonomously work together after the "birth." The presented work seems the first to raise and address this multimodal learning mode without requiring each modal to be developed (or programmed) first and without a handcrafted narrowly applicable task-specific representation. The approach and nature of the problem addressed here touch upon the fundamental issues of grounded [9]–[11], acquisition of multimodal (e.g., visual, auditory, and linguistic) capabilities [1], [12], [13], and internal behaviors [14] (e.g., selective attention).

The following section discusses a few challenging problems of multimodal development. Section III analyzes the proposed multimodal architecture. Section IV gives the algorithms. The experimental results are discussed in Section V. Section VI provides some concluding remarks.

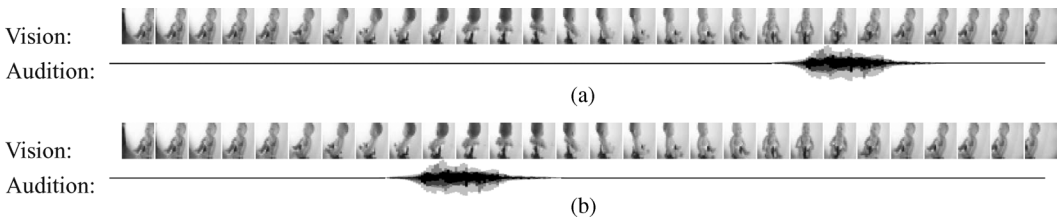

Fig. 2. Direct pattern recognition for multimodal sensory streams will fail—a multimodal developmental architecture is necessary. A typical alignment of an image sequence (the upper sequence) and a spoken utterance (the lower sequence) during (a) learning and (b) performance sessions.

II. Problem Description

Based on the above discussion, suppose that our system's "brain" is "skull-closed." The first implication is that human teachers cannot implant symbolic concepts directly into the "brain." In fact, the human brain never inputs and outputs abstract symbols in the sense of a computer symbol, which assumes the one-to-one correspondence between each symbol and the corresponding meaning—each symbol has only one meaning and each meaning has only one symbol. For example, different instances of an utterance "name?" have different waveforms. Different instances of an action "pick the 'doggy' card" have different trajectories and speeds. In other words, our system only receives and outputs only instances, not abstract symbols. As long as the human (or machine) communicators can correctly interpret each instance with a tolerable variability, the action producer is considered successful.

We discuss a few major problems arising from such a multimodal setting.

A. Time Misalignment

In general, we would like a machine agent to learn to conduct appropriate behaviors based on certain visual-auditory contexts. Particularly, we present a system that learns to answer verbal questions appropriately, given the visual stimuli of dynamically rotating objects. According to the characteristics of the autonomous learning mode discussed above, the learning system should develop visual and auditory perception, and associate audiovisual contexts with behaviors online in real-time. While all the characteristics of the autonomous learning mode are challenging to be realized by a machine agent, we highlight two issues in this section.

In the real world, the visual presence of an object is usually coupled with the related auditory signals, such as a noise made by an object or the verbal name given by a teacher. However, this coupling is not strict because of the following reasons: 1) the visual appearance of an object changes, e.g., the observer may view the object from different angles and the object may rotate as well; 2) for the auditory sensory modality, the signal spreads over many time frames, e.g., the utterance of an object's name covers many auditory frames.

Thus, we have the *double-misalignment* issue (Fig. 2). a) The starting locations of auditory signals in learning and performance sessions do not align with each other with respect to the long visual sequence of an object (e.g., front view in the learning session and back view in the performance session). b) A visual view of an object does not align with a starting location of a long speech sequence (e.g., some views have no auditory signals at all). In other words, if we call a visual-auditory stimulus pair at a particular time instance *an audiovisual context*, it is unlikely that a particular audiovisual context will be exactly repeated in both learning and performance sessions.

Many existing works on multimodal learning rely on the strict coupling between vision and audition information, such as the movement of the lips, the utterance produced, or minimum-mutual-information [15]–[17]. Their success relies on human-designed segmentation scheme of training sequences, a manually assigned association between segments from different modalities, and an atomic symbolic representation. These approaches are not suitable for our autonomous learning problem, since predesigning representation and features are not necessarily applicable to an unknown task.

The misalignment issue is rooted in the fact that an object appears to the learner as a sequence of images captured from different viewpoints. Unless the learner "knows" the sequence of images corresponding to a single object, it will not establish a robust correlation between the visual and auditory stimuli. Fortunately, the physical world has a very important property, i.e., the continuity. For example, the spatio–temporally contiguous views of an object are similar when the capturing speed is high enough. The proposed system forms clusters along the temporal trajectory of the audiovisual context and effectively realizes an "abstraction" procedure to address the misalignment issue. This abstraction is represented by actions, to be explained in our developmental architecture bellow. The underlying mechanism of abstraction is closely related to *object permanence* studied extensively in psychology [18], [19].

B. Sparse Labeling

In order to avoid manual intervene and retain as much sensory information as possible, the proposed system uses the raw input signals as the sensory representation. For the visual modality, the raw-signal representation is essentially the so-called appearance-based representation, which receives support from recent psychophysical and neurophysiological studies [20]–[22] and has a potential to accomplish automation in learning. The problem is that a visual stream is temporally dense, but the labels, with which the teacher tells about the correct answer, are sparse. As shown in Fig. 2, the auditory "name" information only spreads across about 10% of the image frames while the meaning (the real label) is not conveyed until the end of the auditory signals and lasts about 2% of the image frames.

On the other hand, it is incorrect to think that interstimuli interval is unimportant. For example, the interstimuli interval

between a tone and an air-puff needs to fall within the range from 325 to 550 ms for the classical conditioning to be learned effectively by an animal [23], [24].

The sparse labeling issue is also related to the well-known invariance problem. A human learner can generalize the label from a few views of an object to other views when the object is translated, rotated, or scaled. How can a machine learner do this? Similar questions can be asked in speech recognition domain when different speakers or different ways of saying the same phrase are involved.

One solution around this difficulty is to choose an object-based representation, i.e., objects are represented as structural descriptions of their 3-D parts and the relations between those parts in a manner that is independent of the objects' orientation relative to the observer [25]. While this approach solves the invariance problem, it transfers the difficulty to the requirement of establishing internal 3-D models. There is no evidence to support that the brain has an internal monolithic representation (symbolic) about an object in its environments.

Another straightforward solution to this issue is to design a label and manually assign it to the corresponding images one by one. However, this manual transcription is tedious and impractical for an autonomous learning agent.

When a human baby develops, he must take raw sensory streams, not those presegmented and labeled by an human engineer. In fact, the brain of a human baby uses the cooccurance between the sensory frames and his actions (both attention and external actions) [26]–[28]. Our architecture is based on this cooccurrance. If this temporal association is statistically stable, the sensory inputs and motor actions are associated internally on the fly. A similar idea has been used in the association between an reinforcer and a sensorimotor experience [29].

III. Multimodal Architecture

The major multimodal principle we introduce here is that the motor area of a developmental agent is not only for output for generating actions, but also input for internal representations, as shown in Fig. 1(b). Each pathway has direct access to the related sensory end and the motor end. This architecture is supported by the rich top–down connections in the brain from the motor areas back to almost all sensory areas in existing neuroanatomical studies [30]. The motor area of each pathway serves as part of autonomously generated internal state that can be taught externally, supervised directly or shaped through other learning modes.

A. Formulation

Each pathway is formulated as a time-varying (learning) function whose behavior depends on its memory $L(t_n)$. At each time t_n, $n = 0, 1, 2\ldots$ it maps the sensory input $\mathbf{x}(t_n)$ and motor input $\mathbf{a}(t_n)$ to generate its motor output $\mathbf{a}'(t_n)$ and its updated memory $M'(t_n)$

$$(\mathbf{a}'(t_n), M'(t_n)) = g\,(\mathbf{x}(t_n), \mathbf{a}(t_n)|M(t_n))$$

where a vertical bar $|$ indicates the function parameters (i.e., memory). The pathway (module) update for g is performed as $M(t_{n+1}) \leftarrow M'(t_n)$. For the simplicity of time notation, we often simply denote time t as the index n in t_n so that the time takes integer values $t = 0, 1, 2\ldots$.

From Fig. 1(b), we can see that such a system is highly recurrent—motor output is fed back to motor input.

In the asynchronous update mode, the action output $\mathbf{a}'(t_n)$ and the updated memory are not available for other modules in the system within the time window $[t_n, t_{n+1}]$. So every module has their output ready at only the next time instant t_{n+1}.

In the synchronous update mode, the motor output $\mathbf{a}'(t_n)$ from a unimodal module is immediately available for the next bimodal module as part of its sensory input. All the motor outputs from all the modules are available at the motor end of the global system t_{n+1}. We used this synchronous update mode, mainly to reduce the total time required from the sensors to the global system output (i.e., one time step instead of two).

In our experiments, all the actions are meta actions, each action being a sequence of consecutive robot joint positions that is triggered by a single event at the motor. Each meta action is executed continuously without interruption till the action end. During the action execution, all additional actions generated when the motor is busy are flushed (forgotten). Possible interruption of an action by higher priority, later actions is a future research topic.

As the number of actions is not very large, for simplicity the each action uses a canonical representation. Let m be the total number of possible (symbolic) actions. Then the motor vector $\mathbf{m}$ is an m-dimensional vector, where the ith component represents the ith action. Only one action can be executed till the current action ends. Each programed action sequence involves multiple robot joints moving concurrently.

In general, all modules at any time may not necessarily produce actions that are consistent. The motor end for the global system uses a simple mechanism to resolve possible action conflicts. Each component takes the maximum of the "response" from the corresponding components from the three modules, vision, audition. and bimodal. The largest value in the entire resulting motor vector $\mathbf{a}$ represents the output action at this time. But only if it is higher enough (e.g., higher than 0.5), can this action be considered generated by the system. This simulates a winner-take-all mechanism.

B. Active Actions as Part of Internal States

The multimodal architecture provides mechanisms that are not apparent in the illustration in Fig. 1(b).

First, the teacher can teach actions to represent virtually any property. For example, in the brain, the ventral stream [31] may provide the "what" information that drives the verbal pronunciation for the "what." This is because any human communicable concept can be said. On the other hand, the dorsal stream [31] may provide "where or how" properties of an object (e.g., size and location) that arm to reach the object.

Second, since an action at any time is a part of (top–down) input, the action can be taught to represent the equivalent temporal context that the system can use to deal with the above misalignment problem. This is effective as the teacher can interactively teach appropriate actions at different contexts. The following is an example. When an object is placed into the field of

view, the teacher can teach the system to produce an action corresponding to the required visual label (e.g., name-size). When a verbal question is heard, the teacher can then teach the system to produce the corresponding question label as its action. The input to the bimodal module is the primed contexts (discussed later) from both single modality modules. The teacher teaches the bimodal module to produce the correct bimodal answer right after the question is heard. It is important to note that each "label" as action is not symbolic, but numeric with variability in output.

