Acquiring ontological categories through interaction

Luc Berthouze and Adriaan Tijsseling
Neuroscience Research Institute
AIST Tsukuba Central 2
Umezono 1-1-1, Tsukuba 305-8568, Japan
voice: [+81](298)61-5369; fax: [+81](298)61-5841
e-mail: {Luc.Berthouze,Adriaan.Tijsseling}@aist.go.jp

Abstract

We present an experimental framework for the study of a computational agent's acquisition of ontological categories through interaction with a human agent. This interaction takes the form of joint attention in a shared visual environment. Through a game of guessing words, the computational agent adapts its categorization system - a CALM-based neural architecture with a set of Gabor filters - and develops a shared lexicon with the human agent. This paper report initial experimental results.

Keywords: Perceptual categorization, ontology, lexicon formation.

1 Introduction

The social interactionist theory of language stresses the importance of the environment and the interaction with the caretaker in the acquisition of language by an infant. In the early stages, it is pre-verbal communication that provides the basis of communicative competence and entry into linguistic meaning. Joint visual attention - or deictic gaze - and pointing are two important mechanisms. Onset of pointing, at about 11 months, is shown to correlate with the comprehension of object names [HC95]. Children as young as 12-18 months spontaneously check where a speaker is looking when she utters a word that is new to them, and link the word with the object the speaker is looking at [BM96]. In this way, infants avoid many potential word learning errors. However, they still face the so-called *Gavagai problem* [Qui60], i.e. the difficulty of associating a unique meaning to a word. In human communication, feedback is obtained on the communicative success of the interaction, not on the meanings that were used [SK99]. Thus, the infant is left with *acquiring word-meaning and meaning-object relations which are compatible with the word-object co-occurences they overtly observe but without observing word-meaning relations directly.* A natural hypothesis is that the child will ground the meaning-object relationship into perceptual categories - a process called *sensorimotor toil* [CGH00].

In this paper, we describe our experimental framework for the study of this grounding process.

2 Scenario

Taking inspiration from the Talking Heads experiments [SK99], we model the typical mother-child game play as a game of guessing words between a visually grounded computational agent and a human agent, both sharing a common visual scene. This scene displays a variety of two-dimensional objects of various shapes, colors and sizes, pasted on a white board and subject to variations in illumination.

A guessing game proceeds as follows:

1. Using a video camera, the human agent selects and attends to an object in the visual scene. S/he doesn't provide any linguistic cue.

2. Simulating joint visual attention, the computational agent inherits the same viewpoint and categorizes its perception of the visual scene. If its perception results in a known category - a category that was associated to a word in past experiments - the agent displays

the associated word. Otherwise, it produces a random word.

3. The game succeeds if the word produced by the agent is equal to the human agent's word. In case of a failure, the human agent provides his own word to the computational agent, without any extra-linguistic hint. The computational agent subsequently adapts its categorization system.

3 Computational agent

The computational agent has two main components, a perceptual apparatus based on Gabor filters and a neural network architecture based on CALM (Categorization and Learning Modules) networks.

3.1 Perceptual apparatus

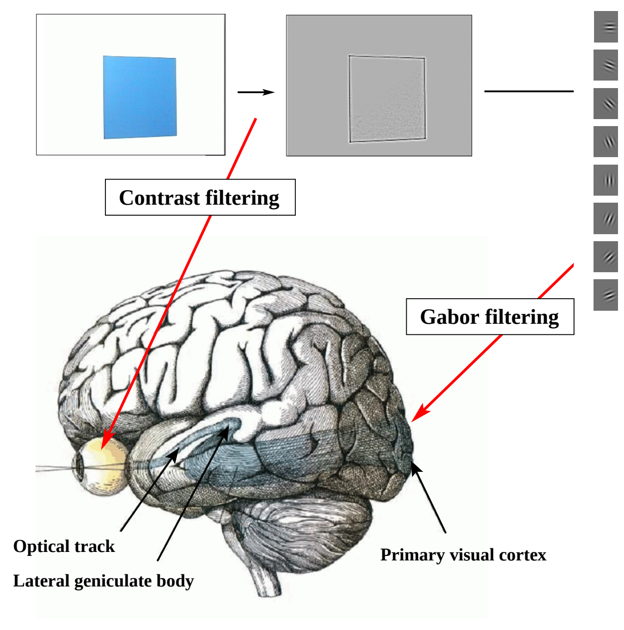

Figure 1: Information flow in the agent's perceptual apparatus and its mapping onto the brain's visual pathway.

