Developing Early Senses About the World: "Object Permanence" and Visuoauditory Real-time Learning

Juyang Weng, Yilu Zhang, and Yi Chen
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48823, USA
Email: weng@cse.msu.edu

***Abstract*— What "constraints" are exactly wired into the human developmental program? What "constraints" are minimally necessary for a developmental robot? These are open questions. In this paper, we propose a mechanism of developing experience-based priming – predicting the future contexts including sensation and action based on the previous experience – as a powerful "constraint" for developmental robots. We present an architecture that develops this priming capability through real-time online interactions with the environment. We report how our SAIL robot developed a sense of novelty in a well-known "drawbridge" experiment which sheds light on the controversial issue of "Object Permanence" in psychology. We further show how the proposed priming mechanism enabled SAIL to deal with a very challenging online learning setting: learning the name and property (e.g., size) of dynamically rotating objects through verbal dialogues.**

I. Introduction

Multimodal priming is the capability to internalize and organize related high-dimensional contexts, retrieve the future ones and associate them to produce corresponding behaviors. This priming mechanism relies on the following three important requirements:

The first is to quickly, effectively and incrementally generate the *internal representation* from real-time experience. In our model, the internal representation corresponds to high dimensional numerical vectors in a series of transformed spaces, simulating to a degree the response patterns of neurons. Each space corresponds to a set of automatically derived (instead of hand-programed) high-dimensional discriminating features. This is essential for the success of generalization in predicting future contexts.

The second one is multimodal integration. Recently this issue has raised great interests and many promising results have been reported [1] [2]. However, existing methods are based on hand-defined symbolic mapping while multimodal integration of real-time sensory streams is still a challenging unaddressed open problem since all perceptual units are highly dependent on each other.

The third one is grounding. This requirement is important because the representation inside an agent should be intimately generated from sensory experience of the external physical world [3] [4] [5] [6] [7]. Grounding becomes very challenging when the representation has to be complete: sufficient for any potential environment, due to the task-nonspecific nature of autonomous mental development (AMD) at the robot's programming time (before "birth").

In this paper we propose a powerful general mechanism called *priming architecture*, which can enable a robot to predict future contexts reliably. The design and implementation of this priming architecture follow the AMD paradigm [8]. One major foundation of AMD is that a robot should not be programmed to conduct known tasks in a known environment. Instead, it should possess a general task-nonspecific learning capability, and develop task-specific representation and skills through real world sensory experience. The architecture is implemented on our house-made human-size mobile robot named SAIL (Fig. 4 (b)). Using the priming architecture and other components in its developmental architecture, SAIL is able to develop perceptual capabilities that have interesting early implications to "Object Permanence" issue and visuoauditory capabilities in a challenging learning setting that a robot has never successfully tried before.

II. Architecture and Algorithm

The proposed architecture is built upon a regression engine called incremental hierarchical discriminant regression (IHDR) [9] [10]. Because space is limited, we cannot go into the details. Briefly, the IHDR technique automatically derives discriminating feature subspaces in a coarse-to-fine manner from a high-dimensional input space to generate a tree architecture of self-organized memory. It can handle high-dimensional data (e.g., thousands) in real-time, which is crucial to a developmental robot.

A. Level-building Element

Using IHDR trees, we designed a basic developmental framework called level-building element (LBE). Shown in Fig. 1 is an example of a LBE taking two channels of sensory inputs, the auditory sensation and the action sensation.

We call the high-dimensional input vector to an IHDR tree the last context $s(t)$, and the output the primed context $p(t)$. The primed context p consists of three parts, a primed sensation vector p_s, a primed action vector p_a, and an associated value. An IHDR tree approximates a mapping g so that:

$$p(t) = g(s(t)).$$

The LBE needs two trees. The upper one is called the reality tree or *R-tree* and the bottom one the priming tree or *P-tree*. The R-tree predicts immediate near future contexts and the P-tree generates far future ones. The prototype updating queue (PUQ) for the P-tree keeps a "live" context trajectory so