It is also important to note that with "action" as part of information for internal states, we still want the system to be time sensitive. Instead of being sensitive to brute-force multimodal sensory inputs sequences, the system is sensitive to the time where action is produced as "softly abstracted state."

Third, the action value can represent the certainty of classification, so that the certainty increases as more views have been observed from the sensory stream.

C. Functions of a Sensorimotor Pathway

Existing neuroanatomical studies reviewed by Felleman & Van Essen [30] indicate that each cortical area has bottom–up sensory input $\mathbf{x}$ and top–down action input $\mathbf{a}$. The input space M of each sensorimotor pathway is the last context $\mathbf{l} \in L$. Each $\mathbf{l}$ includes both the sensory part $\mathbf{x}$ and the action part $\mathbf{a}$, so that $\mathbf{l} = (\mathbf{x}, \mathbf{a})$. For a general-purpose cortex, sensation $\mathbf{x}$ or action $\mathbf{a}$ alone is not sufficient. Without sensation $\mathbf{x}$, there is no basis to generate action $\mathbf{a}$ as the class of $\mathbf{x}$. Without previous action $\mathbf{a}$, there is no basis to generate the next action $\mathbf{a}'$ depends on the nature of previous $\mathbf{a}$.

The internal representation $M(t_n)$ of each sensorimotor pathway includes self-organized clusters of the input space L with $\mathbf{l} = (\mathbf{x}, \mathbf{a}) \in L$, represented as $F = \{\mathbf{v}_i | \mathbf{v}_i = (\mathbf{x}_i, \mathbf{a}_i), i = 1, 2, \ldots, c\}$, as a set of feature clusters. Given any input $\mathbf{l} = (\mathbf{x}, \mathbf{a})$, conceptually each pathway finds the best matched cluster

$$\mathbf{v}_j = \min_{1 \le i \le c} \|\mathbf{l} - \mathbf{v}_i\|_{\text{tree}}$$

where $\|\cdot\|_{\text{tree}}$ denotes the distance measure of the incremental hierarchical discriminant regression (IHDR) tree appeared in [32] and will be outlined later.

We can consider IHDR as a fast neural network whose long term memory has a dynamic number of parameters and thus, never has a problem of over-fitting. For each $\mathbf{v}_j$, IHDR has a link to the next primed (predicted) context $\mathbf{l}' = (\mathbf{x}', \mathbf{a}')$, which is produced as recalled next sensation $\mathbf{x}'$ and next action $\mathbf{a}'$ as soon as $\mathbf{v}_j$ is found as the best match to $\mathbf{l}$. IHDR also updates its features $F(t_n) \subset M(t_n)$ after each input $\mathbf{l}$. In order to be the best match, both parts $\mathbf{x}$ and $\mathbf{a}$ need to match well, to lead to the corresponding joint state $\mathbf{v}_j$. During intermittent training, the action input $\mathbf{a}$ serves as supervised action but while the supervised action is absent, the action input $\mathbf{a}$ serves as action context during autonomous performance (e.g., I am replying, I have replied, etc.).

In summary, in this connectionist network, the numeric action vector serves as a part of the internal state. This part of state is dynamically learned during interactive development, different from the statically designed symbolic state in the finite state machines (FSMs) and its probabilistic versions (hidden Markov models or the partially observable Markov decision process). This way, new concepts can be learned incrementally. The more recent model lobe component analysis (LCA) [33] is dually optimal. IHDR can be replaced by LCA and the effects of such a replacement are subject of future studies.

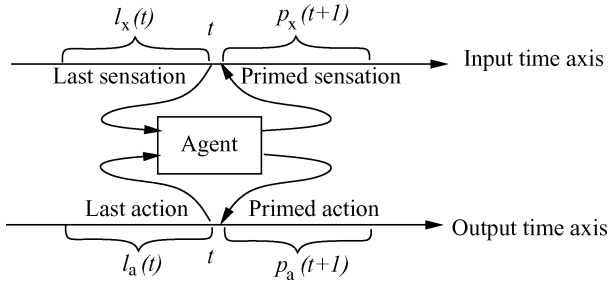

Fig. 3. Contextual view of an agent.

IV. Algorithm

Due to the importance of actions, we define *a context* as the sensation and the action of an agent within a volume of time. With respect to a certain time instance t, the context over a time period up to t is called *the last context*, while the context after t is called *the primed context* (as predicted sensations and actions). The job of an autonomous multimodal learning agent is to internalize and generalize the causal relationship between the last context and the primed context as shown in Fig. 3 so that the agent can predict (prime) what to do when similar context is encountered the next time. We formulate this causal relationship as a mapping

$$p^{(m)}(t+1) = g^{(m)}\left(l^{(m)}(t)\right) \tag{1}$$

where m stands for multimodal learning, $l^{(m)}$ is the last multimodal context, and $p^{(m)}$ is a primed multimodal context.

Because of the uncertainty of the sensation and the action, $p^{(m)}(t+1)$ and $l^{(m)}(t)$ are random variables and (1) can be modeled as a special Markov model called, *observation-driven Markov model (ODMM)* [34], [35], where the transition probability $P(p^{(m)}(t+1)|l^{(m)}(t))$ can be estimated incrementally. In practice, the space of $l^{(m)}(t)$ becomes unbounded as t increases. To keep the problem traceable, we define

$$l^{(m)}(t) = \Big\{l_x^{(m)}(t), l_a(t), l_x^{(m)}(t-1), l_a(t-1) \ldots l_x^{(m)}(t-k), l_a(t-k)\Big\}$$

where $l_x^{(m)}(t)$ and $l_a(t)$ are the last sensation and the last action, respectively, at time instance t.[1] The problem of interest is reduced a kth order ODMM.

This association is not as straightforward to learn as it looks. First, because not all of the information in the last and primed contexts is causally related, the agent has to have certain mechanism to select information between and/or within different sources (vision, audition, touch, or action). Second, since we used the raw and numerical (instead of symbolic) representation

[1]Note while the sensation is sensor specific, the system action is not.

of the sensory inputs, the dimensionality of the context can be as high as hundreds for audition and thousands for vision. It is extremely difficult to extract the high-dimensional mapping in real-time.

We designed a hierarchical architecture to realize the autonomous multimodal learning. At different levels of the hierarchy, the last-prime context mappings were implemented with different details. At the lowest level, we used IHDR tree published in [32] as the association mapping engine. It is a tree version of a cortex-like learner. Composed of two IHDR trees, a level building element (LBE) specialized in one of the three domains, auditory modality, visual modality, and the fusion of the two modalities. The three LBEs together realized the mapping described in (1).

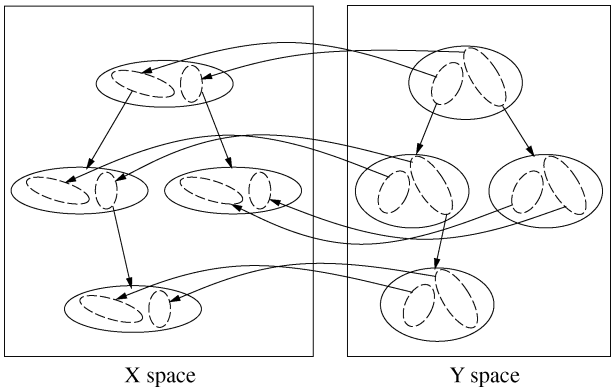

Fig. 4. Hierarchical discriminant analysis.

A. IHDR

Hierarchical discriminant regression (HDR) is a new hierarchical statistical modeling method introduced by Hwang and Weng [36]. IHDR is an algorithm that constructs an HDR tree incrementally [32], which fits our needs for a continuously learning agent. To keep this paper self-contained, we discuss the basic idea of HDR here. The interested reader is referred to the original papers for details.

Consider a general regression problem: approximating a mapping $h : X \rightarrow Y$ from a set of training samples $\{(x_i, y_i) | x_i \in X, y_i \in Y, i = 1, 2 \ldots n\}$. HDR employs the idea of linear discriminant analysis (LDA) [37] to find such a mapping. In a typical usage of LDA, one needs to estimate the between-class and within-class scatter matrices, which require class information. However, for a regression problem, y_i is a numerical vector and there are very few samples sharing a single y_i. In other words, there are very few samples in each class. Therefore, the within-class scatter matrix will be poorly estimated, especially for high-dimensional input data, e.g., a few thousand dimensions in vision problems or a few hundred dimensions in audition problems. HDR handles this issue in a coarse-to-fine way.

At the root of an HDR tree, a few (q) *y-clusters* in the output space are generated using a k-means-like algorithm. The x-part of the samples is accordingly clustered and we call these clusters *x-clusters*. Since q is typically a small number, e.g., 5, it is effective to estimate the between-class and within-class scatter matrices for these q x-clusters. The q x-cluster centers span a $(q-1)$-dimensional space, which is called a *discriminant space* because it characterizes the discriminant information among x-clusters. A probability-based metric called the *size-dependent negative log likelihood* (SDNLL)

$$L(x, c_i) = \frac{1}{2}(x - c_i)^T W_i^{-1}(x - c_i) + \frac{q-1}{2}\ln(2\pi) + \frac{1}{2}\ln(|W_i|) \quad (2)$$

in the discriminant space is used to determine which x-cluster a test sample belongs to with c_i represents the center of the ith x-cluster. The *size-dependent scatter matrix* (SDSM) W_i in (2) is defined as the weighted sum of three matrices

$$W_i = w_e \rho^2 I + w_m \Gamma + w_g \Gamma_i \quad (3)$$

where, ρ is the standard deviation of all samples in the root node, Γ_i is the sample covariance of the ith cluster, Γ is the average of the covariance matrices of q clusters. The weights of these three terms start with large w_e when there are a few samples in the node, and change to large w_m and eventually large w_g as the number of the samples increases. The weights always follow the normalization constraint, i.e., $w_e + w_m + w_g = 1$. Each corresponding pair of x- and y-clusters form a child node of the root and the above process is continued in the child node. If a node receives only a few samples, called primitive prototypes, this node is then a leaf node without children. After all, the tree structure recursively excludes many far-away cases from consideration (e.g., an input face need not be searched among non-faces) (Fig. 4), to reach a logarithmic time complexity.

In summary, the HDR technique realizes a regression in a high-dimensional space. It automatically derives discriminating feature subspaces in a coarse-to-fine manner from the input space to generate a tree architecture of self-organization memory. HDR has little limitation in terms of representation power and potentially can fit any data. HDR can learn without any iterations. As a result, HDR realizes one-instance learning without local minima, i.e., zero error on training data without iterations. Most importantly, HDR can handle high-dimensional data in real-time, which is crucial to an autonomous learner.

