Conjunctive Visual and Auditory Development via Real-Time Dialogue *

Yilu Zhang
Electrical & Control Integration Lab
R&D and Planning
General Motors Corporation
Warren, MI 48090
yilu.zhang@gm.com

Juyang Weng
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
weng@cse.msu.edu

Abstract

Human developmental learning is capable of dealing with the dynamic visual world, speech-based dialogue, and their complex real-time association. However, the architecture that realizes this for robotic cognitive development has not been reported in the past. This paper takes up this challenge. The proposed architecture does not require a strict coupling between visual and auditory stimuli. Two major operations contribute to the "abstraction" process: multiscale temporal priming and high-dimensional numeric abstraction through internal responses with reduced variance. As a basic principle of developmental learning, the programmer does not know the nature of the world events at the time of programming and, thus, hand-designed task-specific representation is not possible. We successfully tested the architecture on the SAIL robot under an unprecedented challenging multimodal interaction mode: use real-time speech dialogue as a teaching source for simultaneous and incremental visual learning and language acquisition, while the robot is viewing a dynamic world that contains a rotating object to which the dialogue is referring.

1. Introduction

Semantics acquisition is a process of internalizing and organizing the context. Accordingly, a communication process is about retrieving the appropriate contexts and producing the corresponding verbal behaviors.

Both of these two processes rely on two important mechanisms. The first is multimodal learning. In the process of cognitive development, children take in and integrate the information from all the senses - sight, hearing, smell, touch, and taste. 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. While the well-known supporting experiment was done on cats (Hirsch and Spinelli, 1971), similar results on human babies were also reported (Bertenthal et al., 1984). Further, the information gathered from one modality is usually limited and may be ambiguous. Here is an example how the multimodal context of the task reduces the ambiguity. When hearing a question "name?", an agent may provide a name of any object known to the agent. Only when a particular object is presented, can the agent respond correctly because the fusion of vision and audition reduces the ambiguity.

The second important mechanism is grounding (Johnson, 1974) (Harnard, 1990) (Chalmers, 1992). Grounding means that representations inside an agent should be connected to their references in the external world. For example, the representation of "dog" should be related to the presence of actual dogs in the environment. Grounding is accomplished through real-time sensory experiences. The fact that people speaking different languages can communicate with each other through physical references is a support for the importance of grounding: since people share the same physical world, we develop similar semantics and we can communicate even we do not have the same spoken language.

In this paper, we will show that, using physical properties of the real world (grounding), an embodied agent may acquire early semantics online in real-time through multimodal interactions (vision, speech and touch) where the strict-coupling assumption is not needed. The design and the implementation of this learning architecture follow the autonomous mental development (AMD) paradigm (Weng et al., 2001). One of the major features of AMD is that a self-learnable robot should not be programmed to conduct one or a few known tasks. Instead, it should possess a general task-nonspecific learning capability, and develop task-specific skills through real world sensory experiences. The work presented

*The work was supported in part by 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.

here focuses on the early stage of language acquisition related to sensorimotor learning using multimodal inputs. It does not include later language acquisition processes that require skills such as joint attention (Baldwin, 1995) (Steels and Kaplan, 1999).

2. Problem Description

We would like a robot to autonomously acquire early semantics from vision and audition, i.e., to learn appropriate behaviors given similar visual-auditory contexts. We are going to present a robot system that can answer verbal questions appropriately, given visual stimuli of a dynamically rotating object. This is an early semantics acquisition process, which requires the robot to develop visual and auditory perception, and associate visuoauditory context with appropriate behaviors, all in real time (both learning and performance sessions). In our previous work, a robot developed real-time audition learning capability (Zhang and Weng, 2001). 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. Since our goal is for a robot to handle various objects without human pre-designed representation, and, thus, learn new things "on the fly," we use the appearance-based representation rather than the monolithic object-based representation. See (Weng, 2002) for a discussion and (Weng et al., 2000) for its usage 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 the noise made by the object or the verbal name given by a teacher. However, this coupling is not strict in the following senses. (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) With the auditory sensory modality, the meaning spreads over many time frames, e.g., the utterance of an object name covers many auditory frames.

