Perception–mediated Learning and Reasoning in the CHILDLIKE System*

Ganesh Mani and **Leonard Uhr**
Computer Sciences Department
1210 W. Dayton St., Madison, WI 53706
{ganesh,uhr}@cs.wisc.edu

Abstract

Intelligent agents interacting with their environments combine information from several sense modalities and indulge in tasks that have components of perception, reasoning, learning and planning. Traditional AI systems focus on a single component. This paper highlights the importance of the integrated perceive-reason-act-learn loop, and describes a system designed to capture this loop. As a first step, it learns about simple objects, their qualities, and the words that name and describe them. The visual-linguistic associations formed serve as a bias in acquiring further knowledge about actions, which in turn aids the system in satisfying its internal needs (e.g., hunger, thirst, sleep, curiosity). Learning mechanisms that extract, aggregate, generate, de-generate and generalize build a hierarchical network (that serves as internal models of the environment) with which the system perceives and reasons.

Introduction

Intelligent agents (embedded in an environment) routinely sense information through one or more sensory channels, engage in various kinds of perceptual reasoning in the process of recognizing a stimulus or a pattern of stimuli, and respond, acting appropriately. Agents also exhibit the capability to learn from repeated interactions with external stimuli. This continuing perceive-reason-act-learn-perceive-... loop is almost certainly central to intelligence, as evidenced by the behavior of animals, including humans. Systems that fully capture this loop will exhibit the essence of mundane everyday reasoning.

The acquisition of knowledge about the environment starting with visual–linguistic learning has begun to gain attention among researchers (e.g., see Nenov & Dyer, 1988; Weber & Stolcke, 1990). Feldman et al. (1990) have dubbed this a touchstone task for cognitive science.

This paper discusses a system that we have designed and are implementing for capturing the perceive-reason-act-learn loop. First, the system learns to recognize and name simple objects, and their parts and qualities. It can then learn, from strings of these perceptually grounded words, about classes, relations, and actions.

Section 2 describes the system's architecture and goals (also see Mani & Uhr, 1991). Section 3 examines the learning mechanisms employed and Section 4 gives details of the acquisition of internal models of the environment.

The Architecture of CHILDLIKE

The CHILDLIKE[1] system is a computational information-processing model (implemented in Common Lisp) designed to learn about objects, their qualities, and the words that name and describe them; and, further, to use this knowledge to satisfy its internal needs (e.g., hunger, thirst, sleep, curiosity).

The system is input sequences of simple "experiences" from which it attempts to learn. An experience contains several different components— for example, a visual pictorial scene, a short language utterance, an abstracted action.

The visual input is a snapshot at a single moment of time. It consists of a 4-by-4 or 8-by-8 image — typically one that a low- or intermediate-level computer vision system might output — that encodes information such as edges, or colors and textures, or simple shapes. These are pre-processed as needed, using a network of convolution-like mask-matching-plus-thresholding operations, to obtain primitive features such as long vertical or horizontal lines, the texture in a large segment of the image, or the color of these significant regions.

The language input consists of short (typically 2-5 words in the current implementation) English language strings (although any language could be used without changing any aspect of the system). An example sequence of inputs to the system is shown in Figure 1.

Visual Input:	**Language Input:**
a) [Picture of an apple]	red apple
b) [Picture of a banana]	banana
c) [Picture of an apple and a banana side by side]	apple and banana
d) [Picture of a table]	brown table
e) [Picture of an apple on a table]	apple on table

Figure 1: A simple input sequence

*This research was supported in part by a Grant-in-Aid of research from Sigma Xi.

[1] which stands for Conceptual Hierarchies In Language Development and Learning In a Kiddie Environment.

Each visual input to the system is a snapshot of the environment in time. Using such simple ordered sequences of inputs, the system is taught how to associate names of objects with their shape, size and other features, and further to associate relational words with visual features that imply them. Once the system is able to ground language symbols in terms of information perceived through the visual channel, it can be further trained by verbal input alone. Thus the system bootstraps itself, by first grounding words in perceptual information, to the point where it can learn from these grounded words only.

