Online Learning for Attention, Recognition, and Tracking by a Single Developmental Framework

Juyang Weng and Matthew Luciw
Department of Computer Science and Engineering
Michigan State University
{weng, luciwmat}@cse.msu.edu

Abstract

It is likely that human-level online learning for vision will require a brain-like developmental model. We present a general purpose model, called the Self-Aware and Self-Effecting (SASE) model, characterized by internal sensation and action. Rooted in the biological genomic equivalence principle, this model is a general-purpose cell-centered in-place learning scheme to handle different levels of development and operation, from the cell level all the way to the brain level. It is unknown how the brain self-organizes its internal wiring without a holistically-aware central controller. How does the brain develop internal object representations? How do such representations enable tightly intertwined attention and recognition in the presence of complex backgrounds? Internally in SASE, local neural learning uses only the co-firing between the pre-synaptic and post-synaptic activities. Such a two-way representation automatically boosts action-relevant components in the sensory inputs (e.g., foreground vs. background) by increasing the chance of only action-related feature detectors to win in competition. It enables develop in a "skull-closed" fashion. We discuss SASE networks called Where-What networks (WWN) for the open problem of general purpose online attention and recognition with complex backgrounds. In WWN, desired invariance and specificity emerge at each of the what and where motor ends without an internal master map. WWN allows both type-based top-down attention and location-based top-down attention, to attend and recognize individual objects from complex backgrounds (which may include other objects). It is proposed that WWN deals with any real-world foreground objects and any complex backgrounds.

1. Introduction

Demonstrated by human cognitive and behavioral development from infancy to adulthood, autonomous development is nature's approach to human intelligence (Piaget 1954 [22], Carey 1990 [3], Elman et al. 1997 [7], Quartz & Sejnowski 1997 [24], Weng et al. 2001 [30]). Humans easily perform vision tasks that no machine can accomplish, and the capacity of the human brain for learning to visually perceive new objects, trajectories, etc. is also unmatched in machines. Therefore, understanding the brain's mechanisms will lead to solving the problems of large-scale online visual representation and learning.

Fundamentally, there remains a large gap between connectionist modeling and symbolic modeling. Connectionist modeling works on low-level sensory data, while symbolic modeling deals with high-level abstract symbols, but there is relatively little study on intermediate representations. This paper focuses on online learning of intermediate representation suited for invariant object recognition, scene classification, spatiotemporal event detection, attention in the presence of complex natural backgrounds, and goal-directed top-down reasoning with pixels. There have been several impressive attempts to model the brain as a symbolic information processor, such as Hecht-Nielsen 2007 [12] and Albus 2010 [1]). However, they are without sufficient, biologically plausible learning to account for the overwhelming brain complexity. As symbol-only modeling is insufficient to deal with uncertainty, the Bayesian probability framework was added to such symbolic models, using either probability models for spatial aspects or Markov chains for temporal aspects (Lee & Mumford 2003 [18], Emami & Jelinek 2005 [8], Tenenbaum et al. 2006 [26], George & Hawkins 2009 [11]). A major advantage of such Bayesian models is that they are intuitive for a human to understand, but they face a fundamental problem: They are not *developmental* —- the symbolic boundaries ("walls") between different internal units (nodes or Markov chains) were handcrafted (defined) by human programmers. By developmental, we mean that the internal representation is fully emergent from interactions with environment, without allowing a human to manually instantiate a task-specific representation.

978-1-4244-7028-0/10/$26.00 ©2010 IEEE

In fact, they all correspond to an "skull-open approach" to the brain — it is the human teacher or programmer who understands a given task and the concepts it needs. Then he directly manipulates (defines) the "brain's" internal representation through its open "skull".

The major limitations of such "skull-open" approaches are 1. that it is labor-intensive to build and 2. it is brittle for real world environments. Given a task, the process of handcrafting a task-specific information processor requires a large amount of man-hours for manual instantiation of an internal representation for each task. The resulting system is known to be brittle due to the inability of a human to sufficiently predict the dynamic real world. Many computer vision researchers thought that the human vision system can be sufficiently modeled by a static object recognizer. It cannot. The brain learns new objects and new object variations all the time.

1.1. Development is Necessary

The necessity of a developmental approach is supported by the following main points:

1. Less brittle. All the traditional machine learning methods, including many neural network methods, require an "skull open" approach. This way, the holistically-aware central controller, at least at the linking ports of the separately trained modules, is the human teacher. This holistically-aware human central controller implants static meaning "walls" which lead to a brittle "brain", because no static meaning "walls" appear sufficient for dynamic real-world environments. For example, he may specify "oriented edges" (Marr 1982 [21]) or "SIFT feature" (Lowe 2004, [20]) as a static representation for the feature module, but a fixed feature type is insufficient for all vision tasks in all dynamic natural environments.