B. Level-Building Element

An IHDR tree realizes the lowest level last-prime context mapping

$$p(t+1) = g(l(t))$$

where $l(t)$ is the last context, $p(t)$ is the primed context, and

$$p(t) = [p_x(t) \quad p_a(t) \quad p_Q(t)]$$

where p_x is a primed sensation vector, p_a is a primed action vector, and p_Q is a primed value associated with the primed action vector. Depending on the type of LBE that an IHDR tree belongs to, the content of the (last or primed) context is different. For example, if the IHDR tree is part of an audition LBE, the last context contains the auditory sensation and the action sensation.

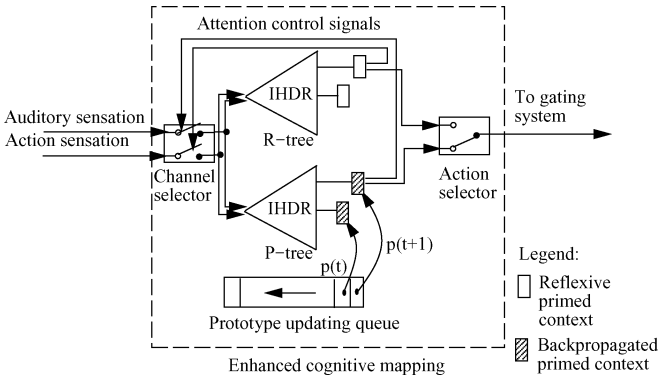

Fig. 5. Level-building element.

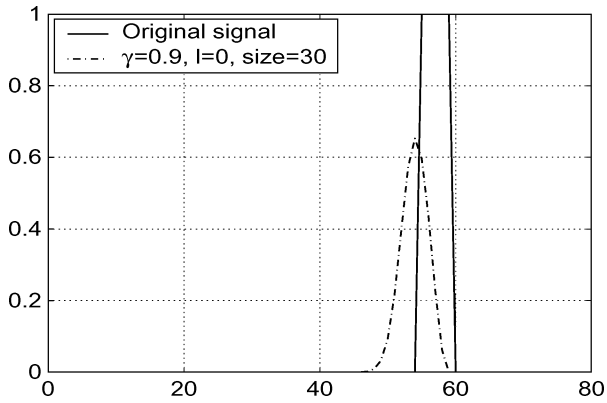

Fig. 6. Behavior of model (5).

For each query, the IHDR tree returns a list of primed contexts, i.e.

$$\{p_1(t+1),\ldots,p_k(t+1)\} = R(l(t)) \tag{4}$$

where R represents the regression realized by the IHDR tree. The primed context with the highest primed value is selected as the ultimate output, i.e.

$$p(t+1) = \arg\max_{\{p_1(t+1),\ldots,p_k(t+1)\}} p_{Q,p_i(t+1)}.$$

An LBE is composed of two IHDR trees. Shown in Fig. 5 is an audition LBE, taking two channels of sensory inputs, the auditory sensation, and the action sensation. The two IHDR trees in an LBE are identical except that the bottom one is associated with a *prototype updating queue* (PUQ). We call the upper one the *reality tree* or the *R-tree*, and the bottom one the *priming tree* or the *P-tree*. The R-tree is necessary here for required fine temporal resolution of sensory processing. This characteristic is similar to that of the dorsal pathway [31] that requires higher temporal resolution during, e.g., hand manipulation. The P-tree does not have a high temporal resolution, but can predict further into immediate future. This characteristic is similar to that of the temporal lobe along the ventral pathway [31] which identifies the type of an attended object. Nether tree can take over the role of the other because of the need for both characteristics in auditory and visual perception. Similar to the dorsal and ventral pathways that lead to the motor area in the frontal cortex, the R-tree and P-tree both lead to actions.

C. P-Tree

The goal of PUQ for the P-tree is to enable a looking-ahead (farther priming) mechanism. The PUQ maintains a list of pointers to the primed contexts retrieved by the P-tree. At every time instance, a pointer to a newly retrieved primed context enters the PUQ while the oldest one moves out. When the pointers are kept in PUQ, the primed contexts they point to are updated with a recursive model adapted from Q-learning [38]

$$p^{(n)}(t) = p^{(n-1)}(t) + \frac{1+w}{n}\left(\gamma p^{(n-1)}(t+1) - p^{(n-1)}(t)\right) \tag{5}$$

where, $p^{(n)}(t)$ is the primed context at time instance t, n represents the number of times $p^{(n)}(t)$ has been updated, and γ is a time-discount rate. w is an amnesic parameter used to give more weight on the newer data points, which is typically positive, e.g., $w = 2$.

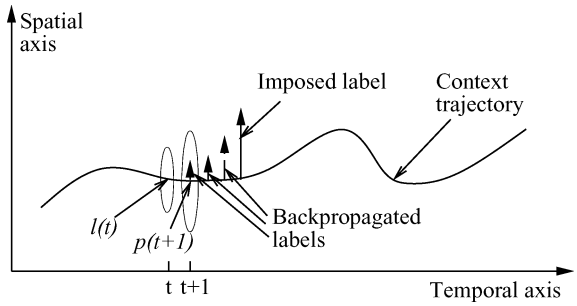

Fig. 7. To an agent, the evolving environment appears to be an ever-extending context trajectory in a spatio–temporal space.

Reorganizing (5), we have

$$p^{(n)}(t) = \frac{n-1-w}{n}p^{(n-1)}(t) + \frac{1+w}{n}\gamma p^{(n-1)}(t+1) \tag{6}$$

which shows that a primed context $p^{(n)}(t)$ is updated by averaging its last version $p^{(n-1)}(t)$ and the time-discounted version of the current primed context $p^{(n-1)}(t+1)$. In this way, the information embedded in the future context, $p^{(n-1)}(t+1)$ in model (5), is recursively backpropagated into earlier primed contexts.

To view this effect more intuitively, we show the behavior of the prediction model in a simple example. Suppose the primed contexts appearing over time are represented by a series of scalers. A scaler with a value of 1 means the primed context contains certain information while 0 means no information is embedded. An example of a series of the primed contexts is shown with a solid line in Fig. 6, where there is certain information over the five consecutive time instances $(t = 55, 56 \ldots 59)$ and nothing elsewhere. Applying the model (5) with $\gamma = 0.9$, $w = 0$, and a PUQ of size 30, we get the dotted line in Fig. 6, which shows that the information is spread over the time interval [46, 59] with the peak appearing at $t = 54$.

The P-tree and the PUQ with model (5) together address the sparse label issue. As illustrated in Fig. 7, to an agent, the evolving environment appears to be an ever-extending context trajectory in a spatio–temporal space where the spatial axis represents the image vector space for the visual modality. The sparse label problem states that it is impractical to assign

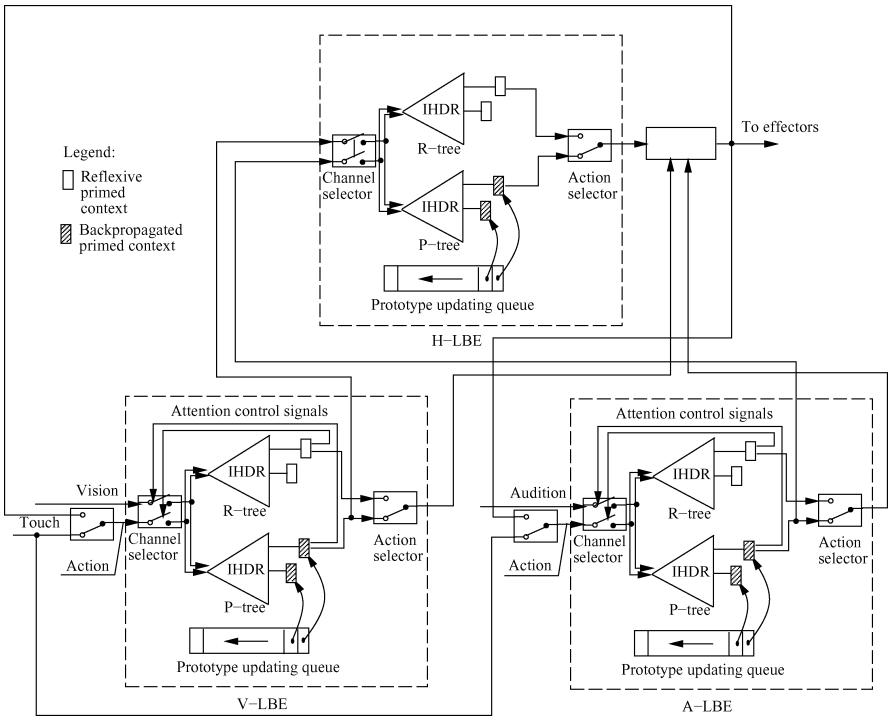

Fig. 8. Multimodal learning system architecture.

labels, the expected actions, to the agent by the teacher at every point on the trajectory. However, when a label *is* given, the teacher usually refers to a vicinity of the contexts around the labeled point. In the P-tree, an x-cluster center (clustering along the spatial axis) approximates the context at time instance t (the last context $l(t)$) and a y-cluster center approximates the context at time instance $t+1$ (the primed context $p(t+1)$). The association between the x-cluster center and the y-cluster center is captured by the IHDR mapping. With model (5) running in the PUQ, the sparse label (action) information is propagated backwards along the context trajectory and embedded into $p(t+1)$. As a result, when a similar context $l'(t) \simeq l(t)$ occurs, the IHDR tree will be able to retrieve the primed context,

$$p'(t+1) = g\left(l'(t)\right) = g\left(l(t)\right) = p(t+1)$$

which contains the appropriate expected action.

The LBE module was designed in order to fulfill a general learning purpose. In multimodal learning, all components in the module were not used as seen in the algorithm below. For simplicity, we do not discuss the LBE components that were not used in the multimodal learning architecture, such as attention control signals, channel selector, and action selector. The interested reader is referred to [39] for detailed discussions.

D. Multimodal Learning

Fig. 8 shows the architecture we used to do multimodal learning. It has three LBE modules, a vision LBE (V-LBE), an audition LBE (A-LBE), and a high-level LBE (H-LBE). Their roles are indicated by the multimodal architecture in Fig. 1(b). Functionally, V-LBE corresponds to the visual pathways in the brain. Their forebrain part is the visual cortex. A-LBE corresponds to the auditory pathways in the brain. Their forebrain part is the auditory cortex. H-LBE corresponds to the frontal cortex which bidirectionally communicates with the vision pathways and the auditory pathways at its sensory end. It also bidirectionally communicates with the premotor and motor areas in itself to generate actions and to learn from actions. The functions of these level-building elements appear necessary for infant-like developmental learning.