Many existing works on multimodal learning rely on the strict coupling between different modalities, such as the lip movement and the produced utterance (de Sa and Ballard, 1998) (Huang et al., 1998). Their success relies on a human-designed segmentation of training sequences and a manually assigned association between each segment and an atomic symbolic representation (e.g, label). This manual transcription is tedious and impractical, and the approaches are not suitable for a robot running continuously in an unstructured and complex environment. Moreover, pre-designed representation is not possible for an unknown task. An interesting recent study (Roy and Pentland, 2002) proposed a minimum-mutual-information-based framework to resolve the imperfect alignment problem. However, in the reported experiment, although the auditory data was very challenging mother-baby verbal interactions, the visual stimuli were static images of objects. In developmental learning, the developmental algorithm does not explicitly specify segmentation. Instead, segmentation is an internal behavior of the agent's when the agent becomes cognitively mature. In our research, we view the imperfect alignment problem as an abstraction issue and use a well-designed architecture to realize this fundamental mechanism.

The imperfect alignment problem is rooted in the fact that an object appears to the robot as a sequence of images captured continuously in time from different viewpoints. Unless the robot "knows" the sequence of images correspond to a single object, it will not establish a robust correlation between the visual stimuli and the auditory stimuli. Therefore, we need a representation upon which the robot can group these different images into a single cluster, i.e. an object, before any object-specific action can be learned.

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. Moving in and out the agent's field of view, two consecutive views of an objects are similar when the capturing speed is high enough. If we filter out the high-frequency components, the images change even more slowly and may be considered as identical in some cases, which is exactly what we need. This underlying mechanism is closely related to the object permanence (Baillargeon, 1987) (Baker et al., 2001).

3. Architecture and Algorithm

The proposed architecture is built upon a regression engine, called hierarchical discriminant regression (HDR) (Hwang and Weng, 2000). Because of space limit, we can not go into the details of the method. Very briefly, the HDR technique automatically derives discriminating feature subspaces in a coarse-to-fine manner from the input space to generate a tree architecture of self-organization memory. It can handle high-dimensional (thousands of dimensions) data in real-time, which is crucial to a robot. Here, we use incremental HDR (IHDR), which is an incremental algorithm of constructing an HDR tree (Weng et al., 2000).

3.1 Level-building Element

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

We call the input to an IHDR tree as the last context, $s(t)$, while the output as the primed context, $p(t)$. A primed context p, consists of three parts, a primed sen-

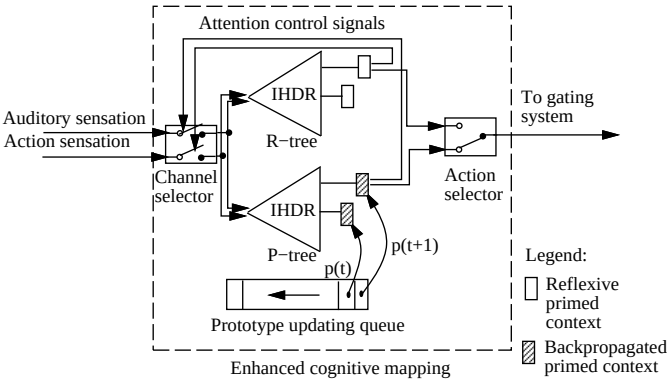

Figure 1: A level-building element.

sation vector p_s, a primed action vector p_a, and a value associated with the primed action vector. An IHDR tree approximates a mapping g so that,

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

The two IHDR trees in an LBE are identical except the bottom one is associated with a *prototype updating queue* (PUQ). We call the upper one the reality tree or *R-tree* and the bottom one the priming tree or *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 (Watkins, 1992):

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

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

Reorganizing Eq. (1), we have:

$$p^{(n)}(t) = \frac{n-1-l}{n} p^{(n-1)}(t) + \frac{1+l}{n}\gamma p^{(n-1)}(t+1), \quad (2)$$

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 Eq. (1), is recursively backpropagated into earlier primed contexts. Therefore, Eq. (1) is effectively a prediction model. When an earlier context is recalled, it contains the expected future information. This property of Eq. (1) has been used in our previous work (Zhang and Weng, 2002) to enable a robot to learn complex behaviors upon acquiring simple ones.