2. Lifetime adaptation and understanding. For humans, the *developmental program* (DP) is responsible for whatever can happen through the entire life. For machines, the DP enables the agent to develop its mental skills (including perception, cognition, behaviors and motivation) through interactions with its environment using its sensors and effectors.

3: More tractable as humans are relieved from task-specific programming. The DP enables machines to learn new tasks that a human programmer does not know about at the time of programming. The DP (i.e., genome) must define relatively simple local learning rules and the statistics of experience drive learning towards intelligence. On the other hand, it has been argued [27] that the intelligence of a special purpose, traditional AI machine is due mainly to the intelligence of its programmer, not really the intelligence of the machine. This is due to the human programmer acting as its external "central" controller — the joint task executor.

4. Scaffolding. Early learned skills should assist in the learning of more complicated new skills in carefully designed settings and later such new skills are further consolidated in later less structured, more general settings (Domjan 1998 [6], Zhang & Weng 2007 [31]). For example, knowledge of a learned object's parts could be used to quickly learn (develop representation for) a new object that contains some of those parts.

2. General Perception: Complex Background Problems

A popular class of problems is called scene classification. Complex features (e.g., patterns or SIFT features) are detected from small patches across the entire image. The locations of all features are discarded, resulting in what is called "bag of features". If the features are sufficiently discriminative for classifying a scene type or for recognizing an object, such methods can be used to classify scenes or even for recognizing objects from general backgrounds (Fei-Fei, 2006) [9], Poggio & coworkers [25]). However, we can expect that the performance will depend on how discriminative the features are. DiCarlo and coworkers [23] reported that such methods have problems in real-world visual object recognition. Recognizing general objects, of which the type is not known at programming time, from complex backgrounds has not been addressed satisfactorily, and this remains an open problem.

Let us consider the formalization of the complex background problem. In the following, we will use image as a temporal sample of all receptors of a sensory modality, visual, auditory, touch, spell or taste. We will concentrate on vision. The set of all possible background images is:

$$B = \{\mathbf{b} \mid \mathbf{b} \in R^d \text{ is an image of the real world}\}$$

which is infinite because of the infinitely large world. Consider a foreground object of type $\mathbf{t} \in T$, location $\mathbf{l} \in L$ and further the vector $\mathbf{w} \in W$ denotes all other possible properties $\mathbf{w}$ (e.g., object orientation, the distance from the viewer, lighting, etc), respectively. The set of all foreground images $\mathbf{f}$ is:

$$F = \{\mathbf{f}(\mathbf{t}, \mathbf{l}, \mathbf{w}) \in R^d \mid , \mathbf{t} \in T, \mathbf{l} \in L, \mathbf{w} \in W\}$$

which is also infinite. Suppose that the pixels in a foreground image that do not receive optical projection of the corresponding object has a unique value "transparent." An input image with background is a composite image $\mathbf{x} = \mathbf{b} \vdash \mathbf{f}$ where the *projection operator* $\vdash$ denotes transparency-based foreground overwrite: Each pixel in $\mathbf{x}$ takes the corresponding pixel value of $\mathbf{f}$ if it is not transparent and otherwise the corresponding pixel value of $\mathbf{b}$. The set of all possible input images *with backgrounds* is then

$$X(B, F) = \{\mathbf{x} \mid \mathbf{x} = \mathbf{b} \vdash \mathbf{f}, \mathbf{b} \in B, \mathbf{f} \in F\}$$

which is again infinite. Through development, an embodied brain, natural or artificial, samples the images in X actively and incrementally, as the natural consequence of its interactions with the physical world.

Problem 1 (Attention-recognition — batch) *Consider a finite set of m training images from the background B and foreground F,*

$$\Sigma(B, F) = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m \mid \mathbf{x}_i \in X(B, F)\},$$

together with the ground truth $G = \{(\mathbf{t}_i, \mathbf{l}_i) \mid \mathbf{t}_i \in T, \mathbf{l}_i \in L, i = 1, 2, ..., m\}$. A test image $\mathbf{x} = \mathbf{b} \vdash \mathbf{f} \in X(B, F)$ is not in the training set, $\mathbf{x} \notin \Sigma(B, F)$, but its foreground part $\mathbf{f}$ is similar to at least some foreground parts of images in the training set but the same is not necessarily true for the background part $\mathbf{b}$. Determine a batch attention-recognition algorithm f_{AR-b} which takes $\Sigma(B, F)$, G, and $\mathbf{x}$ as the input and return the predicted class label $\mathbf{t}$ and location $\mathbf{l}$: $(\mathbf{t}, \mathbf{l}) = f_{AR-b}(\mathbf{x} \mid \Sigma(B, F), G)$.