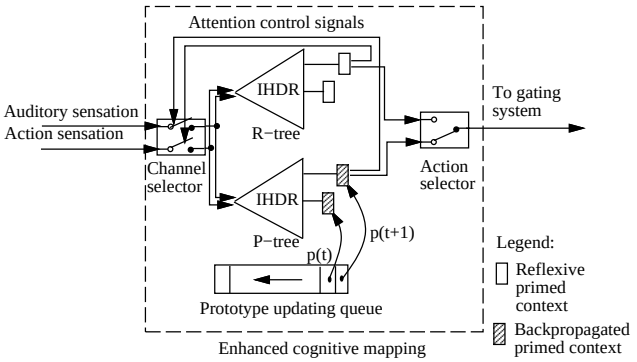


Fig. 1. A level-building element.

that the future contexts can be recursively "blurred" into the "predictor" of early contexts. To accomplish this, PUQ keeps a list of pointers to the consecutive contexts in its first-in-first-out queue. Mathematically, the primed contexts that PUQ point to are updated with a recursive model adapted from Q-learning [11]:

$$p^{(n)}(t) = \frac{n-1-l}{n} p^{(n-1)}(t) + \frac{1+l}{n} \gamma p^{(n-1)}(t+1), \quad (1)$$

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 the time-discount rate. l is an amnesic parameter used to give more weight on the newer data points, which is typically positive, e.g., $l = 2$.

The above equation shows that a primed context $p^{(n)}(t)$ is updated by averaging its last version $p^{(n-1)}(t)$ and a time-discounted version of the current primed context $p^{(n-1)}(t+1)$. In this way, the information embedded in the future context $p^{(n-1)}(t+1)$ is recursively back-propagated into earlier ones through PUQ. Therefore, when an earlier context is recalled, the future information will be primed.

B. Novelty in the Value System

The novelty is measured by the disagreement between what is predicted by the P-tree and what is really seen by the R-tree. If the robot can predict well, the novelty is low. Algorithmically, we define novelty as the normalized distance between the selected primed sensation $p^{(n)}(t) = (p_1^{(n)}(t), p_2^{(n)}(t)...p_m^{(n)}(t))$ and the actual sensation $p(t+1)$ at the next time:

$$n(t) = \sqrt{\frac{1}{m} \sum_{j=1}^{m} \frac{(p_j^{(n)}(t) - p_j(t+1))^2}{\sigma_j^2(t)}}, \quad (2)$$

where m is the dimension of sensory input. Each component is divided by the expected variance σ_j^2, which is the time-discounted average of the squared difference $(p_j^{(n)}(t) - p_j(t+1))^2$. It should be noticed that when the signal vector represents a high-level concept, the novelty is for high-level too. This mechanism enables the robot to build up the perceptual representation of novelty as part of its value system in addition to the rewards.

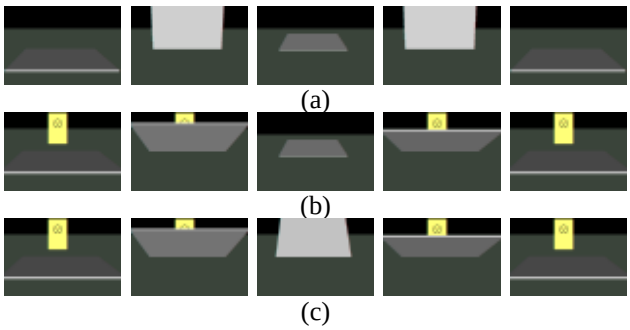


Fig. 2. The animated settings in "drawbridge". (a) 180NB Habituation Event (b) 180B Impossible Event (c) 120B Possible Event.

III. "Object Permanence"

A. "Drawbridge" and "Object Permanence"

The term "Object Permanence" is first introduced by Jean Piaget in 1954, namely, the ability to understand that objects continue to exist even when they are no longer visible. In Piaget's view, "Object Permanence" is a gradual process that lasts from the first month of life through 18-24 months [12]. However, contemporary researchers have suggested that infants may have a sense of "Object Permanence" as young as $3\frac{1}{2}$ months old or earlier [13] [14]. Therefore, whether "Object Permanence" is an innate sense or a later developed ability has been widely debated in psychology. The well-known experiment called "drawbridge" by Baillargeon and her colleagues has been in the center of this debate for many years. Recent studies by Bogartz, Schilling, Cashon, etc. suggested that the infants' behavior in the "drawbridge" experiment reflects the perceptual capacity instead of innate knowledge for "Object Permanence" [15] [16]. Researches in neuroscience also give supports to this perceptual view since a neuronal population in the anterior superior temporal sulcus (STSa) has been identified to be selectively activated by an object as it gradually disappears from view behind an occluding screen [17]. This means that " the prolonged activity during the hidden period appears to arise from the visual event of gradual occlusion rather than from the fact that the object were not visible." The open question is then: what "innate" mechanism if not knowledge is in the infants' developmental program that gives rise to such an early perceptual capability? Before this question is answered clearly in terms of computational neuroscience, which is extremely difficult, we would like to conduct the "drawbridge" experiment on our SAIL robot first. We interpret the above neural activity as novelty detection. Although the consistency between SAIL robot and human infants does not mean they are using the same mechanism, the robot's result does shed some light on this important issue.