The computational agent perceives its environment through a two degrees of freedom articulated robotic head whose motors are controlled via serial cable. The images, decomposed in RGB channels [1], are processed by a series of filters loosely resembling the processing performed in the brain's visual pathway (see figure 1). At first, the contrast is enhanced by a contrast filter, which simulates the profile of retinal ganglion cells. The filter is based on a model of retinal signal processing by Atick and Redlich [AR92] and helps to attenuate low and boost high frequencies while also reducing the effects of noise. The contrast filtered image is subsequently subjected to a set of Gabor filters to extract local orientation selective edge features, similar to the function of simple cells of the cat visual cortex [Dau89]. We generate a local description of the image by taking a set of features, i.e. the responses of a set of Gabor filters distributed over the entire image. This response measures a certain local frequency of the image property at the considered location. We used 32x32 pixel size Gabor filters, spread out over the entire image with an overlap of 4 pixels and tuned to 8 different angle orientations (0.0, 22.5, 45.0, 67.5, 90.0, 112.5, 135, 157.5), a single frequency and a sigma of 2.0. The output of all filters was normalized before presentation to the categorizing network.

3.2 Categorization system

The categorizing network is based on CALM neural networks.

CALM

CALM is a neural network algorithm introduced by Murre et al. [Mur92] and stands for *Categorizing and Learning Module*. A CALM network is built up of several inter-connected modules. While the weights on the links between modules are modifiable - learning proceeds by adapting these weights -, the modules themselves are fixed in structure and function. It is thought that modularity will provide a coarse initial architecture on which learning will impart a finer structure. The idea behind CALM is that the richness of human cognitive behavior is related to the modular structure of both the human brain and human information processing. The latter is characterized by relatively isolated subsystems that can function quite independently of each other. We can, for example,

[1] We used RGB colorspace for Gabor filtering as opposed to opponent color features or HSV colorspace, because the former has been shown to produce high classification results for a wide range of color textures [PKLS00].

speak while driving a car [SML85], but this ability is lost as soon as two behaviors share the same subsystem as when we try to lip-read while driving. Another important property of this modular organization of information processing is that isolated abilities, such as face-recognition [DDH82], may be lost without affecting other cognitive abilities in any way [Lur73]. Structural modularity is not only found at a functional level, but also at the architectural level. The human neural system is characterized by many structurally different cortical and subcortical centers that are only partly connected [LH88, KFAW90]. The neocortex itself is organized into functional columns. Each such cortical column or module forms a micro-circuit that is characterized by strong inhibition between neurons within the same column and a long-range excitation with neurons in other columns. This suggests that signal receiving neurons in a column are in competition with each other [Mou75, Sze75, Ecc81, Koh88]. Further advantages of the CALM architecture are an increased stability of learned representations and a reduced interference by subsequent learning, due to reduced plasticity of the network architecture.

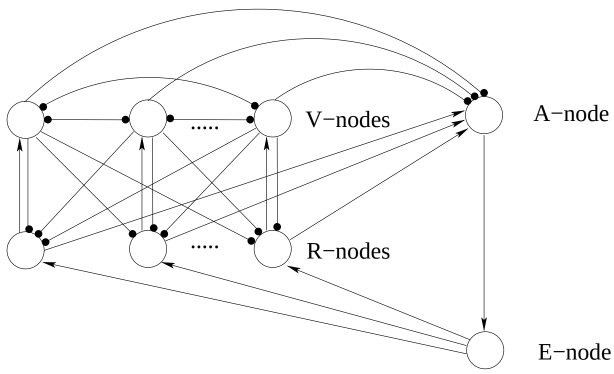

Figure 2: Architecture of a CALM module. Input information is fed to the R-nodes via all-to-all connections. These connections are not represented for readability. An arrow denotes an excitatory connection while a dot-terminated connection denotes an inhibitory connection. There are as exactly as many V-nodes as there are R-nodes.