It is clear that the brain cannot solve this problem in such a batch fashion, as the amount of data in Σ is too large.

Further, continuity of objects as they move in space may also be useful (e.g., object permanence [2]) as the reality typically occurs continuous in real time.

Problem 2 (Attention-recognition — developmental) *After initialization, develop an embodied agent, natural or artificial, through interaction with the real physical world that determines the background B and the foreground F:*

$$(S_{n+1}, M_{n+1}, R_{n+1}, N_{n+1}) \leftarrow f_{AR-d}(S_n, M_n, R_n \mid N_n)$$

for $n = 1, 2, ..., m$, where the discrete index n is for time $t_n = t_0 + n\tau$, $S_n \in X(B, F)$ is the observation of the background and foreground, M_n may occasionally contain ground truth $\mathbf{g}$ but not all the time, R_n the internal response, N_n the adaptive part of f_{AR-d} and $\mathbf{g}$ a part of ground truth related to time t_n. During future times $n = m + 1, m + 2, ..., m + j$, with the testing length $j \geq 1$, without imposition of all the motor effectors, the agent function f_{AR-d} autonomously produces motor outputs $(M_{m+2}, M_{m+3}, ..., M_{m+j+1})$ that are consistent with typical agents in this age group of the species.

Note that "occasionally contain ground truth" allows type bias and location bias from the teacher, if so desirable, but not always so that the teacher can let the agent practice.

3. Brain Architecture

3.1. Neuro-anatomy

How does the brain solve complex background problems? First, let us look at neuro-anatomy. Regulated by the genome, the central nervous system develops very extensive and complex processing hierarchies through experience. Neuro-anatomic studies have produced rich literature about visual pathways (e.g., Fellman & Van Essen 1991 [10]), auditory pathways (e.g., Kass et al. 1999 [15]), and motor pathways (e.g., Felleman & Van Essen 1991 [10]). Each sensing modality (visual, auditory, touch, etc) corresponds to a different sensory pathway. Each may diverge to multiple pathways in the cortex. Each of these pathways occupies different cortical areas and they may converge. Unimodal sensory inputs converge on multimodel association areas. There are three major convergence areas in the cortex: prefrontal, parieto-temporal and limbic cortices (Kandel et al. [16, p. 355]). They all further link with the motor areas (external muscles and internal glands). Based on this and related knowledge, Fig. 1 gives a simplified connection pattern for a multi-sensory, multi-effector developmental brain. Each sensory pathway consists of a network of cortical areas before reaching one of the three major converging areas. Neurons in early cortical areas typically have smaller receptive fields than those in later areas. It is known that each sensory pathway is not a single cascade: For example, V1 connects not only V2, but also V3, PIP, V4, MT, etc.

3.2. SASE Brain Architecture

In traditional artificial intelligence, an agent is modeled as something that senses the external environment and acts on the external environment. We argued that the brain must sense internal environment (inside the brain) and act on the internal environment. The *Self-Aware and Self-Effecting* (SASE) brain model and its learning, discussed here, are biologically supported mainly by neuro-anatomy, with some explicit engineered modifications just in its computer implementations for superior efficiency. A core principle is that SASE incrementally forms its sensorimotor pathways primarily based on the co-firing statistics of received signals, in order to internally represent *relevant* sensory input, prioritized over the *irrelevant*. The relevant input is correlated with each action and behavior, controlled by motor neurons. Due to space limits, we can only present a high-level overview of SASE here.

3.2.1 Neuronal areas

Any set of neurons can be a unit in the SASE brain model. Consider a generic area Y which has its bottom-up (closer to sensors, e.g., pixels) area X and its top-down area Z (closer to motors, e.g., output neurons). Because of the need to address the complex background problem, the SASE model must provide a set of "receptive fields" and "effective fields" that are smaller than X and Z, respectively, as illustrated in Fig 2(b). It is desirable for the receptive field

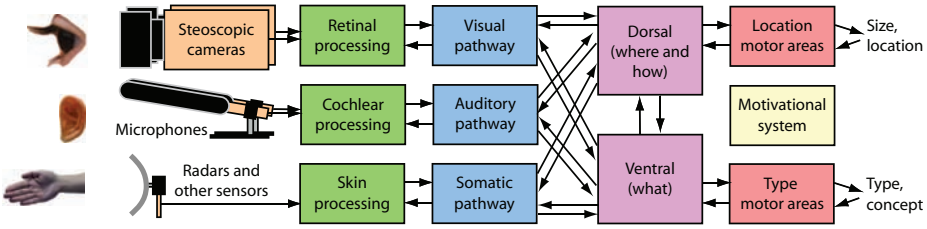

Figure 1. The architecture of a multi-sensory, multi-effector developmental brain in SASE. The multi-sensory and multi-effector integration is achieved through developmental learning. Each area can be served by one or multiple areas. In this figure, "bottom-up" means rightward information flow, and "top-down" means leftward information flow.