B. "Drawbridge" Experiment

For precise control of the temporal traces, we created an OpenGL 3-D animation(Fig. 2) for the "drawbridge" experiment originally designed by Baillargeon et al. [13] [14]. The three events involved settings with a grey table, a silver screen rotating back and forth at either 120° or 180° , a yellow box that was either set in the path of the screen or absent

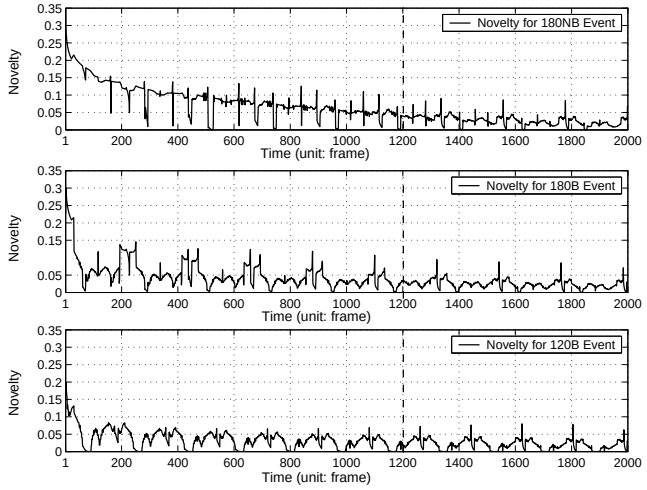


Fig. 3. Novelties measured in the "drawbridge" experiment.

from the event. In the first habituation event (180NB), the box was absent and the screen had a full rotation of $180°$ with 4 sec. away from the robot, 1 sec. pause in the middle and 4 sec. back toward the robot. In the impossible event (180B), the box was presented and the screen moved all the way to $180°$ as if the box were not there. In the possible event (120B), the screen only rotated up to the position of the box at $120°$ and then back. The above three events were shown to the robot sequentially one after another as consistent with the original experiment. Fig. 3 showed the novelty that was detected on SAIL during each of the three events. As shown, after habituating to the 180NB event, the robot found more novelty in the impossible event than in the possible one during the first 1200 image frames of each event. After that, the robot was about habituated and hence, the difference in the novelty decreases. This result is consistent with the psychological findings that human babies looked longer at the impossible event [13] [14]. We have also changed the order at which the last two events were presented and conducted the control condition event in [13]. Our findings indicated that the robot does not show any preference for larger movement and the order of events does matter in the experiment results, which is consistent with the recent more detailed study [18]. Therefore, at least we have established the first developmental robot evidence that the "Object Permanence" reported by Baillargeon can be duplicated by a novelty measurement. If $3\frac{1}{2}$-month-old infant knowledge were so robust as real "Object Permanence," then the order of events should not have altered the relative attention duration as reported in [13] [18]. Of course, we do not claim that our robot uses the same mechanism as human infants do.

IV. Multimodal Learning

A. Problem Description

The above study is important to the debate of "innate physical knowledge" vs. perceptual novelty. However, effective developmental learning for later years must be conducted in much complex multimodal (e.g., vision, audition and touch) contexts. Our previous work has developed the real-time grounded audition learning capability on SAIL robot [6]. However, real-time multimodal learning has raised new challenges:

Visual representation of objects. To perceive objects correctly in the environment, a robot must be able to recognize the objects from different orientations, e.g., when the object is being rotated. Since our goal is for a robot to learn new things "on the fly" without any human pre-designed representation, we use the sensory-signal-centered representation rather than the monolithic object-based representation. See [19] for a discussion and [10] for its use in visual developmental learning.

Imperfect alignment. In the real world, the visual presence of an object is usually coupled with the related auditory signals, such as a verbal name given by a teacher. However, this coupling is not strict in the following senses: (1) Auditory signals spread across many frames. (2) The observer may view an object from different angles when the name was taught and asked at different times. (3) Most of the view angles receive no auditory signals at all.