CHILDLIKE's parallel-hierarchical structure was chosen because it is general; successive compounding can produce any possible set of functions, starting from a universal set of primitives. It also has the potential of great speed and efficiency. And it is a widely used structure for perceptual recognition of objects in 2-dimensional images (Hanson & Riseman, 1978; Li & Uhr, 1987; Tanimoto & Klinger, 1980; Uhr, 1987), parsing trees for 1-dimensional language strings (Chomsky, 1986; Osherson & Lasnik, 1990), and the hierarchical building up of logical functions that accept the combinations of 0-dimensional terms handled by concept formation and similarity-based learning systems (Hunt, 1962; Michalski, 1983; Mitchell, 1982; Quinlan, 1986). It appears that the multilayer, converging "recognition cone" structure of micro-modular processes being developed for perceptual recognition tasks (Uhr, 1978; 1987) can also be used for building linguistic structures and visual-linguistic associations. Building on these visual-linguistic associations, the system can further acquire memory structures that encode the effects of actions and reason about need-fulfillment.

Learning Mechanisms in CHILDLIKE

CHILDLIKE's learning mechanisms extract and incorporate information learned via interactions with the environment. These learning mechanisms can be grouped into to the following major types: *extraction*, *aggregation*, *generation*, *de-generation* and *generalization*. Extraction is involved with the process of carving out potentially useful pieces of information from the visual and verbal input fields. Aggregation puts together pieces. Generation creates new links that encode, contain, and apply this new information.

Extraction takes place by imposing windows on the input array(s) containing perceptual information. This embodies a local-receptive field heuristic that favors extracting connected, compact features from which nodes that detect these features are generated. Evidence that brains' neurons predominantly interact with near neighbors, and empirical evidence that such local receptive fields are superior to random receptive fields for visual tasks in the recognition cone framework (Honavar and Uhr, 1989) support the choice of this heuristic. This locality heuristic is used to aggregate compound features at subsequent levels also.

The network of nodes created as a result of extraction, aggregation and generation serves as an internal representation of the environment. Each node encodes some feature or microfeature (e.g., one leg of a chair, the word "leg"), or some compound feature or class (e.g., four legs, the phrase "four-legged chair," furniture).

De-generation and generalization mechanisms simplify and speed up processing, and combat potential combinatorial explosions.

De-generation involves the discarding of nodes and links that appear to encode useless or wrong information. The value of a link is assessed (e.g., by its associated weight, or processes that estimate how useful it has been), and the network pruned accordingly.

Generalization may involve removing certain links (this corresponds to the *dropping condition rule* of Michalski, 1983, in symbolic similarity-based learning), replacing links to a number of nodes representing explicit entities with a link to a single node which may stand for any of these entities (*turning constants into variables rule*), replacing links to nodes n_k at level i with a link to a node n at level $i+1$ such that a majority of n_k are linked to node n (*climbing generalization tree rule*), or replacing a sub-network with a node that stands for the sub-network (*constructive generalization rule*).

The order of the training sequences also plays a significant role in knowledge acquisition. At each step, experiences are judiciously chosen to slowly build on prior learning. For example, training sequences aimed at teaching relations such as "on," "above" and "and" should contain visual information portraying these relations between already-learned objects, along with the associated language string. An example of such graded training can be found in Figure 1. It should be noted that a number of training sequences of the sort shown in Figure 1 may be required before the relations are effectively learned. Training sequences may also force the learner to focus on a particular aspect of the input by varying all other aspects of the input (see below).

Acquiring Internal Models of the Environment

Reasoning about the world is greatly facilitated by having an internal model of that world. The CHILDLIKE system's internal model consists of subnetworks encoding associations among visual features, words, actions, needs and compound features derived from them.

Learning Words About Objects

CHILDLIKE learns about simple objects by "seeing" them through the visual channel and simultaneously "hearing" a linguistic description of the object through the language channel. Often, the object is not named in isolation but is named along with words describing other properties (e.g.,"green apple") of the object or the scene. Through repeated extractions from the sensory channels and generation of associations, CHILDLIKE learns the word that corresponds to a particular object in the visual field. It conjectures that other words (which do not appear to correspond to whole objects) may refer to parts

BEHAVING PHASE :

1. Input one of the arrays (visual or linguistic).
2. Extract primitive features, aggregate them into compound features, find the best-match for the entities in this channel and find corresponding entities in the other channel.
   {More formally, the extractions and aggregations can be expressed as
   (P stands for a primitive feature and C for a compound feature):
   Level 0 $C_j = A_j(P_i, S)$
   Level 1 : $C_k = A_k(C_i, S)$ and so on for subsequent levels.
   Here, i ranges over a local, connected window, S is an associated strength or weight vector (which gets modified based on errors made), and A_j, A_k, etc., are aggregation functions. Matching is performed by aggregating evidence hierarchically and collecting entities implied at each level.}