Specializing in different domains, the three LBEs realize the last-prime context mappings

$$p^{(v)}(t+1) = g^{(v)}\left(l^{(v)}(t)\right)$$
$$p^{(s)}(t+1) = g^{(s)}\left(l^{(s)}(t)\right)$$

and

$$p^{(h)}(t+1) = g^{(h)}\left(l^{(h)}(t)\right)$$

respectively, where v stands for vision, s stands for sound, and h stands for high-level. The last context of V-LBE $l^{(v)}$ is composed of the visual sensation and the action sensation

$$l^{(v)} = \begin{bmatrix} l_x^{(v)} & l_a \end{bmatrix}.$$

where the visual sensation is the original image captured by a CCD camera. We do not manually derive low-level features such as edge histogram. The last context of A-LBE $l^{(s)}$ is composed of the auditory sensation and the action sensation

$$l^{(s)} = \begin{bmatrix} l_x^{(s)} & l_a \end{bmatrix}$$

where the auditory sensation is captured by a sound blaster card through a microphone. We perform Cepstrual analysis on the original sound signals before entering data into the A-LBE.

Fig. 9. Sample sequence of the visual sensation.

Fig. 10. Sample sequence of the primed visual sensation.

Since sound is a linear signal in the sense that information is distributed over a period of time, each auditory sensation vector actually covers 20 speech frames as will be discussed in the experimental results.

In the first multimodal learning algorithm we propose, the last context of H-LBE $l^{(h)}$ is composed of the primed sensations from the P-trees of both V-LBE and A-LBE

$$l^{(h)} = \begin{bmatrix} p_x^{(s)} & p_x^{(v)} \end{bmatrix}. \tag{7}$$

Because of the recursive averaging in model (5), the primed context of V-LBE and A-LBE changes slowly along the context trajectory. Particularly, the primed sensation part of the primed contexts changes slower compared to that of the corresponding last sensations. For example, instead of receiving the distinguishing images of different views of an object as shown in Fig. 9, H-LBE deals with a sequence of images with reduced variance (Fig. 10). If we imagine the auditory context (the last context of A-LBE) evolves along with the visual context (the last context of V-LBE) in a spatio–temporal space (Fig. 11), the inputs $(l^{(h)})$ to H-LBE are effectively the cluster centers of the last contexts along the context trajectories. Thus, without worrying about the details of the samples within each of the clusters, H-LBE treats the temporally neighboring contexts as one item, which is an abstraction process. As long as the auditory context (the verbal question) overlaps with the visual context (the image sequence of an object), they form *one* audiovisual context from the perspective of H-LBE, which resolves the misalignment issue illustrated in Fig. 2. Note if not learning in real-time,

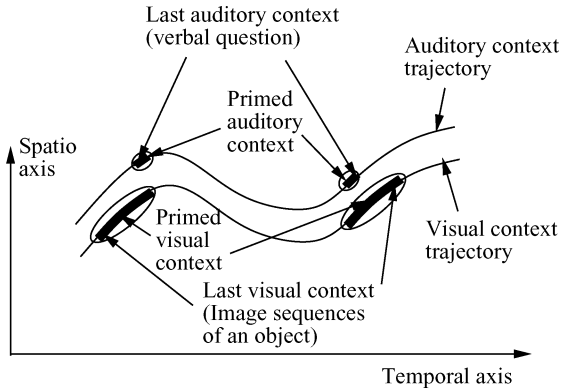

Fig. 11. Primed contexts are cluster centers of the last contexts along the context trajectories. The alignment of the verbal questions and the image sequences of the object are not the same between the left and the right examples. By representing last auditory and last visual contexts with the cluster centers, this misalignment is resolved as long as the verbal question overlaps with the image sequence of an object.

the learner can not enjoy the continuity of the consecutive contexts and the abstraction process will not be effective because clustering does not make any sense.

From Fig. 8, we can see that the visual module V-LBE, the auditory module A-LBEL, and the bimodal module H-LBE all send actions to the motor area. In our agent architecture, the motor area has an internal action called action release [14] so that the agent can learn contexts where a particular action should be immediately released or not. In this work, we adopt a simpler teaching schedule: V-LBE and A-LBE only learn covert

1) Collect the sensation from the auditory sensor, $l_x^{(s)}(t)$, the visual sensor, $l_x^{(v)}(t)$, and the action sensor, $l_a(t)$ (If an action is imposed through the touch sensors, $l_a(t)$ is the imposed action. Otherwise, it is the action produced by the system itself in the last computation loop.). The sensations from different sensors form a last context $l(t)$.
2) Update the P-trees of both V-LBE and A-LBE using the IHDR learning algorithm.
3) Retrieve the P-trees of both V-LBE and A-LBE to get a list of primed contexts, from which the ones with the highest primed value are selected and denoted as $p^{(v)}(t)$ and $p^{(s)}(t)$, respectively.
4) Update the PUQs of both V-LBE and A-LBE using model (5).
5) Do either of the following for algorithm 1 and 2, respectively:
   Algorithm 1: Form $l^{(h)}$ with Eq. (7) as the input to H-LBE.
   Algorithm 2: Form $l^{(h)}$ with Eq. (8) as the input to H-LBE.
6) Update the R-tree of H-LBE the IHDR learning algorithm.
7) Retrieve the R-tree of H-LBE to get a list of primed contexts, from which the one with the highest primed value is selected and denoted as $p^{(h)}(t)$.
8) The motor executes all the action part of every element in $\{p^{(v)}(t), p^{(s)}(t), p^{(h)}(t)\}$, if any of them is overt.

Fig. 12. Multimodal learning algorithm 1 and 2, with the difference in step 5.

actions that are not released to the external motor and the bimodal H-LBE learns overt actions that are released to external motor.

E. Representation of Actions

In general, an agent with n motors have an action vector as an n-dimensional vector, where each component represents the speed or power output of the corresponding motor. In this way, different motors can work together, producing complex sequences of coarticulated actions which can be aborted or altered at any time.

In the experiments reported here, we do not ask the learner to deal with the temporal aspects of the actions. Each action sequence is represented as a programmed symbolic procedure. The motor vector is a vector of n-dimensional vector for n symbolic actions. To active an action, the corresponding component in the motor output vector is 1 and other components are 0. Repeated activation of the same component has no effect until the same action has finished. With this simplified motor representation, the agent cannot abort or modify an ongoing action.

For our experiments, this type of action representation is sufficient. The development of actions that involve the coordination of multiple joints (multiple motor neurons) is a subject for future research. Furthermore, how could a developmental system autonomously develop skills to terminate or alter an ongoing action at any time depending on the environmental context changes? This is also an important subject for future research.

F. Algorithms of the Developmental Agent

A high-level outline of the algorithm (multimodal learning algorithm 1) in each perception–action step is shown in Fig. 12. As one may notice, in this algorithm, the training (featured by words such as "update") and testing (featured by words such as "retrieval") processes are embedded to each other in order to make online learning possible.

Multimodal learning algorithm 1 can be further improved with another abstraction process. The utterances of the same word vary from people to people and, therefore, each word is typically composed of several modes in its auditory representation. So is the primed sensation for the same word. As a result, the decision boundary is complicated in the space spanned by the primed sensations from both V-LBE and A-LBE, as illustrated in the bottom of Fig. 13, which gives H-LBE a hard time. Actually, for the sake of answering questions, H-LBE only needs information to discriminate different verbal questions instead of different modes of uttering the same question. With this in mind, let us take a close look at the behavior of A-LBE. Given a particular question, the primed action of A-LBE is the same as long as the same object is presented. In this sense, as a representation of the question information, the primed action has lower variance than the primed sensation does. Therefore, using the primed action information, A-LBE offers an abstraction process for H-LBE.

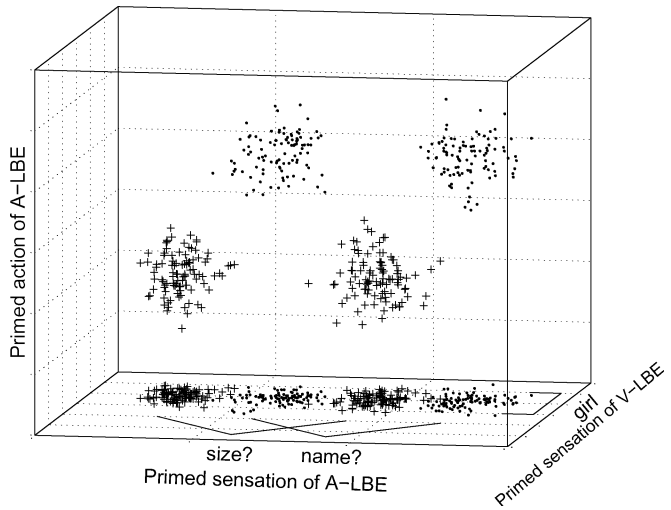

Fig. 13. Illustrative comparison of using and not-using the primed action in decision making.

Following this thinking, we propose the multimodal learning algorithm 2 (shown in Fig. 12), in which the last context of H-LBE $l^{(h)}$ is defined as

$$l^{(h)} = \begin{bmatrix} p_x^{(s)} & p_a^{(s)}(t) & p_x^{(v)} & P_a^{(v)}(t) \end{bmatrix} \tag{8}$$

where $P_a^{(s)}(t)$ and $P_a^{(s)}(t)$ are called *primed action patterns*. We define the primed action pattern of A-LBE as

$$P_a^{(s)}(t) = \sum_{i=1}^{k} p_{Qi}^{(s)}(t) p_{ai}^{(s)}(t) \tag{9}$$

where k is the total number of primed contexts retrieved from the P-trees of A-LBE [see (4)], $p_{Qi}^{(s)}(t)$ is the primed value associated with the ith primed context, and $p_{ai}^{(s)}(t)$ is the primed action of the ith primed context. We use the primed action pattern instead of the primed action in order to cover the situations when the objects are different. Similarly, for V-LBE, we have

$$P_a^{(v)}(t) = \sum_{i=1}^{k} p_{Qi}^{(v)}(t) p_{ai}^{(v)}(t). \tag{10}$$

The inclusion of the primed action pattern in the input to H-LBE well separates the clusters as illustrated in Fig. 13.

The reason that the above additional information helps to improve performance can be explained in terms of information theory. Let the random variables p_x and p_a represent the primed sensation and the primed action, respectively; $f(p_x)$ and $g(p_x)$ are the probability density functions (p.d.f.s.) for "name" and "size," respectively; and $f(p_x, p_a)$ and $g(p_x, p_a)$ are the joint p.d.f.s. for "name" and "size," respectively. We prove in the Appendix that

$$D\left(f(p_x, p_a) \| g(p_x, p_a)\right) \geq D\left(f(p_x) \| g(p_x)\right)$$

where $D(\cdot)$ is the Kullback-Leibler distance (relative entropy) between the two p.d.f.s. In other words, by including primed action in the input to H-LBE, we increase the discriminant power of the representation and thus, expect better performance. The experimental results below show the effectiveness.