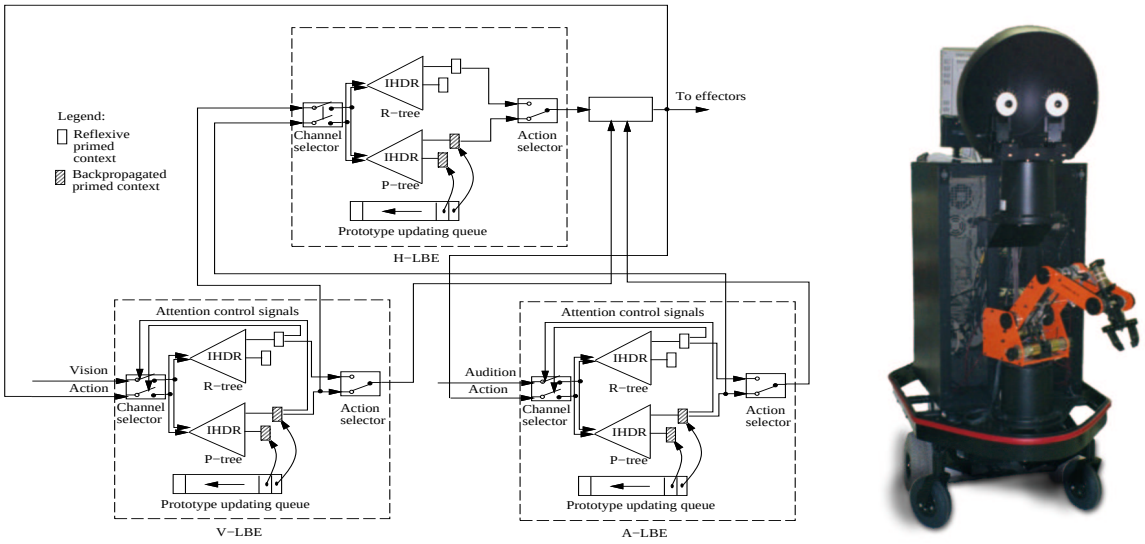
Fortunately, there is a very important property of the physical world we may take advantage of, i.e., the time continuity. In the real world, an object does not emerge from nothing and it does not disappear like a magic. We may make use of the shared image features of the spatiotemporally contiguous views of an object to generate more abstracted representation for multimodal priming.

B. Multimodal Integration through Experiences

Fig. 4 (a) shows the augmented architecture we used for multimodal learning. It has three LBE modules, a vision LBE (V-LBE), an audition LBE (A-LBE), and a high-level LBE (H-LBE). The primed contexts from the P-trees of both V-LBE and A-LBE are inputs to H-LBE. After the low-pass trajectory-wise filtering in PUQ, the primed context only keeps the low-frequency components of the last context. The underlying idea of such an architecture is that while A-LBE and V-LBE may work individually to do certain learning, their combination in H-LBE can resolve the ambiguous situations when neither of the two modalities along, vision or audition, contains enough information for decision-making. For example, when a verbal question ("name?") is asked about an object that the robot is looking at, neither vision nor audition along is able to generate the desired answer.

A high-level outline of the algorithm is as follows:

1) Collect the sensation from the auditory sensor $x_s(t)$, the visual sensor $x_v(t)$, and the action sensor $x_a(t)$.
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, select and denote the ones with the highest primed value as $p_v(t)$ and $p_s(t)$, respectively.
4) Update the PUQs of both V-LBE and A-LBE using equation (1).
5) Take the primed contexts $p_v(t)$ and $p_s(t)$ as the inputs to H-LBE.
6) Update the R-tree of H-LBE using the IHDR learning algorithm.



(a) (b)

Fig. 4. (a) The multimodal learning system architecture. (b) The SAIL robot at Michigan State.

7) Retrieve the R-tree of H-LBE to get a list of primed contexts, select and denote the one with the highest primed value as $p_h(t)$. The primed action part of $p_h(t)$ is sent to effectors for execution.

C. Multimodal Abstraction

The underlying reason that why the above architecture and algorithm work, as shown in the experiment, is that the primed visual sensation is a blurred version of the real visual sensation. As a result, the inputs to H-LBE do not change a lot when the same object is presented. This is an abstraction process, where the cognitive activities reduce the variance of sensations and keep the invariant components.

