The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Perception Coordination Network: A Framework for Online Multi-Modal Concept Acquisition and Binding

You-Lu Xing,[1] Fu-Rao Shen,[2] Jin-Xi Zhao,[2] Jing-Xin Pan,[3] Ah-Hwee Tan[4]
[1]School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[2]Department of Computer Science and Technology, Nanjing University, Nanjing, 210046, China
[3]Medical School, Nanjing University, Nanjing, 210046, China
[4]School of Computer Engineering, Nanyang Technological University, 639798, Singapore.
youluxing@sina.com, {frshen, jxzhao}@nju.edu.cn, jackiejackpan@126.com, asahtan@ntu.edu.sg

Abstract

A biologically plausible neural network model named Perception Coordination Network (PCN) is proposed for online multi-modal concept acquisition and binding. It is a hierarchical structure inspired by the structure of the brain, and functionally divided into the primary sensory area (PSA), the primary sensory association area (SAA), and the higher order association area (HAA). The PSA processes many elementary features, e.g., colors, shapes, syllables, and basic flavors, etc. The SAA combines these elementary features to represent the unimodal concept of an object, e.g., the image, name and taste of an apple, etc. The HAA connects several primary sensory association areas like a function of synaesthesia, which means associating the image, name and taste of an object. PCN is able to continuously acquire and bind multi-modal concepts in an online way. Experimental results suggest that PCN can handle the multi-modal concept acquisition and binding problem effectively.

Introduction

The brain is a hierarchical structure with many function specific modules. For example, neurons which are tuned to a particular color (Livingstone and Hubel 1988), shape (Hegdé and Essen 2000), basic flavor (de Araujo and Simon 2009), and phoneme (Mesgarani et al. 2014) are widely found. These low-level **feature neurons** can correspond to the bottom layer of the brain hierarchy of perception. Then, neurons that respond selectively to particular visual object (Kobatake and Tanaka 1994) and word (Chan et al. 2014) are also discovered. We call them **concept neurons** here, which can be corresponded to the middle layer of the brain hierarchy of perception. Neurons that respond selectively to representations of the same individual across different sensory modalities including vision and audition are detected in the human medial temporal lobe (Quiroga 2012). It is a kind of multi-modal response neuron, which corresponds to the top layer of the brain hierarchy of perception. We name these neurons **associated neurons**. Fig. 1 gives the visual, auditory, gustatory, and olfactory pathways in the brain. The perception in the back of the pathway synthesizes the perception in the front of the pathway, i.e., the sensation becomes complicated through its pathway. Different sensations interact with each other through the areas where different sensory pathways converge. Obviously, such brain structure is efficient and convenient for multi-modal concept acquisition and binding. In (Fuster 1997), an example of concept acquisition and association between vision and touch at the cellular level is given. The explanation is based on the Hebbian theory, which is summarized as "cells that fire together, wire together". As shown in Fig. 2, when a visual and a tactile signal stimulate the network synchronously, a cell-assembly will be formed quickly by the facilitated synapses to associate the visual and tactile sense.

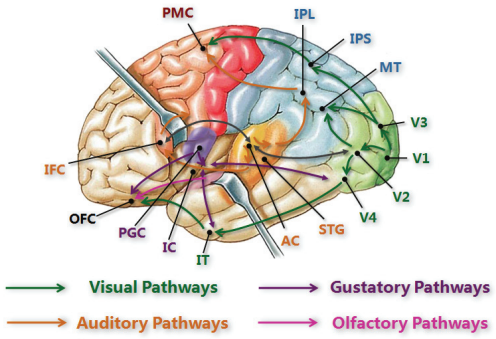

Figure 1: Hierarchical and modularized perception pathway in the brain.

Inspired by the brain's hierarchical structure and the coordination between different functional modules in the structure, a Perception Coordination Network (PCN) is proposed to handle multi-modal concept acquisition and binding between different sensory modules. Briefly, the main contributions are as follows,

(1) Different types of neurons with particular computational models are defined, which makes the hierarchical structure of PCN have a good interpretability.