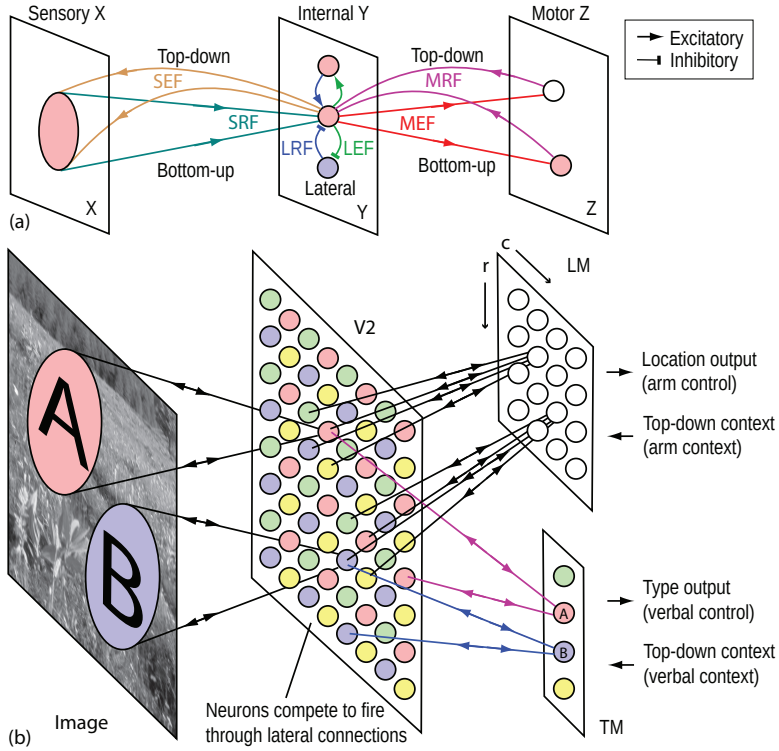

Figure 2. The basic unit of the hextuple representations and an simple Where-Where Network (WWN) example. (a) A basic unit of the hextuple representation that bridges any two arbitrary brain areas. Sensory, lateral or motor receptive or effective fields (e.g., SRF, SEF, etc.) thus are highly recurrent. (b) A simple SASE network for attention with four areas (image, V2, LM and TM) and its hextuple network representation. Each wire connects if the pre-synaptic and post-synaptic neurons have co-fired. The weight is the frequency of pre-synaptic co-firing when the post-synaptic neuron fires. Within each cortical area, each neuron connects with highly correlated neurons using excitatory connections but connect with highly anti-correlated neurons using inhibitory connections. This forces neurons in the same area to detect different features in SRF and MRF. These developmental mechanisms result in the shown connections. Every V2 neuron is *location-specific* and *type-specific*, corresponding to an object type (marked by its color) and to a location block (2×2 size each). Each LM neuron is location-specific and type-invariant (more invariance, e.g., lighting-direction invariance, in more mature nets). Each TM neuron is type-specific and location-invariant (more invariance in more mature nets). Each motor neuron pulls all applicable cases from V2. It also top-down boosts all applicable cases in V2 as top-down context. A two-way arrow means two one-way connections. All the connections within the same area are omitted for clarity. Since V2 is the first area from the image here, V2 does not need explicit SEF connections but all LM and TM neurons have global SEFs.

of a neuron to match the foreground object well, so the response of the neuron is not very sensitive to the irrelevant background.

3.2.2 Output area as another input area

The area Y produces internal representation from bottom-up and top-down connections from lower area X and higher area Z. It senses input from X, but it also produces top-

down signals for X as top-down attention. This is "self-effecting" (internal attention) as Y acts on its sensory area X within the brain. The area Y sends its response to its motor area as its action, but it also receives top-down signals from its motor area Z. This is "self-aware" as Y senses the status of its motor area. In other words, its sensory area is not only its input port but also its output port; similarly its motor area is not only its output port but also its input port.