Another interesting property of Eq. (1) is that it is actually a low-pass filter. With the recursive averaging over the consecutive primed contexts, the primed sensation part of the primed contexts changes slowly compared to the corresponding last sensations. We have discussed that, in a real world, an object does not appear or disappear as a magic. In other words, the presence of an object itself is a low-frequency component, while the orientation changes introduce some high-frequency components. Therefore, we may use the low-pass-filtering property of the model to filter out the high-frequency components in the visual sensation giving a low level abstraction. The resulted slowly-changing primed visual context is not sensitive to orientation changes and enables a robot to tolerate the imperfect alignment.

The LBE module was designed in order to fulfill a general learning purpose. In multimodal learning presented here, some components in the LBE module are not used as you will see in the algorithm below. Because of page limit, we do not discuss the LBE components that are not used in the multimodal learning architecture, such as the attention control signals, the channel selector, and the action selector. If interested, the reader is referred to (Zhang and Weng, 2002) for detailed discussions.

4. Multimodal Learning

4.1 Learning Mechanism

Fig. 2 (a) shows the architecture we used to do early semantics learning. It has three LBE modules, a vision LBE (V-LBE), an audition LBE (A-LBE), and a high-level LBE (H-LBE). The underlying idea of such an architecture is that while A-LBE and V-LBE may work individually to do certain semantics learning, their combination in H-LBE can resolve the ambiguous situations when neither of the two modalities, vision and audition, contains enough information for decision-making.

In our system, the visual sensation is the original image captured by a CCD camera. The program does not have pre-designed code for detecting low-level features such as edge histogram. Instead, important discriminant features are derived automatically by the IHDR trees from experiences that include the labeling information (imposed actions) provided by the trainer. As will be shown in the experiments, only very sparse labels are needed (about 2% in time). The auditory sensation is captured by a sound blaster card through a microphone. We do Cepstral analysis on the original sound signals before the data enters A-LBE. Since sound is a linear signal in the sense the information is distributed over a period of time, each auditory sensation vector actually covers twenty 18.1ms speech frames as shown in the experimental results. The primed sensations from the P-trees of both V-LBE and A-LBE are inputs to H-LBE. After the low-pass filtering in PUQ, the primed sensation only keeps the low-frequency components of the last context.

A high-level outline of the algorithm is as follows. As one may notice, in this algorithm, the training (featured by words such as “update”) and testing (featured