As shown in figure 2, a module consists of four distinct nodes. Information enters a module via the representation nodes (R-nodes). In other words, connections between modules are actually connections between the R-nodes of different modules. The other nodes implement a competitive mechanism such that there is always one winner for each input pattern. Veto- or V-nodes have an inhibiting function. Every V-node receives excitatory activation from exactly one paired R-node and will in turn inhibit all other nodes inside a module to trigger competition. Because some input may be learned before and some input may be new to the network, another mechanism is needed to make it possible to distinguish between learning new information and maintaining old information. The amount of activation of the arousal- or A-node indicates the level of competition in the module, because it receives activation from all R-nodes. A previously learned input already has a winning R-node, whereas new input is likely to activate several R-nodes at once. The level of activation in the A-node in turn determines activation of the external- or E-node. The E-node is important in two ways. First, it sends random activation pulse to all R-nodes based on the amount of activation it receives. These random pulses force the competition mechanism to decide on a winning R-node in event some R-nodes are equally activated (especially at the start of the learning). Second, E-node activation modifies the learning parameter in the module, such that the learning is low when arousal is low and high when arousal is high. In other words, if the pattern presented has been encountered before, weight modification is moderate, but if the pattern is new, weight adaptation should be more dramatic [Tij98].

In their basic form, CALM networks perform unsupervised learning; i.e. the choice of a winning R-node is left to the competition mechanism and the noise injected by the E-node. In our context however, feedback is necessary to model mother feedback when the game is failed. Thus, it is necessary to add a mechanism that enables an external process to specify the winning R-node. We used a mechanism based on Tijsseling [Tij98]. In unsupervised CALM modules, the winner is decided based on previously learned associations and random pulses from the E-node. With supervision, the R-nodes of the module designed to be the output module do not receive random pulses, but instead, the desired winner receives a positive pulse from the E-node while all other R-nodes receive negative pulses. The advantage of using E-node pulses instead of plainly setting the desired winner's activation to 1.0 is that there will still be sufficient competition so that weight changes are more stronger. The learning rate for modifying weights towards this output module is also slightly larger than the learning rate for all other connections. In the case there is a recurrent connection from the output module, connected modules adapt their weights without computing the background activation so that the feedback information becomes a stronger determinant for learning associations.

Implementation

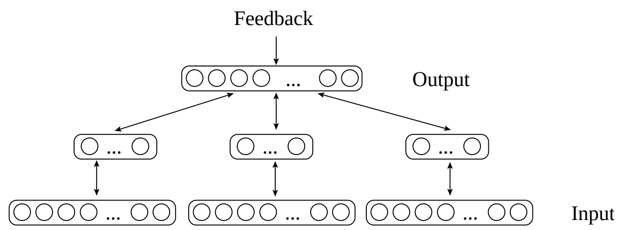

Figure 3: CALM architecture used in this paper. The three input modules correspond to the combined output of the Gabor filters on each RGB channel. These modules are of size 88. Intermediate modules are of size 12. Finally, the output and feedback modules are of size 6, though this size is a parameter deciding the size of the lexicon to be constructed.