(2) Through creating of connections between neurons, PCN learns new concepts and bindings without forgetting of already learned concepts and bindings.

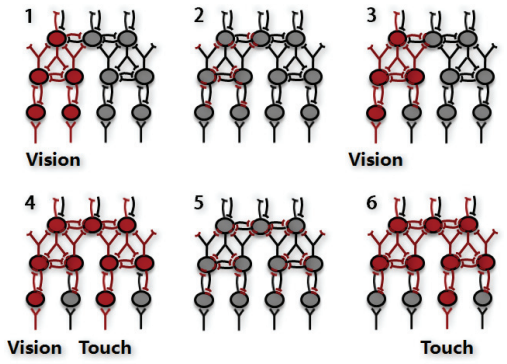

Figure 2: Sensory association at the cellular level. **1.** Two visual inputs coincide in time. **2.** Passive long-term memory formed by the facilitated synapses generated during step **1**, marked in red. **3.** One of the visual inputs activates the subnetwork in step **2**. **4.** A visual and a tactile input coincide. **5.** A bimodal network of long-term memory formed by the facilitated synapses generated during step **4**, marked in red. **6.** The tactile stimulus activates the bimodal network. The figure is reproduced from (Fuster 1997).

Related Work

Many scholars study in sensory integration and multi-modal concept acquisition. In (Schyns 1991), a modular neural network is proposed for concept acquisition, where a self-organizing map (SOM) is used to build concept prototype and a brain-state-in-a-box is used to associate concept names to the output of the SOM. In (Yamauchi, Oota, and Ishii 1999), several neural network modules are fused by some integrating units. The network is able to learn categories of objects by integrating information from several sensors. In (Jantvik, Gustafsson, and Papliński 2011), a multi-modal self-organizing network (MMSON) is proposed for sensory integration with several SOM modules. Positional coordinates of unimodal SOM which receives sensory data are fused by a high level SOM.

In (Nakamura, Nagai, and Iwahashi 2011), a bag of multimodal latent Dirichlet allocation (LDA) is introduced for sensory integration. The bag includes an object categories LDA, a color categories LDA, and a haptic categories LDA. In (Araki et al. 2011), an improved version of multimodal LDA is proposed. When a new object comes to the system, Gibbs sampling is carried out to the new input data iteratively until convergence.

In (Ngiam et al. 2011), a bimodal deep belief network (DBN) is trained to learn a shared representation of visual and auditory input. Firstly, a top restricted Boltzmann machine (RBM) over the pre-trained layers for each modality is used to generate a shared representation of bimodal features. Then, a bimodal deep autoencoder is trained which is initialized with the bimodal DBN weights. Similar approaches are proposed in (Srivastava and Salakhutdinov 2014), which learn joint representation between text features and image features.

In (Parde et al. 2015) and (Thomason et al. 2016), the meaning of words are grounded in visual features by conversations between users and robot. A initial learning phase is needed which leads to the methods cannot deal with words grounding in an totally online incremental way.

Most of the methods above do not have an ability to learn new concepts or new bindings in an online incremental way. But new concepts and bindings always occur in real world. Thus, a better learning system should be able to learn new concepts and bindings continuously, like humans are able to learn new objects and its name without forgetting the previously learned ones throughout their lifetimes. Unfortunately, many learning systems suffer the stability-plasticity dilemma (Carpenter and Grossberg 1988), which means they either cannot learn new knowledge after a period of learning or cannot learn new knowledge quickly without catastrophic forgetting of already learned knowledge. Taking this problem as a target, many online incremental methods for sensory integration are proposed. In (He, Kojima, and Hasegawa 2007), an incremental knowledge robot 1 (IKR1) for word grounding is proposed, where a self-organizing incremental neural network (SOINN) (Shen and Hasegawa 2006) handles the visual module and a vector-quantization (VQ) (Gersho and Gray 1992) system is in charge of the auditory module for words. Integration of words and objects is achieved by associations between SOINN and VQ system. In (Meng, Tan, and Xu 2014), a generalized heterogeneous fusion adaptive resonance theory (GHF-ART) is proposed. It develops the ART (Carpenter and Grossberg 1988) to multi-channel model and can be used for fusion of multi-modal features such as visual and textual features.