3.2.3 LCA: developmental neuronal layers

The biological genomic equivalence principle implies that a cell is a general-purpose machine during its development and operation as far as its genome is concerned. Thus, we consider that every area in the brain is of *general purpose* in the sense of its developmental mechanisms. As this model is formulated from neuro-anatomy, the computation and learning are constrained by this *in-place* learning principle, and each neuron is responsible for its own computation and learning. There does not seem to exist any extracellular mechanisms to e.g., compute the probability density.

We developed the Candid Covariance-free Incremental (CCI) Lobe Component Analysis algorithm (LCA) as the local learning and competition rule for every neuronal layer in the architectures like in Fig. 1. Details of its operation are presented elsewhere [28]. It incrementally updates a layer's internal representation: $L = (V, A, r)$, where $V = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c)$ contains c synaptic vectors, $A = (n_1, n_2, ..., n_c)$ consists of the corresponding firing "ages", and r is the radius of the excitation sphere of each neuron. Each input sample is in the form $(\mathbf{x}, \mathbf{z}) = X \times Z$, where X is the bottom-up space and Z is the top-down input space. In a network, X or Z may have multiple parallel input subspaces. Algorithmically, f_{LCA} takes L as the current level and $(\mathbf{x}, \mathbf{z})$ as the bottom-up and top-down input, respectively, to generate the response vector $\mathbf{y} = (y_1, y_2, ..., y_c)$ and update its representation L:

$$(\mathbf{y}, L) \leftarrow f_{LCA}(\mathbf{x}, \mathbf{z} \mid L).$$

3.2.4 Lateral inhibition

Via approximated lateral inhibition in LCA, only the top-k neurons will have non-zero values in $\mathbf{y}$, and thus $c - k$ neurons do not fire, so that they can keep their long-term memory. This is an important advantage over probability based representation and gradient methods, with regards to "lifetime learning", where long-term memory is crucial.

3.2.5 Top-down connections

Primate cortex is characterized by a high-level of recurrent excitation. Our bidirectional connectivity creates a recurrent network. In a layer in the SASE model, at time t_n, the response vector $\mathbf{z}_n \in Z$ at the motor area Z gives the top-down context, e.g., the goal. The response vector $\mathbf{x}_n \in X$ at the sensory area X gives the bottom-up context, e.g., the image input. The internal area Y has c neurons to represent its input space $P = X \times Z$, in the forms of neuronal synaptic vectors:

$$V = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_c), \mathbf{v}_i \in P = (X, Z), i = 1, 2, ..., c.$$

Each $\mathbf{v}_i$ has a different receptive field and a different effective field, as illustrated in Fig. 2. Using the co-firing learning in LCA, adding an internal area Y between any two areas X and Z leads to prediction capability of the desired (supervised) output in Z and prediction capability of the desired top-down attention (predicted image) in X.

However, it is known that such positive feedback connections can lead to unstable systems (e.g., uncontrollable oscillations). Consider a network running at discrete times $t = t_0, t_1, t_2, ...$ Prescreening (via lateral inhibition) for both bottom-up and top-down signal sources separately is necessary so as to disregard weak and irrelevant responses that are distractors before bottom-up and top-down integration, and to avoid such instabilities. This seems to be what the 6-layer laminar cortex does[29]. Specifically, the bottom-up and top-down "integration" layer takes three signal sources: prescreened bottom-up input $\mathbf{x}(t_{n-1})$ as lower features, lateral input $\mathbf{y}(t_{n-1})$ from its own layer as its last response, and top-down input $\mathbf{z}(t_{n-1})$ from prescreened top-down input, all at time t_{n-1}. Through the feature development function modeled as LCA, the integration layer generates its next response $\mathbf{y}(t_n)$ at time t_n as the attention-selected response and to update its level to $L(t_n)$:

$$(\mathbf{y}(t_n), L(t_n)) = \mathrm{f}(\mathbf{x}(t_{n-1}), \mathbf{y}(t_{n-1}), \mathbf{z}(t_{n-1}) \mid L(t_{n-1})) \quad (1)$$

where f denotes the function of LCA. This process is called *attentive context folding*, folding the spatiotemporal information from the three sources into one response vector and the updated cortical layer.