The Gabor filter responses for each separate RGB channel of each image form an input vector of length 88 and are fed into the corresponding input modules of figure 3. These inputs modules are connected to intermediate modules of size 12, each of which will categorize the local orientation frequencies of one specific color channel. The processed information from each of these intermediate modules is propagated to the output module. The size of the output module is determined by the size of the lexicon to be constructed, since each neuron will be associated to a word. The smaller the size, the more general the resulting categories are. Conversely, a high number of output neurons can result in an over-specialized lexicon or even in redundant categories. In our experiments, a small size (6) was considered. This module is recurrently connected to the intermediate modules in order to have the feedback information correct

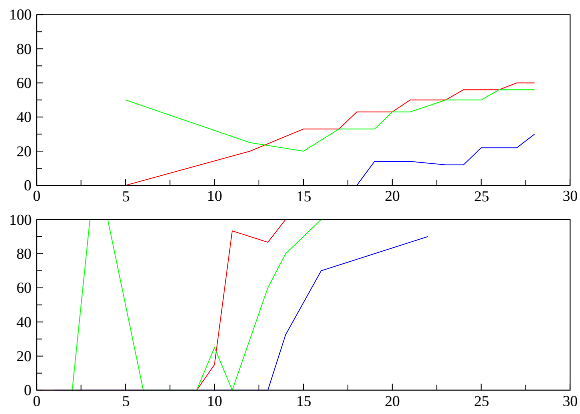

Figure 4: Learning of three basic categories. The horizontal axis denotes the number of games. The vertical axis denotes the success percentages for each word (*green*, *red* and *blue*) in testing (bottom) and actual occurence (top).

possibly unwanted clusterings. The learning rate for the weights to the output module has been set to 0.00125, with the learning rate for all other connections set to 0.0005. Base rate learning has been set to be low, 0.000005, to prevent unwanted shifts in winning nodes.

4 Experiments

The following experimental protocol was used. The shared visual scene consisted of three shapes of different colors: A green triangle, a red disk and a blue square. The human agent elected to label shapes according to their color rather than their shape, i.e. the words *green*, *red* and *blue* were used as feedback in the guessing games. When a game failed - the word suggested by the system did not match the human's word -, the neural network of the agent was trained using the word of the user as feedback. In order to avoid unbalanced weight adaptation, a memory of the last 10 patterns was maintained and included in the learning phase. At each losing game, 10 learning epochs were realized. Learning phases and testing phases (which do not affect the weights) were interleaved so that the quality of learning could be evaluated.

Figure 4 shows the success percentages of each neuron (i.e. of each word) in testing (bottom) and in actual presentations (top) during the acquisition of three labels. The data consistently show one category having an early advantage - af-

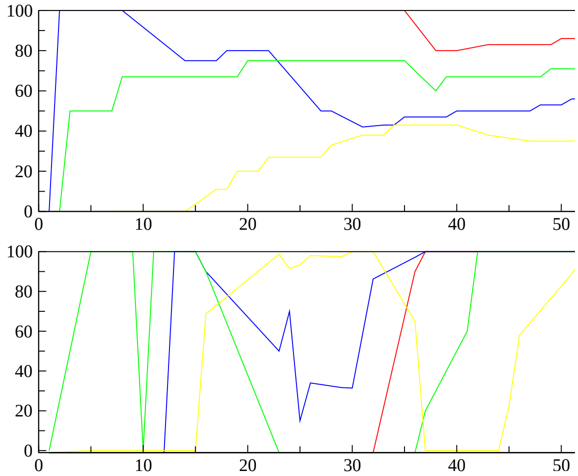

Figure 5: Effect of the addition of a yellow square in the visual scene. The horizontal axis denotes the number of games. The vertical axis denotes the success percentages for each word (*green*, *red*, *blue* and *yellow*) in testing (bottom) and actual occurence (top).

ter about 4 games. This advantage disappears in the next few games and at about 20 games, all categories show an asymptotic increase in success rate. The early advantage of the green category (in this experiment) is due to the competitive nature of the network and the fact that it starts out with equal weights overall. In the initial stages of learning, CALM will converge on one winner for all presented patterns, and only with further learning, specialization occurs. In this case, all shapes were named with the label corresponding to the green object. With more training runs, differences between objects become more apparent and weights are shifted to account for the newly discovered invariants. In addition, the network also has to take into account the provided feedback signals. The success rate of the green object will consequently drop back to the level of the other objects. Repeated exposure to all objects then accounts to the balanced weight adaptation and an increased success rate overall. It is not unlikely that the system initially groups all input under the same label. The lack of previous knowledge prevents finer grained categorization.