V. Experimental Results

We implemented the multimodal learning architecture on the self-organizing, autonomous, incremental learner (SAIL) robot, a human-size mobile robot built in-house at Michigan State University (Fig. 14). SAIL has a drive-base, a six-joint robot arm, a neck, and two pan-tilt units on which two CCD cameras (eyes) are mounted. A wireless microphone functions as an ear. SAIL has four pressure sensors on its torso and a total of 28 touch sensors on its eyes, arm, neck, and bumper. Its main computer is an Xeon 2.2 GHz dual-processor workstation with 1 GB RAM. All of the sensory information processing, memory recall, and updating, as well as effector controls are done in real-time.

We assume that there is only one action from the agent at any time. Therefore, the highest output from the agent's motor vector is considered the current action. When the agent integrates three action outputs from the three modules, vision, audition, and multimodal, the same principle is used: a maximization operation is applied to each component of the three action vectors. This canonical action representation is for simplicity but is static and wasteful. In general, each pattern of the action vector can represent a different action, which is the subject of future extensions.

(a) (b)

Fig. 14. SAIL robot at Michigan State University at (a) a training session and (b) a testing session.

The experiment was done in the following way. After SAIL started running, the trainer mounted objects one after another on the gripper of SAIL and let SAIL rotate the gripper in front of its eyes at the speed of about six second per round for about one round. During training, the properties of each object was taught to SAIL through question-and-answer. First, the trainer verbally asked questions, namely "name?" and "size?" when an object was presented. Then the trainer gave the appropriate answers by pushing the switch sensors of SAIL, where different switch sensor status represented different answers. All the switches here are simply touch sensors, each is linked with an innate (programmed-in) behavior. Developmentally learning behavior decomposition is an important subject, but is beyond the scope of this work. During testing, the objects were presented and the questions were asked again, but no answers were given. SAIL's responses were recorded for performance evaluation. Note SAIL did not stop running until all the objects and questions were taught and tested.

Since the objects were rotated and moved in and out of SAIL's field of view continuously, the orientation and the positions of the objects kept changing. There were hardly any chances that SAIL could see the same images of the objects when the same question was repeated later. In total, 12 objects were presented (Fig. 15) to SAIL. All of these real-world objects consisted of very complex shapes and nonrigid forms, for example, Harry Potter's hair. It was extremely difficult, if not impossible, to model them using 3-D representations.

The video data was captured by a CCD camera and a Matrox Meteor II board as 256-grayscale images at 30 frames/s. The dimension of the images was 25-by-20 pixels. The theoretical number of possible images in this image vector space is $256^{25\times 20} \simeq 1.32 \times 10^{1204}$, an astronomical searching space, which in turn offers a representation that can accommodate rich discriminant information. The auditory data were digitized at 11.025 kHz by a normal sound blaster card. We did Cepstral analysis on the speech data and 13-order mel-frequency Cepstral coefficients (MFCCs) were computed over 256-point wide frame windows. There was an overlap of 56 points between two consecutive frames. Therefore, the MFCCs entered the auditory channel of SAIL at the rate of about 50 Hz. We concatenated 20 consecutive MFCC vectors together as a single auditory sensation vector because a 18.1 ms (200/11.025) speech frame is too short to convey any meaningful information. To compensate the slower capture rate of image data, SAIL used the last captured image to accompany the new vector of MFCC when a new image was not available.

Fig. 15. Objects used in the experiment.

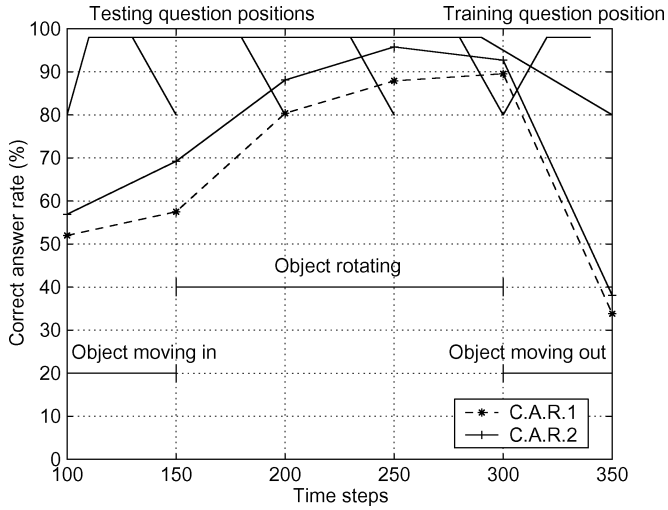

Fig. 16. Results of multimodal learning algorithm 1: the two correct answer rates of SAIL versus the question positions in each image sequence.

Note although we did not manually design discriminant features for either video or sound signals, the signals were first collected by sensors, which are basically band-pass filters. The visual sensors, the photoreceptors on CCD cameras, are visible-light filters that work similarly (but not equivalently) to the cones and the rods in a human's retina. The cones and rods connect to the optical nerve [40]. The Mel-frequency Cepstral analysis does band-pass filtering on sound signals [41], which works similarly (but not equivalently) to the hair cells on the Basilar membrane in a human's cochlea [40]. The hair cells connect to the auditory nerve. The center of human brain does not receive time-domain signals (wave signals), but frequency-domain signals. In this sense, the Mel-frequency Cepstral analysis functions equivalently to the CCD cameras in collecting signals.

A. *Simulation Experiments*

To examine the behavior of SAIL in detail and evaluate the performance, we pursued an experiment on prerecorded data first.

The image data of each object was five image sequences, each image sequence contained 350 frames as follows:

- frame 1–100: background images;
- frame 101–150: an object moving to the center of SAIL's field of view;
- frame 151–300: the object rotating along its center axis;
- frame 301–350: the object moving out of the SAIL's field of view.

The auditory data was a subset of the number data set contributed by 63 people with a variety of nationalities (American, Chinese, French, Indian, Malaysian, and Spanish) and ages (from 18 to 50). Each person made five utterances for each number from one to ten. In the simulation experiment, ten subjects were randomly selected from the total of 63. We used the utterances of "one" to represent the "name" question and "two" to represent the "size" question.

For each subject, each question, and each object, the five utterances were paired with the five image sequences accordingly. Four of the five (image sequence, utterance) pairs were used for training and the remaining one was for testing. So, with ten people, two questions, and twelve objects, SAIL had 960 training (image sequence, utterance) pairs and 240 testing (image sequence, utterance) pairs. The training data were linked one after the other, followed by the testing data, which were also linked one after the other, to form a long and continuous "super-sequence." This is to emulate the scenario that the autonomous learner runs continuously once started. In the training session, the switch sensor inputs (a numerical vector) were given after the utterances were finished and lasted for seven consecutive image frames, which was the time when SAIL was taught the answers. In the test session, we recorded SAIL's responses for performance evaluation.

To emulate the situation that the trainer would not be able to ask questions when the objects were presented with exactly the same orientations and positions in the training and the testing sessions, we randomly chose the points to align the image sequence and the utterances. Fig. 2 shows the typical alignment between an image sequence and an utterance, which is different during training and testing. Specifically, the end points of the utterances was aligned with image number 300 during training. When testing, we aligned the end points of utterances with image numbers 100, 150, 200, 250, and 300. As a result, each testing (image sequence, utterance) pair was used five times. Therefore, in total, SAIL was trained for 960 times and tested $240 \times 5 = 1200$ times.

The behavior of SAIL was evaluated in two different ways. First, we counted the total number of times SAIL responded with certain answers after the question utterances. SAIL actually generated actions at every time frame based on what it could recall with respect to the audiovisual context it received. The actions could be "no response (keeping quiet)" or responding with one of the answers (object names or sizes). Therefore, it is possible that there were more than one responds during each audiovisual sequence. For each response, if it was correct with

TABLE I
SAIL'S RESPONSES ON QUESTION 1 ("NAME?") WHEN THE QUESTIONS WERE ALIGNED WITH IMAGE FRAME NUMBER 250

Objects/ Answers (%)	None	Baby 1	Baby 2	Kitty	Dwarf	Doggy	Girl	Ape	Hugme	Minnie Mouse	Mickey Mouse	Winnie	Harry Potter
None	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baby 1	0.00	93.33	2.50	6.00	5.00	5.00	7.50	5.00	5.00	5.00	5.00	5.00	5.00
Baby 2	0.00	5.00	86.83	0.00	0.00	0.00	0.00	0.00	24.83	0.00	0.00	0.00	0.00
Kitty	0.00	0.00	0.00	94.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dwarf	0.00	0.00	0.00	0.00	92.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Doggy	0.00	0.00	0.00	0.00	0.00	92.50	0.00	0.00	0.00	0.00	0.00	1.67	0.00
Girl	0.00	0.00	0.00	0.00	0.00	0.00	69.83	0.00	0.00	0.00	0.00	0.00	0.00
Ape	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.00	0.00	0.00	0.00	0.00	0.00
Hugme	0.00	1.67	10.67	0.00	2.50	2.50	0.00	0.00	70.17	0.00	0.00	5.83	0.00
Minnie Mouse	0.00	0.00	0.00	0.00	0.00	0.00	6.67	0.00	0.00	92.50	3.33	0.00	0.00
Mickey Mouse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.50	89.17	0.00	0.00
Winnie	0.00	0.00	0.00	0.00	0.00	0.00	16.00	0.00	0.00	0.00	2.50	87.50	0.00
Harry Potter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	95.00
Big	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Small	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TABLE II
SAIL'S RESPONSES ON QUESTION 2 ("SIZE?") WHEN THE QUESTIONS WERE ALIGNED WITH IMAGE FRAME NUMBER 250. THE ITALIC NUMBERS REPRESENT THE CORRECT RATES

Objects/ Answers (%)	None	Baby 1	Baby 2	Kitty	Dwarf	Doggy	Girl	Ape	Hugme	Minnie Mouse	Mickey Mouse	Winnie	Harry Potter
None	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baby 1	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baby 2	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kitty	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dwarf	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Doggy	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Girl	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00	0.00
Ape	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00	0.00
Hugme	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00	0.00
Minnie Mouse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00	0.00
Mickey Mouse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00	0.00
Winnie	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00	0.00
Harry Potter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.00
Big	0.00	*93.00*	*93.00*	*93.00*	0.00	1.67	15.50	0.00	*93.00*	2.00	*91.00*	*93.00*	2.00
Small	0.00	0.00	0.00	0.00	*93.00*	*91.33*	*77.50*	*93.00*	0.00	*91.00*	2.00	0.00	*91.00*

respect to the object presented at the time, we counted it as correct. So, we obtained the first correct answer rate (C.A.R.1)

$$\text{C.A.R.1} = \frac{n_c}{n_t}$$

where n_c is the number of correct responses and n_t is the total number of responses.