Perception Coordination Network

Fig. 3 gives the structure of the PCN. It is a hierarchical structure and functionally divided into the primary sensory area (PSA), the primary sensory association area (SAA), and the higher order association area (HAA). There are three types of neurons including feature neurons, primary concept neurons, and associated neurons, which perform different functions. External stimulus is categorized into order independent stimulus (OIDS) and order dependent stimulus (ODS). Note that the figure only takes vision (OIDS) and audition (ODS) for example, other sensations can be also involved in the structure. In the following, we first give an overview of the network structure. Then, we give the learning process of the PCN.

Network Structure

Primary Sensory Area (PSA): The PSA includes "feature neurons", which respond to particular features, e.g., color feature, shape feature, or syllable feature, see the bottom layer in Fig. 3. Feature neurons which respond to the same type feature are located in the same area α and we use set N^{F_α} to store them. As mentioned above, α can be the color area, the shape area, or the syllable area. $N_i^{F_\alpha}$ is used to denote feature neuron i and $N_i^{F_\alpha} \triangleq \{\boldsymbol{w}_i, \sigma_i\}$, where $\boldsymbol{w}_i$ and σ_i represent the weights and the activation times of $N_i^{F_\alpha}$ respectively. The activating domains of $N_i^{F_\alpha}$ are defined as

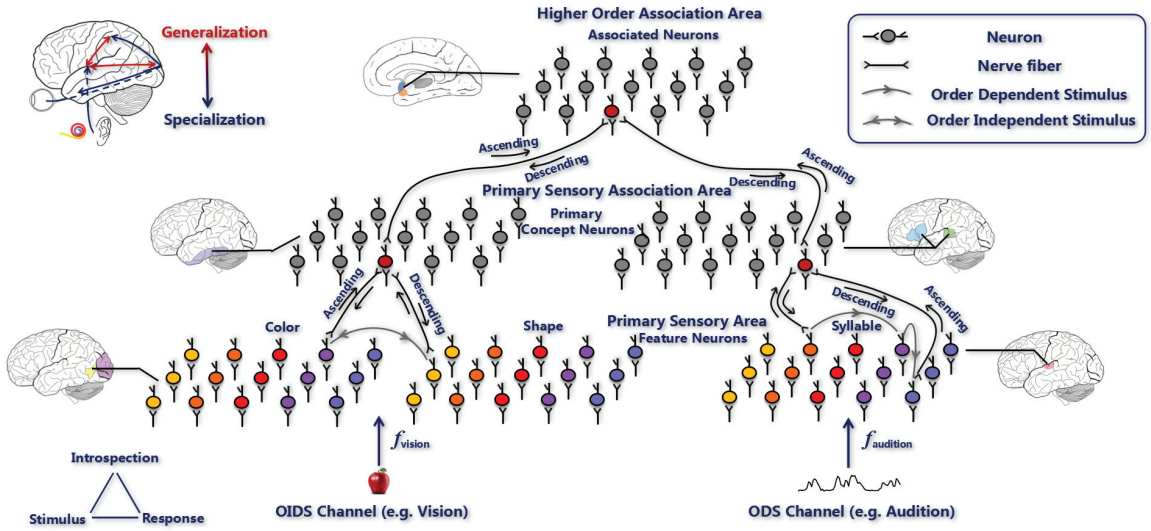

Figure 3: Structure of the Perception Coordination Network.

follows,

$$\boldsymbol{ADs}(N_i^{F_\alpha}) \triangleq \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{w}_i\| \leq \theta\} \tag{1}$$

Where θ is a parameter that controls the response range of the ADs.