Why do we send not just primed sensations but also primed actions to H-LBE? Before answering this question, let us first define the primed action pattern of A-LBE as:

$$P_s(t) = \sum_{i=1}^{n} q_{si}(t) p_{si}(t),$$

where n is the total number of primed contexts retrieved from the P-trees of A-LBE, $q_{si}(t)$ is the primed value associated with the ith primed context. Similarly, for V-LBE, we have:

$$P_v(t) = \sum_{i=1}^{n} q_{vi}(t) p_{vi}(t).$$

Then let the random variables P_s and P_a represent the primed sensation and primed action, respectively; $f(P_s)$ and $g(P_s)$ represent the p.d.f.s for "name" and "size," respectively; and $f(P_s, P_a)$ and $g(P_s, P_a)$ the joint p.d.f.s for "name" and "size," respectively. With the above definitions, the reason can then be explained in terms of information theory since we prove in the appendix:

$$D(f(Ps, Pa)||g(Ps, Pa)) \geq D(f(Ps)||g(Ps)),$$

where $D(\cdot)$ is the Kullback-Leibler distance (relative entropy) between the two p.d.f.s. In other words, by including primed actions in the inputs to H-LBE we increase the discriminant power of the representation.

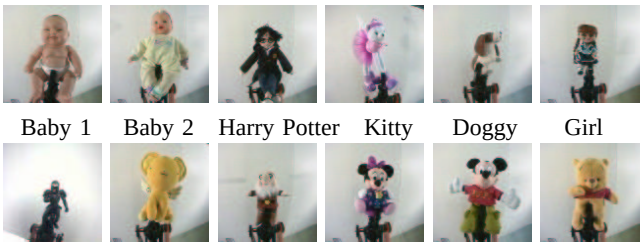


Fig. 5. The objects used in the experiment.

D. Visuoauditory Experiments

The experiment was done in the following way. After the robot started running, the trainers mounted objects one after another onto the gripper of the robot and let the robot rotate the gripper in front of its eyes at the speed of about 3.6s per round. During rotation, the trainers verbally asked the questions of "name?" and "size?" and then appropriate answers were given by the trainers through switch sensors on the robot. Different switch sensor status represented different answers. Since the objects were being rotated, moved in and out of the robot's field of view continuously, the orientation and the positions of the objects kept changing. The robot could then hardly see the same image when the same question was asked again. Totally 12 objects were presented (Fig. 5) to the robot and we expected it to correctly answer questions after being taught by the trainers. A sample video sequence seen by the robot is shown in Fig. 6.

To examine the behavior of the robot in detail and evaluate the performance, the experiment was first done on pre-recorded sensory streams. The image data included five video sequences of every object with 350 frames in each sequence. The auditory data was collected from 63 people

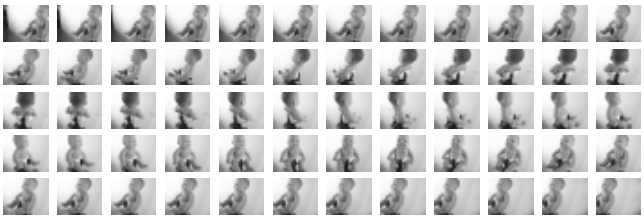


Fig. 6. Part of a sample image sequence.

with a variety of nationalities (American, Chinese, French, Indian, Malaysian, and Spanish) and ages (from 18 to 50). During training, ten subjects were randomly selected and the switch sensor inputs (numerical vectors) were given after the utterances were finished. Of all the five sets of image and speech data, we used four for training and the left-out one for testing. So, with 12 objects, ten people, and two questions, the robot was taught 960 times in training and evaluated for 240 times in testing.

To emulate the situation that the trainers would not be able to ask questions consistently to synchronize the object views, the end point of each question was aligned with image No. 300 during training and with image Nos. 100, 150, 200, 250, and 300, respectively, during testing (Fig. 7).

The correct answer rate (C.A.R.) for evaluating the robot's behaviors can be written as:

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

where n_c is the number of image sequences with correct majority responses and n_t is the total number of image sequences. We denote the rate for the algorithm using only primed sensation as C.A.R.1 and that using both primed sensation and primed action as C.A.R.2. The correct answer rates from SAIL robot are reported in Fig. 8.