4. Where-What Networks (WWN)

We called SASE networks for developmental visual attention, recognition and tracking Where-What networks, due to the two paths of "What" and "Where". Let us first consider an example: A child is staring at a novel car (indicated by pattern A in Fig. 2) and his brain (via e.g., the pulvinar) suppresses other background sensing neurons as he attends. This leads to the firing of pink V2 neuron in Fig. 2 that best matches the "car" image at the correct retina location. At the same time, his mother repeats "car, car," which excites, through the child's auditory stream, the child's motor neurons for pronouncing "car". (This association should have established before since when the child motor pronounced "car", his auditory stream heard his own "car" —

co-firing.) This corresponds to the firing between the V2 neuron and the pink motor neuron in TM in Fig. 2. Their synapse (both-way) is connected with the Hebbian increment yp_i where p_i is each active V2 neuron. The learning of LM is analogous.

Thus, as car appears at different "retinal" locations, the "car" neuron in TM adds "location" connections while all firing LM neurons add their "car" connections. Suppose the response y is an approximated probability for the event that the neuron detects to occur at the current time. Then the above learning expression incrementally updates the synapse as the sample probability for the pre-synaptic neuron to fire conditioned on that the post-synaptic neuron fires.

All "loser" neurons are not updated and their ages do not advance, serving as the long term memory relative to this context $\mathbf{p}$. Therefore, the role of each neuron as working-memory or long-term memory is dynamic and relative. If it fires, it is part of the current working memory and updates. Otherwise, it is part of the long term memory. Therefore, forgetting occurs only in the details of the nearest matched memory for "unconscious" refinement of skills.

4.1. Experimental Result

We trained an implemented Where-What Network as a simplified but a general purpose vision system for recognizing general objects from complex backgrounds. We used images like those in Fig. 4(a). To simulate a shortage of neuronal resource relative to the input variability, we used a small network, five classes of objects, with images of a single size, and many different natural backgrounds. Both the training and testing sets used the same 5 object classes, but different background images. As there are only 3 V2 neurons at each location but 15 training object views, the WWN is 4/5 = 80% short of resource to memorize all the foreground objects. Each V2 neuron must deal with various misalignment between an object and its receptive field, simulating a more realistic resource situation. Location was tested in all $20 \times 20 = 400$ locations.

Without top-down inputs from motor areas, the network operates in the free-viewing mode. This mode is also called *bottom-up* attention [13, 14] — a salient learned object "pops up" from backgrounds. With WWN, saliency is learned, consistent with experience-dependent saliency reported by Lee et al. [19]. As reported in Fig. 4(b), the network gave respectable performance after only the first round (epoch) of practice. After 5 epochs of practice, the network reached an average location error around 1.2 pixels and a correct disjoint classification rate over 95%. This is the first solution to the joint attention-recognition problem in unknown complex backgrounds with a practical-grade performance in free-viewing mode.

It is known [5, 4, 17] that visual *top-down* attention as operational bias has two types, *location* based and *object* (or feature) based. As we discussed above, the top-down signal from a motor action can represent any human communicable concepts, the deliberate reasoning scheme below is applicable to general abstract concepts.

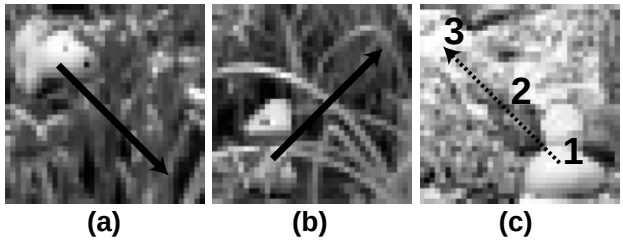

Figure 3. Trajectory examples for tracking. (a) and (b) show continuous trajectories (object is placed at all possible positions along the arrow) while (c) shows a three-frame "jumping" discontinuous trajectory.

We tested the above learned WWN for top-down attention with two competing objects in each image, at four possible quadrants to avoid overlapping. As shown in Fig. 4(e), the success rates are 96% from given type context to predict location and 90% from given location context to predict type. This capability can be thought of as low-level reasoning, an abstract capability, shown here for the first time to develop in a connectionist fashion using low-level signals.

To allow the network to self-generate its own top-down contexts (i.e., abstract "thoughts") like an autonomously "living" animal, we use its *homeostatic mode*. The currently two firing motor neurons in LM and TM get suppressed (simulating temporal depletion of synaptic vesicles which package neural transmitters) and relatively other neurons are boosted concurrently (simulating the disappearance of lateral suppression from the previous winner). WWN correctly deliberatively reasoned for the "runner-up" object (in LM and TM) under this *homeostatic mode* with an average success rate 83% (see Fig. 4(d)).