Since feedback does not include any information as to what characteristic of the shape should be considered in the categorization, we tested the agent's behavior when a new object is added, whose characteristics only partially overlap those of an already categorized object. Figure 5 describes the effect of adding a yellow square to the visual scene (at game 5). The first presentation of the yellow square produced the label *blue*, i.e. the label associated with the blue square. The figure shows that learning to differentiate the two differently colored squares required a relatively large number of epochs. The tight correlation between the success rates of the neurons *blue* and *yellow* clearly indicate the conflicting information that both shapes are squares. After about 55 games, the agent can successfully discriminate between the two objects.

Although one may argue that shape information is often prevalent over color information in natural scenes [BJ88], the network's difficulty in separating the two differently colored squares is more a consequence of the restricted network architecture that proved too biased to shape information. The unichrome Gabor filtering already puts more weight on local orientation information as compared to the differences in the red, green, and blue color components. In future work, we will investigate alternative network architectures that alleviates this imbalance. Only then can we reliably verify the stability and flexibility of categories by teaching the agent that, for example, the label *blue* indicates the color of an object and not the shape or both the color and shape of an object.

Our last experiment aimed at assessing the generalization capability of the network and the categories it constructs. In a first phase, the network was presented with four shapes, two of them being labelled *disk* (see figure 6, left). After 39 games, the network was able to succesfully categorize the four shapes. Two new disks (see figure 6, right) were then added to the scene. Figure 7 shows a temporary decrease of the performance of the la-

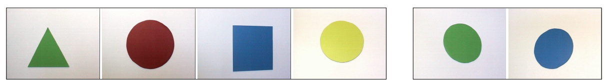

Figure 6: Shapes used in the disk experiment. In the first stage of the experiments, only the first four shapes were used. The last two disks were used to test the generalization ability of the network *disk* category.

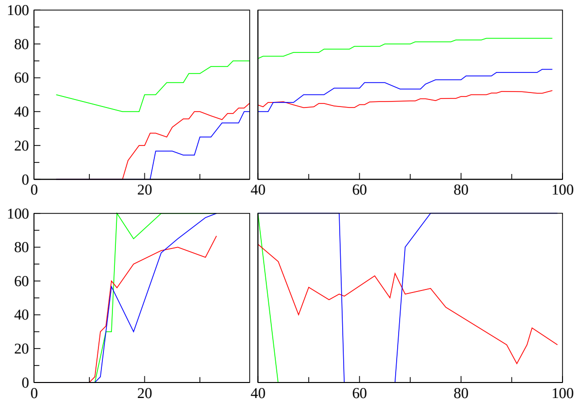

Figure 7: Learning the word *disk* (red time-series). Initial phase with two instances of disks (left) and effect of the addition of two new instances of disks (right). The horizontal axis denotes the number of games played. The vertical axis denotes the success percentages for each word (*green*, *disk* in red and *blue*) in testing (bottom) and actual occurence (top).

bel *disk*, along with a slight re-balancing of the weights, followed by an asymptotic increase.

Though the experiment seems to indicate a potential for generalization, we believe that further testing is necessary. Gabor filtering is neither location-invariant nor scale-invariant therefore it is more likely that the network actually learns to associate the label *disk* to the four shapes involved and the extent of generalization will be bound by the network's capacity.

5 Conclusion

Mother-child interaction - also called motherese - is thought to play an important role in the acquisition of a lexicon by young children. Infants use nonverbal cues of other's attentional focus, e.g. gaze, to determine novel linguistic referents. It is our hypothesis that infants use sensorimotor toil to ground word-meaning relations into perceptual categories. In this paper, we described an experimental framework to study this acquisition of an ontological categorization. We proposed a visually grounded computational agent, capable of categorizing its perceptions with its neural architecture based on CALM neural networks. Together with a human agent sharing a same visual environment, the agent played games of guessing words, thus simulating early mother-child interaction. Our preliminary experimental results show that the platform is suitable to study the dynamics of ontological acquisition. Categories have been shown to be flexible yet stable over time. The issue of generalization capability has been touched but results are inconclusive at this stage.