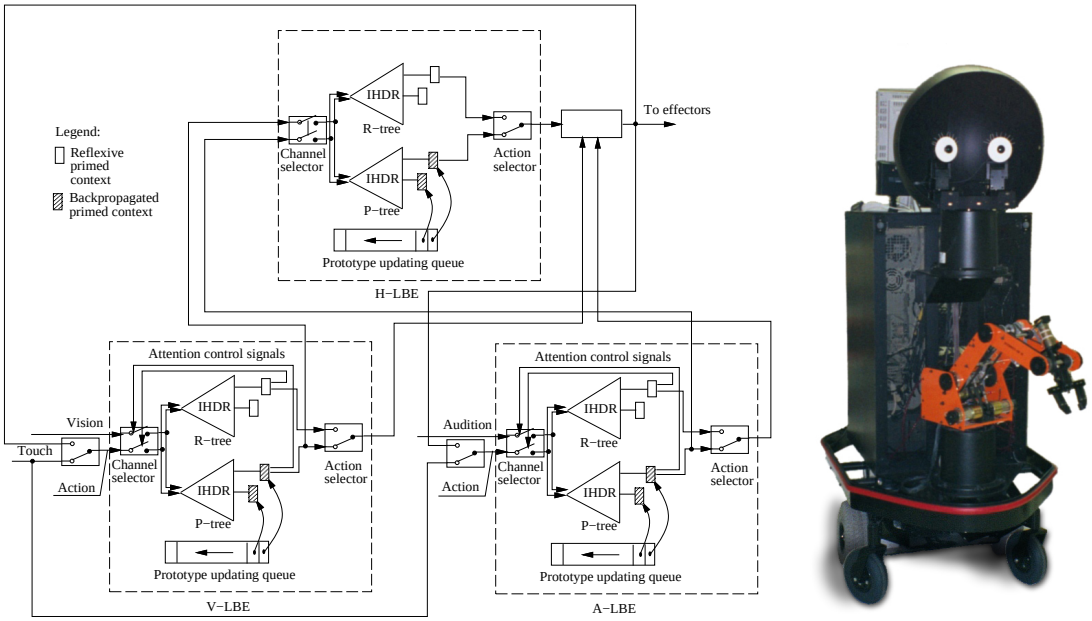

Figure 2: (a) The multimodal learning system architecture; (b) The SAIL robot at Michigan State University.

by words such as "retrieval") processes are embedded to each other in order to make online learning possible.

1. Collect the sensation from the auditory sensor, $x_s(t)$, the visual sensor, $x_v(t)$, and the action sensor, $x_a(t)$ (If an action is imposed through the touch sensors, $x_a(t)$ is the imposed action. Otherwise, it is the action produced by the system itself in the last computation loop.).

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 Eq. (1).

5. Take the primed sensation part of $p_v(t)$ and $p_s(t)$ as the input to H-LBE.

6. Update the R-tree of H-LBE using 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 primed action part of $p_h(t)$ is sent to effectors for execution.

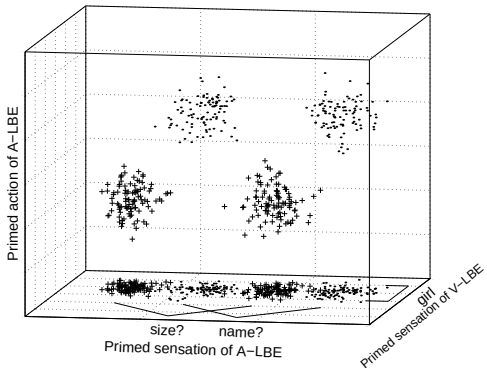

Figure 3: The illustrative comparison of using and not-using the primed action in decision making.

4.2 Multimodal Abstraction

The underlying reason that 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 was presented. This is an abstraction process, where the cognitive activities reduce the variance of the sensation, keep the invariant components, and eventually generate the acquired knowledge (e.g., the correct answers) for other equivalent contexts.

The primed actions from the lower-level LBEs was not used in the above algorithm. In fact, the primed actions contain very useful information. For example, the utterances of a same word vary from person to person and each word is typically composed of several modes. The primed sensations of A-LBE consists of multiple modes

too as shown in Fig. 3. Consequently, the decision boundary is complicated in the space spanned by the primed sensations from both V-LBE and A-LBE, which gives the recognizer a hard time. It is easy to imagine that while the utterances of "name?" and "size" vary from people to people, the internal responses of A-LBE do not. In other words, because A-LBE does discriminate the verbal question "size?" from "name?" the behavior (the output) of A-LBE completes another abstraction process, i.e., reducing the variance of auditory sensation (different persons' utterances) by mapping it to A-LBE's internal responses. By adding the axis of the primed action from A-LBE, the clusters are well-separated as illustrated in the 3D space in Fig. 3. Following this thought, we improve the multimodal learning architecture above by feeding a primed action pattern from the two lower-level LBEs to H-LBE. The primed action pattern of A-LBE is given by,

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

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,

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

The reason that the above additional information helps to improve performance can be explained in terms of information theory. Let PS and PA represent the primed sensation and the primed action, respectively. Both PS and PA are random variables. Let $f(PS)$ and $g(PS)$ be the pdf for "name" and "size", respectively. Let $f(PS, PA)$ and $g(PS, PA)$ be the joint pdf for "name" and "size", respectively. 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 two pdf. 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.

5. Experimental Results

We implemented the multimodal learning architecture on our house-made human-size mobile robot (Fig. 2 (b)). The robot 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. Our robot has four pressure sensors on its torso and 28 touch sensors on its eyes, arm, neck, and bumper. Its onboard main computer is an Xeon 2.2GHz dual-processor workstation with 1GB RAM. All the sensory information

Figure 4: The objects used in the experiment.