LEARNING PHASE:

1. Input visual and linguistic arrays (V and L respectively).
2a. Extract primitive features and aggregate them into compound features hierarchically.
   b. Match known entities across the visual and linguistic channels (call the matched parts v and l respectively).
   c. Generate all possible links, subject to a resource constraint, between novel features (corresponding to $(V - v)$ and $(L - l)$) across the visual and linguistic channels, plus q (a changeable parameter) links involving already known entities (corresponding to v and l).
3. If normal-training
   Change link weights using a Hebbian-like learning mechanism.
   {More precisely, the learning mechanism can be expressed as $W_{new}(N1, N2) = W_{old} + \eta f(N1, N2)$
   where $N1$ is a (primitive or compound) visual feature and $N2$ is a linguistic feature. $f(N1, N2)$ is non-zero only if the features $N1$ and $N2$ are both present in the current input instance; currently a normalized frequency count is used.}
   else (attention-focusing training)
   Modify weights explicitly using a high learning rate.
4. Adjust link weights, delete nodes and links, and form compact structures or subnets using de-generation and generalization.
   {De-generation heuristics and generalization mechanisms are described in the text.}

Figure 2: The algorithm for learning about objects, object qualities and words that refer to them

of objects, their qualities, or even relationships among objects.

Learning Words About Parts and Qualities of Objects

An explicit feature in the visual field may correspond to certain qualities of the object (such as color or texture) or may constitute a sub-part of the object. Words about these features often occur in the language field of the input. Associations between words and their corresponding features are generated by the initial experience, and their weights strengthened (or weakened) by subsequent ones. Qualities and parts of objects will often also tend to be highly correlated with (hence will be linked to) the names of the object; for example, the color "yellow" from the visual field — in addition to being linked to the word "yellow" with a strong weight — may also be linked to the word "banana" and to the words for other yellow objects with moderate weights. But to the extent that the system is input experiences where an object's name is always present, and the names of qualities and parts are present for other objects as well, it will learn the information needed to assign names correctly.

The association between something like the color "yellow" and the word "yellow" can be further strengthened by a process called *attention-focusing* training. This involves making the system concentrate on a particular sub-concept via a training sequence or set of examples that focuses on the particular sub-concept. For example, to force the learner to form a reliable association between the color "yellow" and the word "yellow," one of the following strategies can be used: A) a yellow object is input through the visual channel and explicit feedback (which is specially marked, or input through a specific explicit-feedback channel) refers to the word "yellow" on the language channel. Such specially marked training instances are processed in a special way, resulting in the required link's weight being strengthened rapidly (using a higher learning rate). Note that although ad hoc, this is simply a fast alternative to repeating a large number of training examples such as those described next. B) a number of yellow objects, with other attributes (such as shape and texture) entirely unrelated to those of the objects experienced so far, are input through the visual channel along with the word "yellow" on the language channel. This helps in strengthening the required association without adversely affecting other useful, previously learned associations. Any new associations that are formed (embedding the irrelevant features) get de-generated eventually because they are not reinforced, and hence fall below a dynamically-adjusted threshold.[2]

[2]Threshold adjustments are done when it is evident that the system has made an error (e.g., the visual node for a ba-

Learning Spatial Relations

Relations among object parts or between objects are much harder to learn if the notion of a relation (and its arity) is not built into the system. In CHILDLIKE (which is designed to learn as much as possible, rather than use built-in knowledge), priority is given to associating words from the language channel with explicit features or objects in the visual channel. Where a word and a sub-object have already been learned and linked together, they are considered to be accounted for. Each of the additional words is tentatively linked to each of the additional sub-objects. The arity of the relation is tentatively assumed to be the number of structures in the visual image already linked to words that are present.

When an example such as

> Visual input:
> Picture of yellow banana and brown table, with the center of the banana having a y-coordinate 2 units (a unit is a certain number of rows depending on the resolution of the input image) higher than that of the table (the x-coordinates of the 2 objects are the same or differ by less than 0.1 unit, a parameter which could be a function of things such as object-size).
>
> Language input:
> banana on table

is presented for learning the relation "on," the system finds the best match (the highest level structure implied across the visual and language input) for each component, and hypothesizes "on" to be the quality of one object and also to be a relation between the two constituent structures in the visual image that had matching words. Subsequent training using examples such as

> Visual input:
> Picture of red apple and white table, with the center of the apple having a y-coordinate 1 unit higher than that of the table (same x-coordinates).
>
> Language input:
> apple on table

reinforces the latter hypothesis.