Irrespective of its validity in accounting for the acquisition of ontological categorization in children, this work illustrates that given a proper sensorimotor and neural apparatus, interactive systems, such as humanoid robots for example, could acquire ontological categorization and develop a common lexicon with their users. This is an interesting development as computer systems are nowadays evolving into personal interactive systems, active in both our physical and social environment.

Acknowledgments

T. Kurita made valuable suggestions regarding the use of contrast and Gabor filters. M. Pic contributed to the development of the software platform. A. Tijsseling was supported by a fellowship of the Science and Technology Agency (STA) of the Japanese government.

References

[AR92] J.J. Atick and A.N. Redlich, *What does the retina know about natural scenes?*, Neural Computation **4** (1992), no. 2, 196–210.

[BJ88] I. Biederman and G. Ju, *Surface versus edgebased determinants of visual recognition*, Cognitive psychology **20** (1988), 38–64.

[BM96] D. Baldwin and L. Moses, *The ontogeny of social information-gathering*, Child development **67** (1996), 1915–1939.

[CGH00] A. Cangelosi, A. Greco, and S. Harnad, *From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories*, Connection science **12** (2000), 143–162.

[Dau89] J. Daugman, *Entropy reduction and decorrelation in visual coding by oriented neural receptive fields*, Transactions on Biomedical Engineering **36** (1989), no. 1, 107–114.

[DDH82] A.R. Damasio, H. Damasio, and G.W. Van Hoesen, *Prosopagnosia: Anatomic basis and behavioral mechanisms*, Neurology **32** (1982), 331–341.

[Ecc81] J.C. Eccles, *The modular operation of the cerebral neocortex considered as the material basis of mental events*, Neuroscience **6** (1981), 1839–1855.

[HC95] A. Harris and L. Campbell, *Historical syntax in cross-linguistic perspective*, Cambridge: Cambridge University Press., 1995.

[KFAW90] S.M. Kosslyn, R.A. Flynn, J.B. Amsterdam, and G. Wang, *Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes*, Cognition **34** (1990), 203–277.

[Koh88] T. Kohonen, *Self-organization and associative memory*, 2nd ed., Berlin: Springer Verlag, 1988.

[LH88] M. Livingstone and D.H. Hubel, *Segregation of form, color, movement, and depth: Anatomy, physiologym and perception*, Science **240** (1988), 740–749.

[Lur73] A.R. Luria, *The working brain: an introduction to neuropsychology*, New York: Basic, 1973.

[Mou75] V.B. Mountcastle, *The mindful brain*, ch. An organizing principle for cerebral function: The unit module and the distributed system, Cambridge, MA: MIT Press, 1975.

[Mur92] J.M.J. Murre, *Learning and categorization in modular neural networks*, Hillsdale, NJ: Lawrence Erlbaum, 1992.

[PKLS00] C. Palm, D. Keysers, T. Lehmann, and K. Spitzer, *Gabor filtering of complex hue/saturation images for color texture classification*, Proceedings of the joint conference on information sciences - International conference on computer vision, pattern recognition and image processing, Atlantic City, NJ (USA) **2** (2000), 45–59.

[Qui60] W. Quine, *Word and object*, MIT Press, Cambridge Ma., 1960.

[SK99] L. Steels and F. Kaplan, *Linguistic evolution through language acquisition: formal and computational models*, ch. Bootstrapping Grounded Word Semantics, Cambridge: Cambridge University Press, 1999.

[SML85] T. Shallice, P. McLeod, and K. Lewis, *Isolating cognitive modules with the dual-task paradigm: Are speech perception and production separate processes?*, The quarterly journal of experimental psychology **37A** (1985), 507–532.

[Sze75] J. Szentágothai, *The "module concept" in the cerebral cortex architecture*, Brain research **95** (1975), 475–496.

[Tij98] A. Tijsseling, *Connectionist models of categorization: a dynamical approach to cognition*, Ph.D. thesis, Southampton University (England), 1998.