Primary Sensory Association Area (SAA): The SAA includes "primary concept neurons", which connect the feature neurons to represent a unimodal concept, e.g., to form visual concepts by connecting shape and color feature neurons, to form auditory concepts such as words by connecting syllable feature neurons, see the middle layer in Fig. 3. Concept neurons in a sensory association area β are stored in set N^{C_β}, β can be the visual association area v, auditory association area a or other sensory association area. $N_i^{C_\beta}$ is used to denote concept neuron i in area β. The connection between a concept neuron $N_i^{C_\beta}$ and a feature neuron $N_j^{F_\alpha}$ is defined as follows,

$$c_{(i,j)} \triangleq \{N_i^{C_\beta}, N_j^{F_\alpha}, \rho_{(i,j)}\} \tag{2}$$

Where $\rho_{(i,j)}$ is the cumulative activation times of the connection. And α represents the feature areas that the sensory association area β connects. The activating domains of $N_i^{C_\beta}$ is the activations of the feature neurons that it connects. The external stimulus is divided into two types, which are Type I. Order InDependent Stimulus (OIDS); and Type II. Order Dependent Stimulus (ODS). For example, different visual features of an object belong to the OIDS. Because different orders of visual features, such as color and shape, do not affect the activation of their corresponding visual concept; The syllables contained in a voice wave belong to the ODS. Because different orders of the same group of syllables may refer to different concepts such as the words "[bō luó]" and "[luó bō]". Correspondingly, two kinds of activating domains of the concept neurons are defined as follows,

$$\boldsymbol{ADs}(N_i^{C_\beta}) \triangleq \begin{cases} \overrightarrow{(N_1^{F_\alpha}, N_2^{F_\alpha}, ..., N_n^{F_\alpha})}, & \text{ODS} \\ (N_1^{F_\alpha}, N_2^{F_\alpha}, ..., N_n^{F_\alpha}), & \text{OIDS} \end{cases} \tag{3}$$

Where feature neurons $N_1^{F_\alpha}, N_2^{F_\alpha}, ..., N_n^{F_\alpha}$ connect the concept neuron $N_i^{C_\beta}$. The arrow over the vector means $N_i^{C_\beta}$ can be fired only by the firing of $N_1^{F_\alpha}, N_2^{F_\alpha}, ..., N_n^{F_\alpha}$ through this direction.

Higher Order Association Area (HAA): The HAA includes "associated neurons", which connect primary concept neurons in different SAA, see the top layer in Fig. 3. Associated neurons are stored in set N^A and N_i^A is used to denote associated neuron i. The connection between a concept neuron $N_m^{C_\beta}$ and another concept neuron $N_n^{C_\beta}$ through N_i^A is defined as follows,

$$c_{(m,i,n)} \triangleq \{N_m^{C_\beta}, N_i^A, N_n^{C_\beta}, \rho_{(m,i,n)}\} \tag{4}$$

Where $N_m^{C_\beta}$ and $N_n^{C_\beta}$ come from different PSA, e.g., $N_m^{C_\beta}$ can be a visual concept neuron and $N_n^{C_\beta}$ can be an auditory concept neuron, then a view and a name of an object are associated. $\rho_{(m,i,n)}$ is the cumulative activation times of the connection. The activating domains of N_i^A is the primary concept neurons that N_i^A connects, i.e.,

$$\boldsymbol{ADs}(N_i^A) \triangleq \{N_1^{C_\beta}, N_2^{C_\beta}, ..., N_n^{C_\beta}\} \tag{5}$$

Note that the ADs of the associated neuron is a set of primary concept neurons, it means any concept neurons in the set can activate it, i.e., the associated neuron has a multimodality activation mode.

Learning Process

The overall learning process of the PCN is as follows: When a pair of sample comes, e.g., a pair of visual and auditory input, the PSA will first extract features. Then competitive learning is conducted and firing neurons will transmit activated signals to the SAA. After the SAA gets the signals, competition among concept neurons will be executed to activate some concept neurons and the activated signals are

transmitted to the HAA, meanwhile the SAA is updated. When the HAA receives the signals, an unconscious impulse process will be triggered firstly, which uses one channel's signal to wake its corresponding concepts in other channels. Then an introspection process is conducted, which aims to check the consistency between the current input pair and the learned knowledge from the past input pairs. Finally, the SAA will be updated according to the results of the introspection process. Now the process for the current input finishes, and PCN will go to the next input. In the following, we use a pair of vision (image of an object) and audition (name of an object) input to describe the method in detail.