Finally, rules like the following one are acquired as a result of the training regime explained above (the rule is a hierarchical subnetwork within the system; we have re-expressed the rule here in a more symbolic format for clarity).

> on(L,M) $\longleftrightarrow$
> L(y-pos = Y + (0..2), x-pos = X $\pm$ delta, color = *)
> M(y-pos = Y, x-pos = X, color = *)

When necessary, spatial relations are also taught the system in a more explicit way, using the attention-focusing training mechanism described above. Several experiences are input that contain the same objects, but

nana being active along with the node for the word "apple").

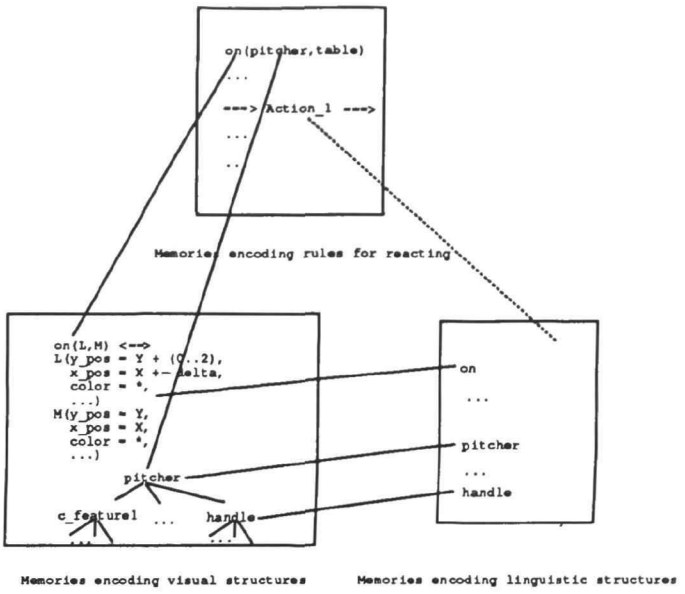

Figure 3: An overview of the memory structures acquired by CHILDLIKE

at different locations. Since other properties (such as color and texture) do not change when an object moves, the system learns to associate the (relative) positional information with relational words.

This will handle scenes whose descriptions are, for example, "glass on three-legged stool," as well as scenes involving non-spatial relations, for example ones described by "apple and yellow banana."

Learning About Actions and Need-fulfillment

Reasoning about actions and their effects is an important aspect of intelligent behavior. In CHILDLIKE, sequences of visual frames are used to learn about the effect of actions. From two visual frames, and an intervening action, the system learns to associate an action (say $A1$) with the description of the two frames (based on a *temporal proximity* heuristic). Needs are sensed internally (in the same way that visual and linguistic features are sensed externally) and action sequences that change need levels favorably (indicative of need-fulfillment) are learned and stored for future use.

Discussion and Conclusions

We have successfully experimented with the current version of CHILDLIKE using 5–10 object classes (such as fruit, food and furniture) with 5–10 distinct objects in each class (a large number of instances of each object are possible since objects can vary in position and also in their features and feature values). We are in the process of testing the learning abilities of the system on larger sets of objects. Our preliminary explorations appear to indicate that the underlying approach is sound; the system should not exhibit brittleness as the number of objects and their properties grows.

A snapshot of the memories acquired by CHILDLIKE

is shown in Figure 3. Distinct memories are used to encode the actions-related rules and their components; these are linked to memories containing encodings of related visual structures and words. Thus a pre-condition like "on(pitcher,table)" can occur in the action memories as part of a rule, and is connected to the corresponding visual structures and through them to words. (Only the highly weighted linkages are shown in Figure 3; links with small weights exist, for example, between the visual structure for "pitcher" and the word "table" since they co-occur in the same training instance. Links whose weights fall below a dynamically adjustable threshold are periodically removed from the memory structures.) The dotted line in Figure 3 represents an example of the kind of links that would be formed after words about actions are also learned.