In the second evaluation, we counted each audiovisual sequence as one trial. During each trial, if the majority of the responses were correct, we counted this trial as correct. Otherwise, we counted it as wrong. Here is the second correct answer rate (C.A.R.2)

$$\text{C.A.R.2} = \frac{s_c}{s_t}$$

where s_c is the number of audiovisual sequences with the correct majority responses and s_t is the total number of audiovisual sequences.

1) Results of Algorithm 1: The first experiment was conducted using multimodal learning algorithm 1. We plot the correct answer rates with respect to the question positions during testing in Fig. 16.

Since the objects were mounted on and off the gripper from one image sequence to another image sequence and the objects were rotated by the gripper, when the questions were asked, the audiovisual scenes were never exactly the same during testing as those during training. This difference was further increased with the increase of the question-position difference between training and testing. Overall, SAIL maintained a high correct answer rate when the question-position difference between training and testing is within the expected range. The correct answer rate drops with the question-position difference, but still reasonably well—average above 90% from frame 210 to frame 300 across a time span of about three seconds, compared with a range of 0.33 s to 0.55 s in eyeblink animal classical conditioning learning. As the system needs to respond to any sensory events, two events are not necessarily related if they do not occur concurrently in real-time within a small time window. This is important to avoid temporal hallucination. For example, when the time separation is large (e.g., when objects were moving in or out of SAIL's field of view), SAIL's association rate should be very low.

Then, one may ask, how can a mature human relate two events that occur across a span of larger time scales (e.g., in the scale of days in criminal investigations)? In fact, they still must be recalled from the brain's long term memory so that they coactivate in the brain within a short time span in real-time (e.g., within a second) for the brain to associate them. For example, a sentence from the supervisor mentions both events within a

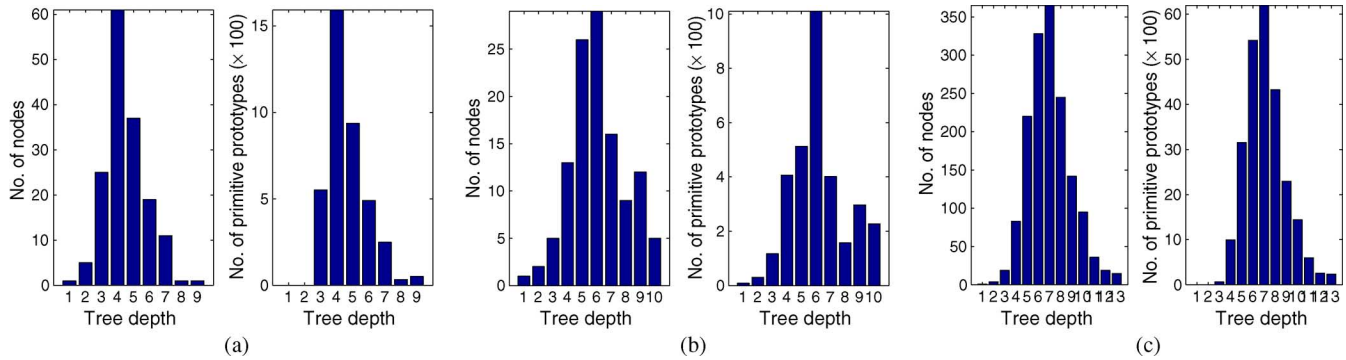

Fig. 17. After trained using multimodal learning algorithm 1, the node distribution (left) and primitive prototype distribution (right) in the trees: (a) P-tree of A-LBE; (b) P-tree of V-LBE; (c) R-tree of H-LBE.

TABLE III
THE CORRECT ANSWER RATE 2 (MAJORITY CORRECT RATE) OF SAIL WHEN THE QUESTIONS WERE ALIGNED WITH IMAGE FRAME NUMBER 250

Objects/ Questions(%)	Baby 1	Baby 2	Kitty	Dwarf	Doggy	Girl
"Name?"	90.0	90.0	90.0	100.0	100.0	90.0
"Size?"	100.0	100.0	100.0	100.0	100.0	80.0
Objects/ Questions(%)	Ape	Hugme	Minnie Mouse	Mickey Mouse	Winnie	Harry Potter
"Name?"	100.0	70.0	100.0	100.0	90.0	100.0
"Size?"	100.0	100.0	100.0	100.0	100.0	100.0

second (e.g., "Is event A and event B related?") or the agent can think so autonomously based on its experience. The work here models the capability of general multimodal association within a moderate time span within a second. The association of events across larger time span requires learning through language, as the above supervisor example indicated. Such a mode of developmental learning is called *communicative learning* [42] and is beyond the scope of the work here.

Detailed confusion tables (Table I and Table II) show SAIL's C.A.R.1 on different objects and different questions when the questions were aligned with image frame number 250. The average C.A.R.1 is 87.9%. The "None" category in the tables corresponds to the period of time when the verbal questions were not asked. Since SAIL never responded during those periods, which is desirable, the percentage is 100. Table III shows SAIL's C.A.R.2 with an average rate of 95.8%.

Comparing Table I with Table II, careful readers may ask why the upper part of Table II does not show the similar confusion pattern as the upper part of Table I. Note the input to H-LBE is the concatenation of the primed sensations from V-LBE and A-LBE, which forms a long vector of $(25\times 20+13\times 20=760)$ dimensions, where 25×20 is the image size, 13 is the order of MFCC, and 20 is the number of MFCC vectors included in one auditory sensation vector as described in page 13. In Table I, (image "Baby 2" + verbal "name") is confused with (image "Hugme" + verbal "name") because they share the "name" part. (Image "Baby 2" + verbal "size") is recognized mistakenly as (image "Baby 2" + verbal "name") because they share the "Baby 2" part. However, the chance that (image "Baby 2" + verbal "size") is confused with (image "Hugme" + verbal "name") is very low because they do not share anything. Therefore, we see the "faultless" diagonal in the confusion Table II.

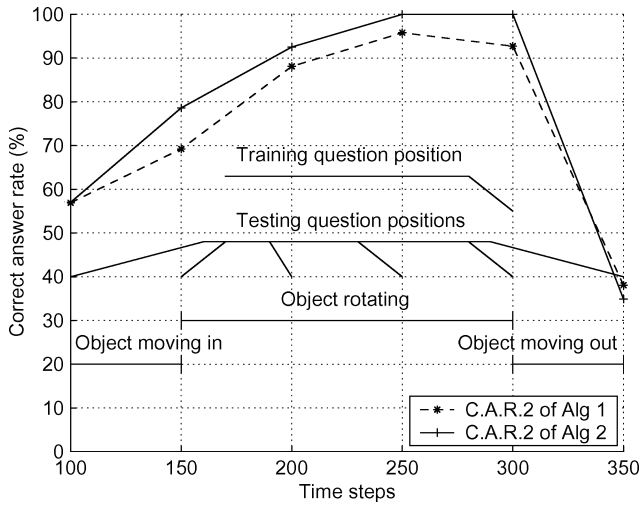

Fig. 18. Comparison of the majority correct answer rates of multimodal learning algorithm 1 and 2.

The size of the whole "brain" after training was 806 MB. The shape of the three major trees of the three LBEs are shown in Fig. 17. Because of the tree structure, the average execution time at each time step is 3.4 ms, which is much lower than 18.1 ms, the interval of a single speech frame and the upper-bound of one execution step.

2) Results of Algorithm 2: The second experiment was conducted using multimodal learning algorithm 2. The majority correct rates for both algorithm 1 and 2 are plotted in Fig. 18 for comparison, in which the improvement is visible. Particularly, when the questions were aligned with image number 250, the performance of the robot improved from 95.8% to 100%. The major difference between algorithm 2 and algorithm 1 is that, in algorithm 2, the inputs to H-LBE contain the primed action pattern. As explained in Section IV-D, the primed action pattern catches the characteristic information, i.e., the different questions instead of the different ways of uttering the same question. This abstraction process enabled a better performance. In addition, as shown in Fig. 19, the size of R-tree in H-LBE is significantly reduced compared to that of multimodual algorithm 1 shown in Fig. 17 because the decision boundary was simplified due to the inclusion of the primed action pattern in the input to H-LBE.

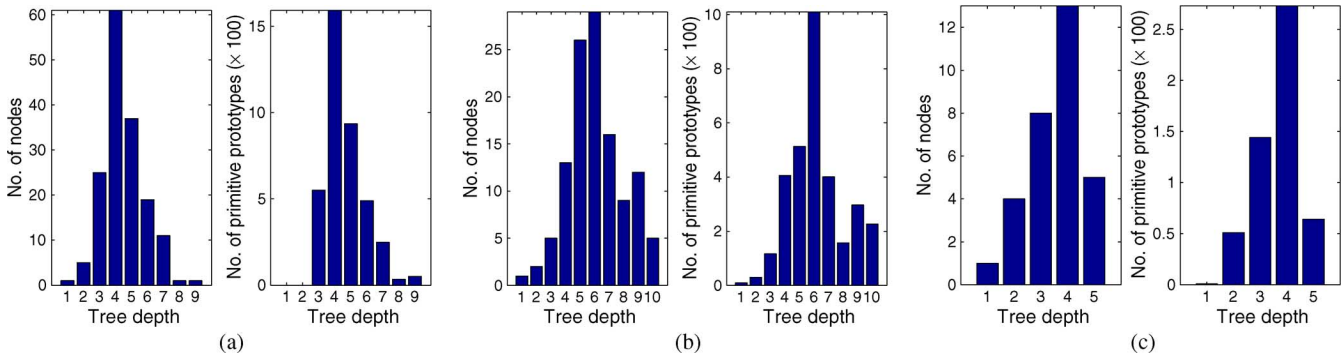

Fig. 19. After trained using multimodal learning algorithm 2, the node distribution (left) and primitive prototype distribution (right) in the trees: (a) P-tree of A-LBE; (b) P-tree of V-LBE; (c) R-tree of H-LBE.