Figure 5: Part of a sample image sequence.

processing, memory recall and update, as well as effector controls are done in real-time on board.

The experiment was done in the following way. After our robot started running, the trainer mounted objects one after another on 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 trainer verbally asked the questions, "name?" and "size?" And then the trainer gave the appropriate answers by pushing the switch sensors of the robot. Different switch sensor status represented different answers. Particularly, one of two sizes, large or small, was assigned to each object. Since the objects were rotated and moved in and out of the robot's field of view continuously, the orientation and the positions of the objects kept changing. There were hardly chances that the robot could see the same images of the objects when the same question was asked again. A sample video sequence seen by the robot is shown in Fig. 5. Totally 12 objects were presented (Fig. 4) to the robot. All these real-world objects were of very complex shape and of non-rigid form (e.g., Harry Potter's hair). It was extremely difficult, if not impossible, to model them using 3-D representations. We expected the robot to correctly answer the taught questions when the objects were presented and the question were asked the next time.

The images were captured by a Matrox Meteor II board as gray-scale images at 30 frames per second. The dimension of the images was 25-by-20. The speech data were digitized at 11.025kHz by a normal sound blaster card. Cepstral analysis was performed on the speech stream 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 the robot at the rate of about 50Hz. Twenty consecutive MFCC vectors together form a single auditory sensation vector.

To examine the behavior of the robot in detail and evaluate the performance, we pursued an experiment on prerecorded data first. The image data of each object were five video sequences of the object moving in the robot's field of view, rotating for roughly one round, and then moving out of the robot's field of view. Each image sequence contained 350 frames. Frame 1-50: background images. Frame 51-100: an object moving to the center of the robot's field of view. Frame 101-300: the object rotating along its center axis. 301-350: the object moving out of the robot's field of view.

The auditory data was taken from the number data set contributed by 10 people, each making five utterances for each of the ten numbers, one to ten. We used the utterances of "one" to represent "name" and "two" to represent "size." During training, the switch sensor inputs (a numerical vector) were given after the utterances were finished, which was the time the robot was taught the answers. Of all the five sets of image and speech data, we used four of them in training and the left-out one for testing. So, with 12 objects, ten persons, and two questions, the robot was taught 960 times in training and evaluated for 240 times in testing.

To emulate the situation that the trainer would not be able to ask questions consistently to synchronize the object views, we randomly choose the point to align the image sequences and speech sequences (Fig. 6). Specifically, the end point of questions was aligned with image No. 300 during training. When testing, it was aligned with image No. 100, 150, 200, 250, and 300, respectively.

The behavior of the robot was evaluated in the following way. We counted the number of robot's responses after each question utterance, which is usually larger than one. If the majority of the responses were correct, we counted that the robot did correctly in this image sequence. Otherwise, we counted it as wrong. Here came the correct answer rate (C.A.R.),

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

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 the primed sensation as C.A.R.1 and that using both the primed sensation and the primed action as C.A.R.2. The correct answer rates are shown in Fig. 7.

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. In the real world, the visuoauditory scenes during testing were never exactly same as those during training when the questions

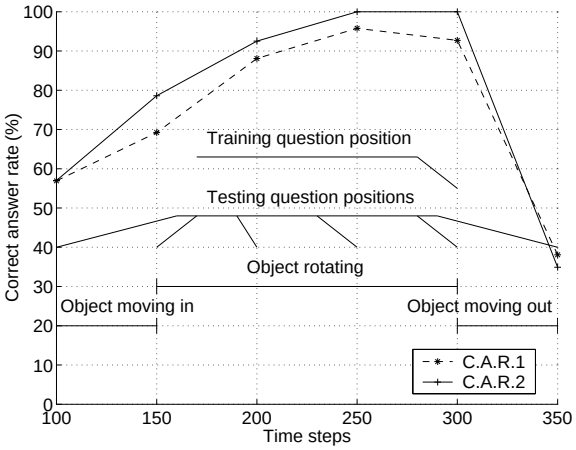

Figure 7: The two correct answer rates of the robot v.s. the question positions in each image sequence.

were asked. 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 question-position difference, the correct answer rate dropped gradually. Also, during the time the objects moving in or out of the robot's field of view, the robot's performance was low because the current robot does not have an attention mechanism to locate the object off the center of its field of view.