Particularly, when the questions were aligned with image frame No. 250, the C.A.R.1 and C.A.R.2 of the robot are 95.77% and 100%, respectively. When the question-position difference between training and testing was not large, the robot maintained a high correct answer rate. With the increase of the difference, the correct answer rate dropped gradually. Also, the robot's performance was low during the time when the objects were moving in or out of the robot's view field since no attention mechanism was used here.

With the recursive averaging over the consecutive primed contexts, the primed sensation was a blurred version of the real visual sensation (Fig. 9). This low-pass-filtering property of PUQ in V-LBE helps to filter out the high-frequency components in the visual sensation and gives a low level abstraction. Therefore, the robot was able to answer the question correctly even it was taught while another pose of the object was seen.

In the real-time experiment, the visual data were captured by a CCD camera at 30 frames per second and speech data were digitized at 11.025kHz and the Mel-frequency Cepstral Coefficients (MFCC) were then captured. For each object, we issued the questions five to six times until we went through three objects (baby 1, dwarf, and girl). Then the objects were mounted again and questions asked. We repeated the above

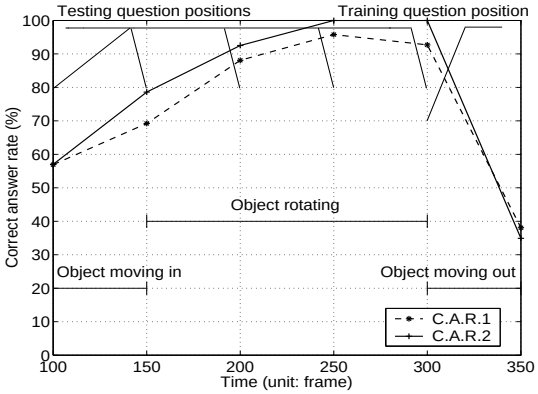


Fig. 8. The two correct answer rates of the robot v.s. the question positions in each image sequence.

Fig. 9. A sample sequence of the primed visual sensation.

process for ten times and the robot responded correctly at about 90% of the time for all three training objects.

Fig. 10 shows that the execution time of each step grew at the beginning and flattened out after about 100 seconds. The execution time of each step is lower than 18.1ms, the required interval of a single speech frame. The size of the "brain" containing three LBEs is about 60MB after the above training process.

V. Conclusions

In this paper, we introduced a developmental architecture that enables a robot to generate single and multiple modality representation in high-dimensional feature spaces. The resulting representation is an abstract nonsymbolic internal generalization by taking advantage of the spatiotemporal continuity of the real world. The effective architecture design together with the use of IHDR retrieval engine enable the robot to handle

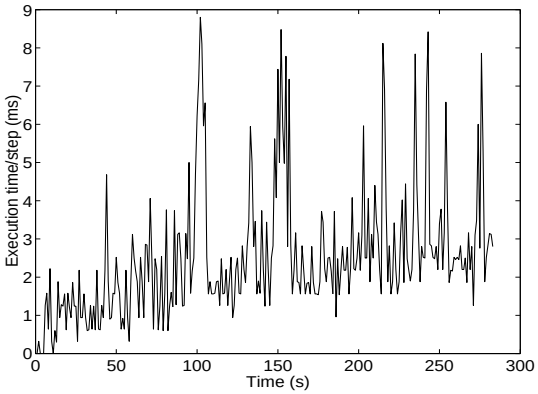
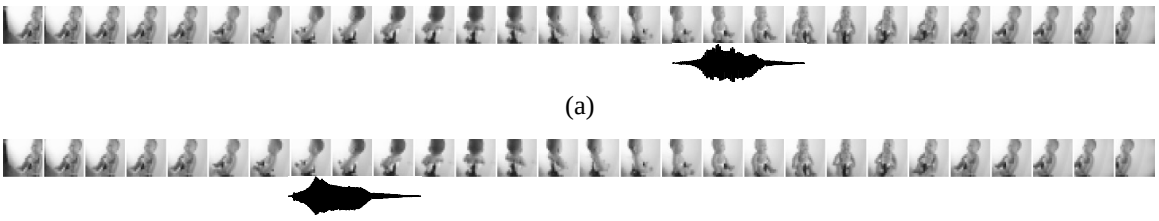


Fig. 10. The average execution time of the multimodal learning system.



(a)

(b)

Fig. 7. The misalignment of image and speech data (a) during training (b) during testing.

very high-dimensional multimodal sensory inputs online in real time. This progress is a solid step towards our ultimate goal of autonomous mental development on a robot, to learn complex cognitive and behavioral capabilities effectively with a low training cost.