Primary Sensory Area: Currently, the normalized Fourier descriptors $\boldsymbol{d}$ of the object's boundary and the color histogram $\boldsymbol{g}$ of the object's area are used for visual features. The Mel-Frequency Cepstral Coefficients $\boldsymbol{m}$ of a syllable are used for the auditory features, where short-time energy and short-time zero crossing are used to extract syllables in an input word. After feature extraction, competitive learning is executed. First, a winner neuron in each area α is found as follows,

$$N_f^{F_\alpha} = \underset{N_i^{F_\alpha} \in N^{F_\alpha}}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_i\|, \quad \text{where } \alpha \in \{b, c, s\} \tag{6}$$

Where $\boldsymbol{x}$ is the feature vector such as $\boldsymbol{d}$, $\boldsymbol{g}$ and $\boldsymbol{m}$. $\boldsymbol{w}_i$ is the weights of feature neuron i. $\|\cdot\|$ is the distance function, for vision Euclidean distance is used and for audition dynamic time warping is used. The superscribe F_b, F_c and F_s represent the shape, color and syllable PSA respectively. If $\boldsymbol{x}$ belongs to the ADs of neuron $N_f^{F_\alpha}$, i.e., $\|\boldsymbol{x} - \boldsymbol{w}_f\| \leq \theta$, the weights of $N_f^{F_\alpha}$ are updated as follows,

$$\sigma_f = \sigma_f + 1; \;\; \boldsymbol{w}_f = \boldsymbol{w}_f + (\boldsymbol{x} - \boldsymbol{w}_f)/\sigma_f \tag{7}$$

[1]

Then $N_f^{F_\alpha}$ is activated. If $\boldsymbol{x}$ does not belong to the ADs of $N_f^{F_\alpha}$, a new neuron $N_{\text{new}}^{F_\alpha}$ will be created for $\boldsymbol{x}$, i.e.,

$$N_{\text{new}}^{F_\alpha} = \{\boldsymbol{x}, 1\}, \quad \text{where } \alpha \in \{b, c, s\} \tag{8}$$

Then $N_{\text{new}}^{F_\alpha}$ is activated. Finally, the activated signals of the firing feature neurons are transmitted to the their corresponding SAA: For vision, $(N_{f_b}^{F_b}, N_{f_c}^{F_c}) \xrightarrow{\text{Signal}}$ Visual SAA, where $N_{f_b}^{F_b}$ and $N_{f_c}^{F_c}$ represent the activated neurons in the shape area and color area respectively. For audition, $\overrightarrow{(N_{f_1}^{F_s}, N_{f_2}^{F_s}, ..., N_{f_k}^{F_s})} \xrightarrow{\text{Signal}}$ Auditory SAA, where $N_{f_k}^{F_s}$ is the k-th firing syllable feature neuron, and k is the syllable number of the input word.

Primary Sensory Association Area: In the SAA, the received signals are firstly checked whether equal to any concept neuron's ADs, i.e., to solve the following equations,

$$\boldsymbol{ADs}(N_i^{C_v}) = (N_{f_b}^{F_b}, N_{f_c}^{F_c}), \qquad N_i^{C_v} \in N^{C_v} \tag{9}$$

$$\boldsymbol{ADs}(N_i^{C_a}) = \overrightarrow{(N_{f_1}^{F_s}, N_{f_2}^{F_s}, ..., N_{f_k}^{F_s})}, \quad N_i^{C_a} \in N^{C_a} \tag{10}$$