To see why the robot was able to respond when the questions in testing were not asked at the exactly same time as in training, we show the primed sensation of V-LBE in Fig. 8. Since the operation done in the PUQ of V-LBE was a low-pass filtering, the primed visual sensation was a blurred version of the real visual sensation. The result was that the visual inputs to H-LBE did not change a lot in consecutive frames when the same object was presented. Thus, the robot was able to answer the question correctly even it was taught while another pose of the object was seen.

Overall, the improvement from not using primed action to using is visible in Fig. 7. As we have explained in Section 4.2, the primed action pattern catches the characteristic of the inputs to A-LBE, i.e., the questions, although the best primed action of A-LBE is not likely to be right. The primed action pattern contains less variance than that of the primed sensation. Therefore, the primed action pattern provides another level of abstraction.

In the real-time experiment, the verbal questions ("name?" and "size?") were asked followed by the answers imposed through the switch sensors of the robot. For each object, we usually issued each question five to six times. To make it easy for the trainer to see the response of the robot, we manually mapped the robot's action vectors to the names of the objects and used Microsoft text-to-speech software to read out the names. After going through three objects (baby 1, dwarf, and girl), the objects were mounted on the gripper in turn again and the questions were asked without giving answers. We

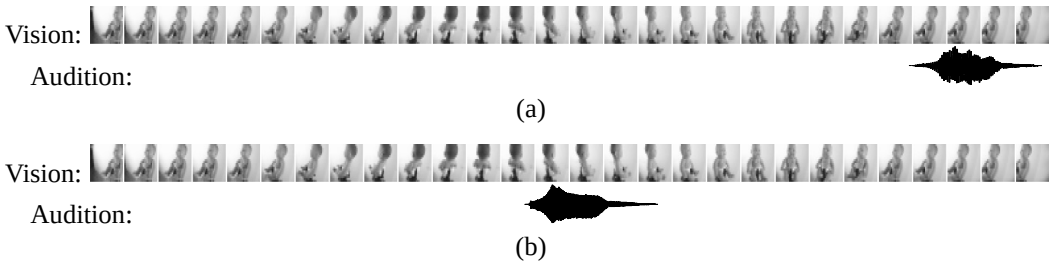

Figure 6: The alignment of image and speech data (a) during training (b) during testing.

Figure 8: A sample sequence of the primed visual sensation.

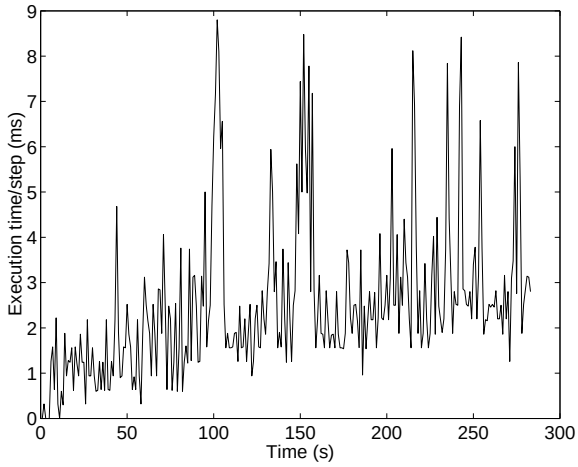

Figure 9: The average execution time of the multimodal learning system at each time step is much shorter than 18.1ms, the required interval of each speech frame.

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

Fig. 9 shows that the execution time of each step grew at the beginning and it flattened out after about 100 seconds. The short surging period around 100s, 150s and 210s were the times when 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 is lower than 18.1ms, the required interval of a single speech frame. The size of the whole "brain" containing three LBEs is about 60MB after the above training process.

6. Conclusions

In this paper, we introduced a learning architecture that enables a robot to learn visual and auditory perception and association from multiple modalities. Unlike most of the existing works, this architecture does not require the strict coupling between visual stimuli and auditory stimuli. With this architecture, a robot was able to pursue real-time semantics learning for early cognitive development. After online multimodal interactive dialogue training, the robot was able to answer the vision-related questions correctly even when the orientation of the objects was changed. This process emulates the way a human child learns concepts of the physical world through verbal instructions.