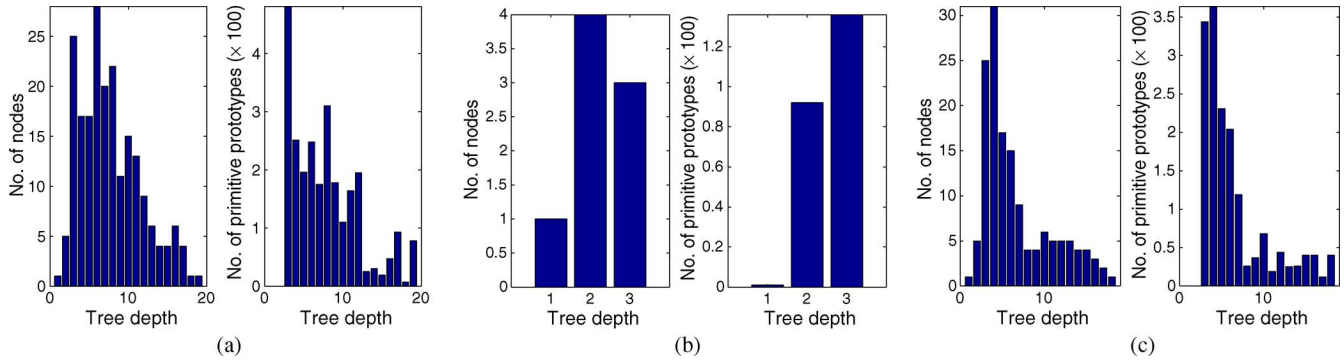

Fig. 20. Node distribution (left) and primitive prototype distribution (right) in the trees: (a) P-tree of A-LBE; (b) P-tree of V-LBE; (c) R-tree of H-LBE.

B. Real-Time Experiment

In the real-time experiment, the verbal questions ("name?" and "size?") were asked followed by the answers imposed through the switch sensors of SAIL. For each object, we usually issued each question five to six times. To make it easier for the trainer to see the response of SAIL, we manually mapped SAIL's action vectors to the names of the objects and used Microsoft text-to-speech software to speak the names. After going through three randomly chosen objects (baby 1, dwarf, and girl), the objects were mounted on the gripper again and the questions were asked without giving the answers. We repeated the above process ten times and SAIL responded correctly approximately 90% of the time for the three trained objects.

The shape of the three major trees of the three LBEs are shown in Fig. 20. The P-tree of V-LBE is fairly small compared to the P-tree of A-LBE because SAIL's eyes focused on a small field of view covering the object and did not experience dramatic changes. In contrast, the microphone of SAIL collected the conversation of the trainer with his lab mate in addition to the verbal questions. The size of the whole "brain" containing three LBEs is about 60 MB after the above training process.

While extending the real-time experiment to more objects, the system experienced unacceptable delay. We recorded the execution time to get an idea of the speed performance of SAIL. Fig. 21(a) shows that the average execution time of each step over 50 consecutive steps grew at the beginning and flattened out after about 100 s. The short surging periods around the 100 s, 150 s, and 210 s were during the times we changed the objects. Since the visual context changed a lot at the time, the trees conducted extensive learning and required more time in each execution step. But even in these periods, the execution time of each step was lower than 18.1 ms, the interval of a single speech frame and the upper-bound of one execution step.

Failing to discover the reason for delay, we plotted the execution time in each step without averaging in Fig. 21(b). As we can see, in a few time steps, SAIL's step execution time bursts to as high as 95 ms, which is five times a single speech frame. The frequency of this kind of burst grows with the size of total memory consumption as shown in Fig. 21(b). So is the longest step execution time. For example, after the 806 MB "brain" is trained on recorded is loaded for testing, the longest step execution time grows to as high as about 600 ms. We believe that these bursts of step execution time is related to memory paging. Frequent long steps accumulate and caused the long delay we mentioned above. Without accessing the kernel of the operating system, we were not able to conduct real-time experiments on more objects.

C. Discussion on Scalability

There are two types of scalability of concern. The first one is the computational complexity. At the component level of the system, HDR has the time complexity as low as $O(\log n)$, where n represents the number of distinguishing training data [36]. The space complexity of HDR is $O(n)$, which is not an issue since the digital storage is getting very cheap. At the architecture level, when the length of the audiovisual context required

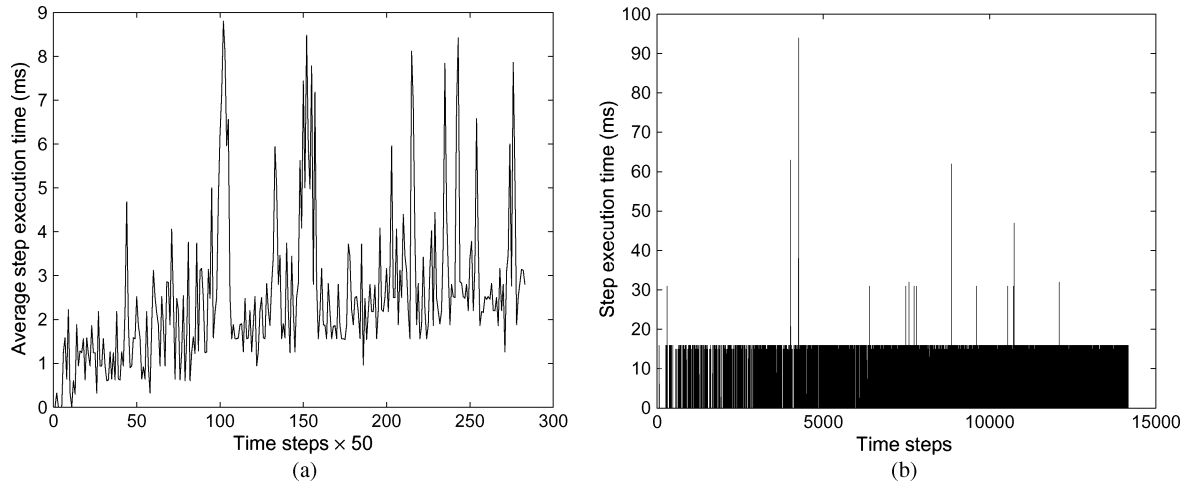

Fig. 21. Average step execution time of over 50 consecutive time steps (a) and the actual step execution time (b) of the multimodal learning system.

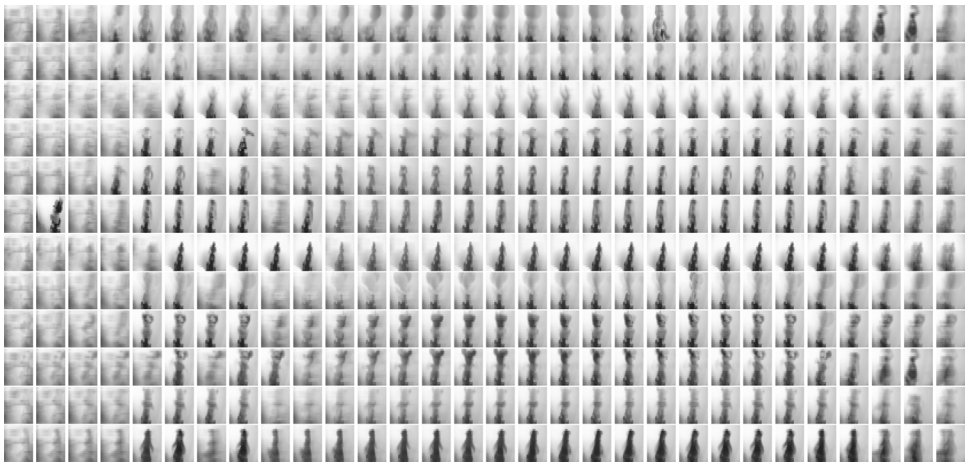

Fig. 22. Sample sequences of the primed sensation of 12 objects.

to answer the questions gets longer, (e.g., a long question sentence), the system potentially needs even higher levels to handle, which is still an ongoing research topic. But adding more short questions, such as "color," will not pose any serious problems.

The second scalability concern is regarding the discriminant power due to the clustering along the temporal context trajectory. We have discussed that the theoretical number of possible images in the image vector space is about 1.32×10^{1204}, which means a high representation power. It is hardly possible for the same object to generate two exactly same images. This is the warranty of the scalability of the appearance-based method. In this sense, the major issue of the appearance-based method is to find the invariance of different views of the same object instead of the scalability. One of the popular approaches to this is the subspace method, such as using PCA to find eigenface. HDR uses the idea of discriminant analysis and the detailed discussion of the features extracted by HDR can be found in [43]. However, the appearance-based method requires the alignment of images.

The proposed method of clustering along the temporal context trajectory reduces the requirement of the exact alignment as discussed and shown in previous sections. Inevitably, doing clustering will reduce the resolution of the visual representation. However, be aware that the cluster centers are generated by multiple images along the temporal trajectory of each individual moving object and, therefore, they contain extra temporal information unique to individual objects. In addition, the clustering threshold can always be controlled to ensure the resolution. While not yet addressed in the proposed system, we believe the threshold should be context-dependent and can be learned by the system.

To give the reader some intuitive idea of the affect of the clustering, we plot one sample sequence of primed sensation for each object in Fig. 22. Due to space limit, we can not show the images with higher resolution. But the difference between the primed sensation of different object is still visible, and, therefore, distinguishable by the system.

It is important to note that the learning results are highly related to teaching experience. For example, the direction of rotation is also an important aspect of experience. Think of child's "count down" skill, which need to be learned separately by the child.

VI. Conclusion

In this paper, we introduce a multimodal autonomous learning architecture that enables a machine learner to develop

and integrate vision and audition concurrently in a dynamic, interactive training setting. With this architecture, after being taught the answers to verbal questions upon the presence of objects, the SAIL robot was able to answer the questions correctly even when the orientation of the objects was changed. It is worth emphasizing that the system did not have any prior knowledge about the objects or the verbal questions before it started running ("birth"). Nor does it require the sensory sequences to be transcribed, in contrast with almost all existing speech recognition systems which used an open "skull" approach. As far as we know, there has been no such published multimodal developmental learning system, while earlier preliminary work of this line of work appeared in 2003 [44]. The work presented here seems the first attempt focusing on computational multimodal developmental learning. The sensitivity and a desired amount of tolerance to interstimuli interval have been discussed and demonstrated.

This represents a major departure from traditional modes of machine learning that uses handcrafted features. For example, unlike many traditional speech processing methods, it does not use audio silence as a cut point to isolate words, since continuous speaking sentences often do not have any interword silence. Unlike many traditional visual processing methods, it does not use a motion static point as a key frame for segmenting video sequences, since such static points do not always exist or reliable in real-world events. The capability of learning directly from raw, untranscribed multimodal sensory streams is necessary for autonomous development.

The multimodal developmental paradigm, the architecture, the level-building elements for the corresponding functions of the cortical areas, and the experimental results presented here indicate that multimodal learning directly from raw sensory and motor streams are computationally feasible.