Where N^{C_v} and N^{C_a} represent the visual concept neuron set and auditory concept neuron set respectively. If some concept neurons $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ are found by Eq. 9 and Eq. 10 respectively, then $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ are activated and the activation times of the connections between the concept neurons and the feature neurons will increase, i.e., $\rho_{(f_v,f_b)} = \rho_{(f_v,f_b)}+1$, $\rho_{(f_v,f_c)} = \rho_{(f_v,f_c)}+1$, and $\rho_{(f_a,f_i)} = \rho_{(f_a,f_i)}+1$, where $1 \leq i \leq k$. If no concept neuron is found according to Eq. 9 or Eq. 10, new concept neuron $N_{\text{new}}^{C_v}$ or $N_{\text{new}}^{C_a}$ will be created as follows,

$$\text{For vision:} \quad \boldsymbol{ADs}(N_{\text{new}}^{C_v}) = (N_{f_b}^{F_b}, N_{f_c}^{F_c}) \tag{11}$$

$$\text{For audition:} \quad \boldsymbol{ADs}(N_{\text{new}}^{C_a}) = \overrightarrow{(N_{f_1}^{F_s}, N_{f_2}^{F_s}, ..., N_{f_k}^{F_s})} \tag{12}$$

And connections of the new neuron are created as follows,

$$c_{(f_v,f_b)} = \{N_{f_v}^{C_v}, N_{f_b}^{F_b}, 1\}; \quad c_{(f_v,f_c)} = \{N_{f_v}^{C_v}, N_{f_c}^{F_c}, 1\} \tag{13}$$

$$c_{(f_a,f_i)} = \{N_{f_a}^{C_a}, N_{f_i}^{F_s}, 1\}, \quad \text{where } 1 \leq i \leq k \tag{14}$$

Then the new established neurons are activated. Finally, the activated signals are transmitted to the HAA, i.e., $(N_{f_v}^{C_v}, N_{f_a}^{C_a}) \xrightarrow{\text{Signal}}$ HAA, where $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ represent the activated neurons in visual and auditory SAA.

Higher Order Association Area: When HAA receives signal $(N_{f_v}^{C_v}, N_{f_a}^{C_a})$, an unconscious impulse process will be triggered firstly.

Taking the vision as the start point, to find the associated neuron which connects $N_{f_v}^{C_v}$ by solving the following equation,

$$N_{f_v}^{C_v} \in \boldsymbol{ADs}(N_i^A), \quad \text{where } N_i^A \in N^A \tag{15}$$

Assuming associated neuron N_{vf}^A is found, then the neuron is activated, and the primary auditory concept neurons that connect N_{vf}^A will be unconsciously activated, which are

$$N_u^{C_a} = \{N_i^{C_a} \mid N_i^{C_a} \in \boldsymbol{ADs}(N_{vf}^A)\} \tag{16}$$

Where set $N_u^{C_a}$ is used to store these primary auditory concept neurons.

Meanwhile, taking the audition as the start point, find the associated neuron which connects $N_{f_a}^{C_a}$ as follows,

$$N_{f_a}^{C_a} \in \boldsymbol{ADs}(N_i^A), \quad \text{where } N_i^A \in N^A \tag{17}$$

Assuming associated neuron N_{af}^A is found, and the primary visual concept neurons that connect N_{af}^A will be unconsciously activated, which are

$$N_u^{C_v} = \{N_i^{C_v} \mid N_i^{C_v} \in \boldsymbol{ADs}(N_{af}^A)\} \tag{18}$$

Where set $N_u^{C_v}$ is used to store these primary visual concept neurons.

After the unconscious impulse process, an introspection process will be executed. This process is divided into four conditions based on the results of the unconscious impulse.

(a) If some associated neuron N_{af}^A is found through Eq. 17 and no associated neuron is found through Eq. 15, it means that the view of the object is new to PCN, but the voice is met by PCN which may be used to call some other views. The current visual input should be like the views symbolized by the primary concept neurons in set $N_u^{C_v}$ according to the current input voice. Therefore, PCN will enquire:

[1]The dimension of the weights of different syllable feature neurons is usually not same, so we do not update them using this way.