An important feature of the system is its ability to learn and reason using language input alone (such as "apple is fruit," "banana is fruit," and "fruit is sweet"). Once a few words and what they refer to have been learned, new concepts can be learned using these words alone. This is salient in human learning, and can greatly speed up learning.

It should be noted that the performance of the system is not tied to specific object classes, specific words about them, or specific needs. The same program can be used with different objects, object-classes, needs, or/and words from a different language.

In CHILDLIKE, the effects of actions are learned from experience. The representation in CHILDLIKE's action memories is akin to that used by planning systems (starting with STRIPS, Fikes & Nilsson, 1971). This acquired knowledge, which is already in a convenient representation, can be used by a powerful planning module to enhance CHILDLIKE's reasoning capabilities.

We have described a system that attempts to capture the perceive-reason-act-learn loop which is central to intelligence, as evidenced by animal and human behavior. A perception–mediated approach enables efficient acquisition of concepts that are descriptive as opposed to those that simply classify and discriminate. The multi-level description of objects, including parts and qualities and words about them, facilitates different kinds of reasoning. Words and small sentences about the visual scene are grounded in terms of visual entities, and vice versa. The system can also learn to reason about simple action sequences that can potentially satisfy needs. All of this learning and reasoning is performed using micro-modular structures and general-purpose mechanisms such as extraction, aggregation, generation, de-generation and generalization.

We are extending CHILDLIKE in several ways. These include using a more sophisticated inheritance paradigm, refining the way actions are handled, and the inclusion of a deliberation module to arrive at approximate plans for fulfilling needs. Lastly, we note that the current functionality of the system is very good considering how little information is built in to the system a priori; all the rules employed in recognizing objects and their parts, in grounding words in objects and relations, and in reasoning are learned from experiences, starting with the most simple ones.

References

1. Chomsky, N. 1986. *Knowledge of language: Its nature, origin and use.* New York: Praeger.
2. Feldman, J.A., Lakoff, G., Stolcke, A. and Weber, S.H. 1990. Miniature Language Acquisition: A touchstone for cognitive science. *Proc. of the 12th Annual Conference of the Cognitive Science Society.*
3. Fikes, R. E. and Nilsson, N. J. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence 2 (3/4).*
4. Hanson, A.R. and Riseman, E.M. 1978. Visions: A Computer System for Interpreting Scenes. In: *Computer Vision Systems* (A. R. Hanson and E. M. Riseman, Eds.). New York: Academic Press.
5. Hunt, E.B. 1962. *Concept Formation: An Information Processing Problem.* New York: Wiley.
6. Honavar, V. and Uhr, L. 1989. Generation, Local Receptive Fields and Global Convergence Improve Perceptual Learning in Connectionist Networks. *Proc. of the Eleventh IJCAI,* San Mateo, CA: Morgan Kaufmann.
7. Li, Z.N. and Uhr, L. 1987. Pyramid vision using key features to integrate image-driven bottom-up and model-driven top-down processes. *IEEE Trans. on Systems, Man, and Cybernetics 17.*
8. Mani, G. and Uhr, L. 1991. Integrating Perception, Language Handling, Learning and Planning in the CHILDLIKE System. Working notes of the *AAAI Spring Symposium on Integrated Intelligent Architectures.* Also to appear in *SIGART.*
9. Mitchell, T.M. (1982). Generalization as Search. *Artificial Intelligence 18(2).*
10. Nenov, V.I. and Dyer, M. G. 1988. DETE: Connectionist/Symbolic Model of Visual and Verbal Association. *Proc. of the IEEE International Conference on Neural Networks. Vol. 2.*
11. Osherson, D.N. & Lasnik, H. 1990. *Language: An Invitation to Cognitive Science, Vol. 1.* Cambridge, MA: MIT Press/Bradford Books.
12. Quinlan, J.R. 1986. Induction of Decision Trees. *Machine Learning 1.*
13. Tanimoto, S. and Klinger, A. (Eds). 1980. *Structured Computer Vision.* New York: Academic Press.
14. Uhr, L. 1978. 'Recognition Cones' and some test results. In: *Computer Vision Systems* (A. R. Hanson and E. M. Riseman, Eds.). New York: Academic Press.
15. Uhr, L. (Ed.) 1987. *Parallel Computer Vision.* Boston: Academic Press.
16. Weber, S. H. and Stolcke, A. 1990. L_0: A Testbed for Miniature Language Acquisition. ICSI TR-90-010. Berkeley, CA.