Tracking an object in WWN occurs via imposed What motor, designating the type of the object the network should look for. There are no built-in constraints, such as assuming that the object will remain nearby its current location or not deform too much. It emerges from observing the object's behavior. Both short range motion and long range motion are therefore tolerated. When we see an object go behind an occlusion, our natural assumption is too assume it has smoothly moved from one spot to the other in the context of its natural movement [2]. It is common to not perceive an uninterrupted trajectory even though it is the case (long-range motion). We testing WWN on two types of tracking: continuous (examples are (a) and (b) in Fig. 3), which uses smoothly changing position, and discontinuous (seen in Fig. 3(c)), which jumps between positions. Each object moved from one corner to another. In the continuous case, the object's position changed slightly between frames.

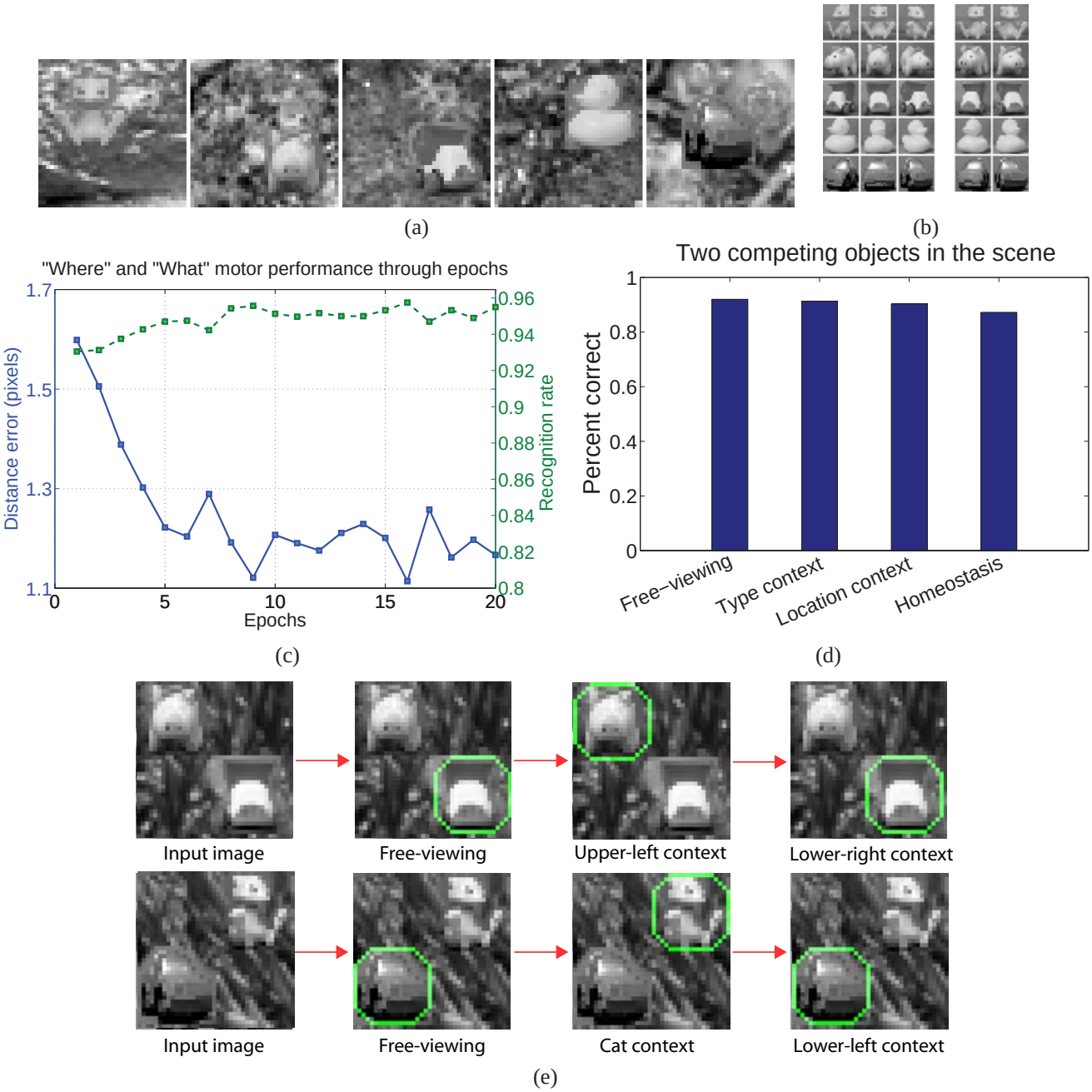

Figure 4. WWNs for the joint attention-recognition problem under the free-viewing mode and reasoning modes. (a) Sample image inputs. (b) Foreground training images (left three) for each object and test images (right two) from different viewing angles. (c) The average errors of the reflexive actions, reaching and telling the type (recognition) during free-viewing in unknown complex natural backgrounds, which improve through epochs of learning experience. (d) Performance when input contains two learned objects: reflexive (free-viewing), two types of deliberative reasoning (top-down type-context and location-context), and fully autonomous deliberative reasoning (homeostasis). (e) Some examples of deliberative reasoning by a trained WWN. "Context" means top-down context. A green octagon indicates the location and type action outputs. The octagon is the default receptive field before synapse adaptation where individual synaptic weights can reduce from LCA learning.