APPENDIX

The relative entropy, or Kullback–Leibler distance between two densities f and g is defined by

$$D\left(f(x)\|g(x)\right) = \int f(x)\log\frac{f(x)}{g(x)}dx.$$

Thus

$$\begin{aligned}
&D\left(f(x,y)\|g(x,y)\right)\\
&\quad = \int f(x,y)\log\frac{f(x,y)}{g(x,y)}dxdy.\\
&D\left(f(x,y)\|g(x,y)\right) - D\left(f(x)\|g(x)\right)\\
&\quad = \int f(x,y)\log\frac{f(x,y)}{g(x,y)}dxdy - \int f(x)\log\frac{f(x)}{g(x)}dx\\
&\quad = \int f(x,y)\log f(x,y)dxdy - \int f(x)\log f(x)dx\\
&\qquad + \int f(x,y)\log\frac{g(x)}{g(x,y)}dxdy\\
&\quad = h(X,Y) - h(X) + \int f(x,y)\log\frac{g(x)}{g(x,y)}dxdy\\
&\quad = h(Y|X) + \int f(x,y)\log\frac{g(x)}{g(x,y)}dxdy\\
&\quad \geq 0 \qquad (11)
\end{aligned}$$

where $h(\cdot)$ is the differential entropy. The strict inequality holds except for the degenerated case where the second term in (11) is equal to zero, which requires that $f(x,y)\log(g(x)/g(x,y))$ equals zero almost everywhere.

REFERENCES

[1] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, pp. 599–600, Jan. 26, 2001.

[2] R. A. Brooks, "A robust layered control system for a mobile robot," *IEEE J. Robot. Autom.*, vol. RA-2, no. 1, pp. 14–23, Mar. 1986.

[3] R. C. Arkin, *Behavior-Based Robotics*. Chambridge, MA: MIT Press, 1998.

[4] N. Mercer, *Words and Minds: How We Use Language to Think Together*. London, U.K.: Routledge, 2000.

[5] H. Hirsch and D. Spinelli, "Modification of the distribution of receptive field orientation in cats by selective visual exposure during development," *Exp. Brain Res.*, vol. 13, pp. 509–527, 1971.

[6] B. Bertenthal, J. Campos, and K. Barrett, "Self-produced locomotions: An organizer of emotional, cognitive, and social development in infancy," in *Continuities and Discontinuities in Development*, R. Emde and R. Harmon, Eds. New York: Plenum, 1984.

[7] J. Weng, W. Hwang, Y. Zhang, C. Yang, and R. Smith, "Developmental humanoids: Humanoids that develop skills automatically," in *Proc. 1st IEEE-RAS Int. Conf. Humanoid Robot.*, Boston, MA, Sep. 7–8, 2000.

[8] Y. Zhang and J. Weng, "Grounded auditory development of a developmental robot," in *Proc. INNS-IEEE Int. Joint Conf. Neural Netw.*, Washington, DC, Jul. 14–19, 2001, pp. 1059–1064.

[9] M. Johnson, *The Body in the Mind: The bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: Univ. Chicago Press, 1974.

[10] S. Harnard, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[11] D. Chalmers, "Subsymbolic computation and the Chinese room," in *The Symbolic and Connectionist Paradigms: Closing the Gap*, J. Dinsmore, Ed. Hillsdale, NJ: Erlbaum, 1992, pp. 25–48.

[12] K. Allan, *Natural Language Semantics*. Malden, MA: Blackwell, 2001.

[13] P. Violi, *Meaning and Experience*. Bloomington, IN: Indiana Univ. Press, 2001.

[14] J. Weng, "On developmental mental architectures," *Neurocomputing*, vol. 70, no. 13–15, pp. 2303–2323, 2007.

[15] V. de Sa and D. Ballard, "Category learning through multimodality sensing," *Neural Commun.*, vol. 10, pp. 1097–1117, 1998.

[16] T. Huang, L. Chen, and H. Tao, "Bimodal emotion recognition by man and machine," in *Proc. ATR Workshop Virtual Commun. Environments*, Kyoto, Japan, Apr. 1998.

[17] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cogn. Sci.*, vol. 26, no. 1, pp. 113–146, 2002.

[18] R. Baillargeon, "Object permanence in 3.5- and 4.5-month-old infants," *Develop. Psychol.*, vol. 23, pp. 655–664, 1987.

[19] C. I. Baker, C. Keysers, J. Jellema, B. Wicker, and D. Perrett, "Neuronal representation of disappearing and hidden objects in temporal cortex of macaque," *Exp. Brain Res.*, vol. 140, pp. 375–381, 2001.

[20] H. Bulthoff and S. Edelman, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," *Nat. Acad. Sci.*, vol. 89, pp. 60–64, Jan. 1992.

[21] N. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current Biol.*, vol. 5, no. 5, pp. 552–563, 1995.

[22] J. Pauls, E. Bricolo, and N. Logothetis, "View invariant representations in monkey temporal cortex: Position, scale, and rotational invariance," in *Early Visual Learning*, S. K. Nayar and T. Poggio, Eds. Oxford, U.K.: Oxford Univ. Press, 1996, pp. 9–41.

[23] J. T. Green, R. B. Ivry, and D. S. Woodruff-Pak, "Timing in eyeblink classical conditioning and timed-interval tapping," *Psychol. Sci.*, vol. 10, no. 1, pp. 19–25, 1999.

[24] A. B. Steinmetz and C. R. Edward, "Comparison of auditory and visual conditioning stimuli in delay eyeblink conditioning in healthy young adults," *Learn. Behav.*, vol. 37, pp. 349–356, 2009.

[25] D. Marr and H. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Roy. Soc. London. Series B: Biol. Sci.*, vol. 200, pp. 269–294, 1978.

[26] I. Pavlov, *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. London, U.K.: Oxford Univ. Press, 1927.

[27] M. Cole and S. R. Cole, *The Development of Children*, 3rd ed. New York: Freeman, 1996.
[28] P. Bloom, *How Children Learn the Meaning of Words*. Cambridge, MA: MIT Press, 2000.
[29] R. Sutton and A. Barto, *Reinforcement Learning—An Introduction*. Chambridge, MA: MIT Press, 1998.
[30] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cereb. Cortex*, vol. 1, pp. 1–47, 1991.
[31] M. Mishkin, L. G. Unterleider, and K. A. Macko, "Object vision and space vision: Two cortical pathways," *Trends Neurosci.*, vol. 6, pp. 414–417, 1983.
[32] J. Weng and W. Hwang, "Incremental hierarchical discriminant regression," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 397–415, Feb. 2007.
[33] J. Weng and M. Luciw, "Dually optimal neuronal layers: Lobe component analysis," *IEEE Trans. Autonom. Mental Develop.*, vol. 1, no. 1, pp. 68–85, May 2009.
[34] D. R. Cox, "Statistical analysis of time series: Some recent developments," *Scand. J. Statist.*, vol. 8, no. 2, pp. 93–115, 1981.
[35] S. Zeger and B. Qaqish, "Markov regression models for time series: A quasi-likelihood approach," *Biometrics*, vol. 44, no. 4, pp. 1019–1031, Dec. 1988.
[36] W. Hwang and J. Weng, "Hierarchical discriminant regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1277–1293, Nov. 2000.
[37] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[38] C. J. Watkins, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.
[39] Y. Zhang and J. Weng, "Action chaining by a developmental robot with a value system," in *Proc. IEEE 2nd Int. Conf. Develop. Learn.*, Cambridge, MA, Jun. 12–15, 2002, pp. 53–60.
[40] E. Kandel, J. Schwartz, and T. Jessell, *Principles of Neural Science*, 3rd ed. Norwalk, CT: Appleton, 1991.
[41] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, 2nd ed. New York: IEEE Press, 2000.
[42] J. Weng, "Developmental robotics: Theory and experiments," *Int. J. Humanoid Robot.*, vol. 1, no. 2, pp. 199–235, 2004.
[43] J. Weng and W. Hwang, "An incremental learning algorithm with automatically derived discriminating features," in *Proc. 4th Asian Conf. Comput. Vis.*, Taipei, Taiwan, Jan. 8–9, 2000, pp. 426–431.
[44] Y. Zhang and J. Weng, "Conjunctive visual and auditory development via real-time dialogue," in *Proc. 3rd Int. Workshop Epigenetic Robot.*, Boston, MA, Aug. 4–5, 2003, pp. 974–980.

Yilu Zhang (S'99–M'02–SM'08) received the B.Sc. and M.Sc. degrees in electrical engineering from Zhejiang University, Zhejiang, China, in 1994 and 1997, respectively, and the Ph.D. degree in computer science from Michigan State University, East Lansing, in 2002.

He joined General Motors Global R&D Center, Warren, MI, in 2002, and currently holds the position of staff researcher. His research interests include statistical pattern recognition, machine learning, signal processing, and their applications, including human machine interactions and integrated vehicle health management.

Dr. Zhang served as an Associate Editor of the *International Journal of Humanoid Robotics* from 2003 to 2007, the Publication Chair for the IEEE 8th International Conference on Development and Learning 2009, and is currently a member of the Autonomous Mental Development Technical Committee of the IEEE Computational Intelligence Society. In 2008, he received the "Boss" Kettering Award—the highest technology award in GM—for his major contribution to Connected Vehicle Battery Monitor, a remote vehicle diagnostics technology.

Juyang Weng (S'85–M'88–SM'05–F'09) received the B.Sc. degree in computer science from Fudan University, Shanghai, China, in 1982, and the M.Sc. and Ph.D. degrees in computer science from the University of Illinois at Urbana-Champaign, in 1985 and 1989, respectively.

He is currently a Professor of Computer Science and Engineering at Michigan State University, East Lansing. He is also a faculty member of the Cognitive Science Program and the Neuroscience Program at Michigan State University. Since the work of Cresceptron (ICCV 1993), he expanded his research interests in biologically inspired systems, especially the autonomous development of a variety of mental capabilities by robots and animals, including perception, cognition, behaviors, motivation, and abstract reasoning skills. He has published over 200 research articles on related subjects, including task muddiness, intelligence metrics, mental architectures, vision, audition, touch, attention, recognition, autonomous navigation, and other emergent behaviors.

Dr. Weng is an Editor-in-Chief of the *International Journal of Humanoid Robotics* and an Associate Editor of the IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, as well as a member of the Executive Board of International Neural Network Society. He was a Program Chairman of the NSF/DARPA funded Workshop on Development and Learning 2000 (1st ICDL), the Chairman of the Governing Board of the International Conferences on Development and Learning (ICDL) (2005–2007), Chairman of the Autonomous Mental Development Technical Committee of the IEEE Computational Intelligence Society (2004–2005), a Program Chairman of 2nd ICDL, a General Chairman of 7th ICDL (2008) and 8th ICDL (2009), an Associate Editor of the IEEE TRANSACTIONS ON PATTERN RECOGNITION AND MACHINE INTELLIGENCE, and an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He and his co-workers developed SAIL and Dav robots as research platforms for autonomous development.