The proposed learning architecture takes the advantage of the spatiotemporal continuity of the real world. A more "abstract" numeric representation is realized through sensory trajectory filtering (priming) and the use of internal primed action distribution. While the sensory inputs vary greatly, the internal responses to the inputs vary less, providing an abstract nonsymbolic representation. The effective architecture design together with the use of HDR retrieval engine enable the robot to handle very high-dimensional multimodal sensory inputs online in real time. This progress is a solid step towards our ultimate goal of autonomous mental development by a robot, to learn complex cognitive and behavioral capabilities effectively with a low training cost.

With the current implementation, the robot did not discriminate the foreground from the visual background. In other words, the robot did not really have a clear object concept. It essentially treated the whole image as a pattern, with which the audition signals and behaviors were associated. To achieve object concept learning, among other requirements, the system needs a sophisticated attention mechanism to establish the bound of the objects from the background. This voluntary segmentation is beyond the scope of this paper. Another limitation of this implementation is that the action of the system was designed to be the output of one of the three LBEs, namely H-LBE. When the robot becomes cognitively more mature, it should be able to choose the LBE, from which the action is taken. This will be our future work.

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,

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

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

where $h(\cdot)$ is the differential entropy.

References

Baillargeon, R. (1987). Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology*, 23:655–664.

Baker, C. I., Keysers, C., Jellema, J., Wicker, B., and Perrett, D. (2001). Neuronal representation of disappearing and hidden objects in temporal cortex of macaque. *Experimental Brain Research*, 140:375–381.

Baldwin, D. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P., (Eds.), *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associatives, Publishers, Hillsdale, NJ.

Bertenthal, B., Campos, J., and Barrett, K. (1984). Self-produced locomotions: an organizer of emotional, cognitive, and social development in infancy. In Emde, R. and Harmon, R., (Eds.), *Continuities and Discontinities in Development*. Plenum Press, New York, NY.

Chalmers, D. (1992). Subsymbolic computation and the chinese room. In Dinsmore, J., (Ed.), *The Symbolic and Connectionist Paradigms: Closing the Gap*, pages 25–48. Lawrence Erlbaum Associates, Hillsdale, NJ.

de Sa, V. and Ballard, D. (1998). Category learning through multimodality sensing. *Neural Communication*, 10:1097–1117.

Harnard, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hirsch, H. and Spinelli, D. (1971). Modification of the distribution of receptive field orientation in cats by selective visual exposure during development. *Experimental brain research*, 13:509–527.

Huang, T., Chen, L., and Tao, H. (1998). Bimodal emotion recognition by man and machine. In *Proc. ATR Workshop on Virtual Communication Environments*, Kyoto, Japan.

Hwang, W. and Weng, J. (2000). Hierachical discriminant regression. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(11):1277 –1293.

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

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

Steels, L. and Kaplan, F. (1999). Situtated grounded word semantics. In *Proc. Int'l Joint Conf. on Artificial Intelligence*, San Francisco, CA.

Watkins, C. J. (1992). Q-learning. *Machine Learning*, 8:279–292.

Weng, J. (2002). A theory for mentally developing robots. In *Proc. IEEE 2nd International Conference on Development and Learning (ICDL 2002)*, pages 132–140, MIT, Cambridge, MA.

Weng, J., Hwang, W., Zhang, Y., Yang, C., and Smith, R. (2000). Developmental humanoids: humanoids that develop skills automatically. In *Proc. The First IEEE-RAS International Conference on Humanoid Robots*, Boston, MA.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., and Thelen, E. (2001). Autonomous mental development by robots and animals. *Science*, 291:599–600.

Zhang, Y. and Weng, J. (2001). Grounded auditory development of a developmental robot. In *Proc. INNS-IEEE International Joint Conference on Neural Networks*, pages 1059–1064, Washington, DC.

Zhang, Y. and Weng, J. (2002). Action chaining by a developmental robot with a value system. In *Proc. IEEE 2nd International Conference on Development and Learning (ICDL 2002)*, pages 53–60, MIT, Cambridge, MA.