In the discontinuous case, there were only three positions per trajectory. For each frame, if the network guessed position outside of 4 pixels or was wrong in type, we considered it wrong. Otherwise, it was considered correct. Over 5 different tests, with changing backgrounds, the average performance for the continuous tracking was 91.5% with a

deviation of 0.8%. For the discontinuous tracking, the mean performance was 89.5% with deviation of 2.3%.

5. Conclusions

We presented a high-level overview of the SASE brain model for developmental learning. Development is essential to scale up to human level visual learning capabilities. This is the first general purpose developmental model for recognizing general objects from complex backgrounds. This seems to be the first general purpose model that is brain plausible, "skull-closed" and which does not require a human to implant internal representations. A larger variety of tasks need to be learned and tested using the presented SASE model to further demonstrate its generality and computational efficiency. The general purpose deliberative reasoning with both the abstract and the concrete, demonstrated here, has a potential to demonstrate in the future how a machine autonomously learns to think and discover.

References

[1] J. S. Albus. A model of computation and representation in the brain. *Information Science*, 180(9):1519–1554, 2010.

[2] R. Baillargeon. How do infants learn about the physical world? *Current Directions in Psychological Science*, 3:133–140, 1994.

[3] S. Carey. Cognitive development. In D. N. Osherson and E. E. Smith, editors, *Thinking*, pages 147–172. MIT Press, Cambridge, Massachusetts, 1990.

[4] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neural Science*, 3:201–215, 2002.

[5] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[6] M. Domjan. *The Principles of Learning and Behavior*. Brooks/Cole, Belmont, California, fourth edition, 1998.

[7] J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. *Rethinking Innateness: A connectionist perspective on development*. MIT Press, Cambridge, Massachusetts, 1997.

[8] A. Emami and F. Jelinek. A neural syntactic language model. *Machine Learning*, 60:195–227, 2005.

[9] L. Fei-Fei. One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[10] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[11] D. George and J. Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLos Computational Biology*, 5(10):1–26, 2009.

[12] R. Hecht-Nielsen. *Confabulation Theory*. Springer, Berlin, 2007.

[13] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

[14] L. Itti, G. Rees, and J. K. Tsotsos, editors. *Neurobiology of Attention*. Elsevier Academic, Burlington, MA, 2005.

[15] J. H. Kaas, T. A. Hackett, and M. J. Tramo. Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, 9(2):164–170, 1999.

[16] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors. *Principles of Neural Science*. McGraw-Hill, New York, 4th edition, 2000.

[17] E. I. Knudsen. Fundamental components of attention. *Annual Reviews Neuroscience*, 30:57–78, 2007.

[18] T. S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 20(7):1434–1448, 2003.

[19] T. S. Lee, C. F. Yang, R. D. Romero, and D. Mumford. Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, 5(6):589–597, 2002.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[21] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, New York, 1982.

[22] J. Piaget. *The Construction of Reality in the Child*. Basic Books, New York, 1954.

[23] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLOS Computational Biology*, 4:151–156, 2008.

[24] S. Quartz and T. J. Sejnowski. The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences*, 20(4):537–596, 1997.

[25] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[26] J. B. Tenenbauma, T. L. Griffithsb, and C. Kemp. Theory-based bayesian models of inductive learning and reasoning. *Science*, 290:2319–2323, 2000.

[27] J. Weng. Task muddiness, intelligence metrics, and the necessity of autonomous mental development. *Minds and Machines*, 19(1):93–115, 2009.

[28] J. Weng and M. Luciw. Dually optimal neuronal layers: Lobe component analysis. *IEEE Trans. Autonomous Mental Development*, 1(1):68–85, 2009.

[29] J. Weng, T. Luwang, H. Lu, and X. Xue. Multilayer in-place learning networks for modeling functional layers in the laminar cortex. *Neural Networks*, 21:150–159, 2008.

[30] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291(5504):599–600, 2001.

[31] Y. Zhang and J. Weng. Task transfer by a developmental robot. *IEEE Transactions on Evolutionary Computation*, 11(2):226–248, 2007.