In the "drawbridge" experiment, our results that are consistent with the recent human infant studies [18] showed that a developmental robot can be a useful tool for computational psychology and neurophisiological studies, although the robot and humans do not use exactly the same computations.

With the current implementation, our robot did not really have a clear object concept, but rather treated the whole image as a pattern. The future work will then incorporate our attention mechanism for voluntary segmentation, which is currently beyond the scope of this paper.

ACKNOWLEDGEMENT

This work is supported in part by the National Science Foundation under grant No. IIS 9815191, DARPA ETO under contract No. DAAN02-98-C-4025, and DARPA ITO under grant No. DABT63-99-1-0014.

APPENDIX

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

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

Thus,

$$D(f(x,y)||g(x,y)) = \int f(x,y) \log \frac{f(x,y)}{g(x,y)} dxdy.$$

$$\begin{aligned}
& D(f(x,y)||g(x,y)) - D(f(x)||g(x)) \\
= & \int f(x,y) \log \frac{f(x,y)}{g(x,y)} dxdy - \int f(x) \log \frac{f(x)}{g(x)} dx \\
= & \int f(x,y) \log f(x,y) dxdy - \int f(x) \log f(x) dx \\
+ & \int f(x,y) \log \frac{g(x)}{g(x,y)} dxdy \\
= & h(X,Y) - h(X) + \int f(x,y) \log \frac{g(x)}{g(x,y)} dxdy \\
= & h(Y|X) + \int f(x,y) \log \frac{g(x)}{g(x,y)} dxdy \qquad (3) \\
\geq & 0,
\end{aligned}$$

where $h(\cdot)$ is the differential entropy. The strict inequality holds except for the degenerated case where the second term in (3) is equal to zero, which requires that $f(x,y) \log \frac{g(x)}{g(x,y)}$ equals zero almost everywhere.

REFERENCES

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

[2] T.S. Huang, L.S. Chen, and H. Tao, "Bimodal emotion recognition by man and machine," in *Proc. ATR Workshop on Virtual Communication Environments*, Kyoto, Japan, April, 1998.

[3] M. Johnson, *The body in the mind: the bodily basis of meaning, imagination, and reason*, The University of Chicago, Chicago and London, 1974.

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

[5] N. Almassy, G.M. Edelman, and O. Sporns, "Behavioral constraints in the development of neuronal properties: a cortical model embedded in a real-word device," *Cerebral Cortex*, vol. 8, pp. 346–361, June, 1998.

[6] Y. Zhang and J. Weng, "Grounded auditory development of a developmental robot," in *Proc. INNS-IEEE International Joint Conference on Neural Networks*, Washington, DC, July 14-19, 2001, pp. 1059–1064.

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

[8] 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, January 26, 2001.

[9] W. Hwang and J. Weng, "Hierachical discriminant regression," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1277 –1293, 2000.

[10] J. Weng, W.S. Hwang, Y. Zhang, C. Yang, and R. Smith, "Developmental humanoids: humanoids that develop skills automatically," in *Proc. The First IEEE-RAS International Conference on Humanoid Robots*, Boston, MA, September 7-8, 2000.

[11] C. J. Watkins, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[12] J. Piaget, *The construction of reality in the child*, New York, Basic Books, 1954.

[13] R.Baillargeon, E. S.Spelke, and S.Wasserman, "Object permanence in five-month-old infants," *Cognition*, vol. 20, pp. 191–208, 1985.

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

[15] C. H. Cashon and L. B. Cohen, "Eight-month-old infants' perception of possible and impossible events," *Infancy*, vol. 1, no. 4, pp. 429–446, 2000.

[16] R. S. Bogartz, J. L. Shinskey, and T. H. Schilling, "Object permanence in five-and-a-half-month-old infants?," *Infancy*, vol. 1, no. 4, pp. 403–428, 2000.

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

[18] K. Durand and R. Lecuyer, "Object permanence observed in 4-month-old infants with a 2d display," *Infanct Behavior and Development*, vol. 25, pp. 269–278, 2002.

[19] J. Weng, "A theory for mentally developing robots," in *Proc. IEEE 2nd International Conference on Development and Learning (ICDL 2002)*, MIT, Cambridge, MA, June 12-15 2002, pp. 132–140.