"The name $N_{f_a}^{C_a}$ can also represent this view?" Then an external signal γ from users is needed to help making judgment. When $\gamma = 1$, which means the new view is also called $N_{f_a}^{C_a}$, then $N_{f_v}^{C_v}$ is added to the ADs of N_{af}^{A},

$$\boldsymbol{ADs}(N_{af}^{A}) = \boldsymbol{ADs}(N_{af}^{A}) \cup N_{f_v}^{C_v} \tag{19}$$

And the connection between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ will be created as follows,

$$c_{(f_v,af,f_a)} = \{N_{f_v}^{C_v}, N_{af}^{A}, N_{f_a}^{C_a}, 1\} \tag{20}$$

When $\gamma = 0$, it means the current view is not called $N_{f_a}^{C_a}$. Then a new associated neuron will be created to connect $N_{f_v}^{C_v}$ but without connecting any name, i.e.,

$$\boldsymbol{ADs}(N_{\text{new}}^{A}) = \{N_{f_v}^{C_v}\} \tag{21}$$

(b) If some associated neuron N_{vf}^{A} is found through Eq. 15 and no associated neuron is found through Eq. 17, it means the current input voice is new to PCN, but the view of the current object is met before. The object should be called with the voices symbolized by the primary auditory concept neurons in set $N_u^{C_a}$ if $N_u^{C_a}$ is not empty. Then PCN will enquire: "This object is called $N_u^{C_a}$ previously. Is it also named $N_{f_a}^{C_a}$?" Then an external signal γ is needed. When $\gamma = 1$, which means the object is also called $N_{f_a}^{C_a}$, PCN will put $N_{f_a}^{C_a}$ into the ADs of N_{vf}^{A},

$$\boldsymbol{ADs}(N_{vf}^{A}) = \boldsymbol{ADs}(N_{vf}^{A}) \cup N_{f_a}^{C_a} \tag{22}$$

Meanwhile, the connection between this new combination will be created as follows,

$$c_{(f_v,vf,f_a)} = \{N_{f_v}^{C_v}, N_{vf}^{A}, N_{f_a}^{C_a}, 1\} \tag{23}$$

When $\gamma = 0$, which means the object is not called $N_{f_a}^{C_a}$, the primary auditory concept neuron $N_{f_a}^{C_a}$ will be rejected by the network.

(c) If some associated neuron N_{vf}^{A} and N_{af}^{A} are found through Eq. 15 and Eq. 17, it means PCN has seen the object and heard the voice. The coherence between current vision and audition input should be checked firstly. If $N_{vf}^{A} = N_{af}^{A}$, it means $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ activate the same associated neuron. The current input pair is consistent with previous pattern combinations. Then the activity of the connection between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ through the associated neuron N_{vf}^{A} will be increased to strengthen the association, i.e., $\rho_{(f_v,vf,f_a)} = \rho_{(f_v,vf,f_a)} + 1$. If $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ activate different associated neurons, i.e., $N_{vf}^{A} \neq N_{af}^{A}$, it means the current combinations between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ is inconsistent with some previous combinations. PCN will enquire to the user to get an answer γ. When $\gamma = 1$, which means the view of the object and the name is an expected combination, then $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ are added to the ADs of N_{af}^{A} and N_{vf}^{A},

$$\begin{aligned} \boldsymbol{ADs}(N_{vf}^{A}) &= \boldsymbol{ADs}(N_{vf}^{A}) \cup N_{f_a}^{C_a} \\ \boldsymbol{ADs}(N_{af}^{A}) &= \boldsymbol{ADs}(N_{af}^{A}) \cup N_{f_v}^{C_v} \end{aligned} \tag{24}$$

And the connections between the new combinations will be created as follows,

$$\begin{aligned} c_{(f_v,vf,f_a)} &= \{N_{f_v}^{C_v}, N_{vf}^{A}, N_{f_a}^{C_a}, 1\} \\ c_{(f_v,af,f_a)} &= \{N_{f_v}^{C_v}, N_{af}^{A}, N_{f_a}^{C_a}, 1\} \end{aligned} \tag{25}$$

When $\gamma = 0$, it means the combination is not expected, and no operation will be done by PCN.

(d) If no associated neurons are found through Eq. 15 and Eq. 17, it means the combination of $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ has not been met before. And a new associated neuron N_{new}^{A} will be created to associate this combination as follows,

$$\boldsymbol{ADs}(N_{\text{new}}^{A}) = \{N_{f_v}^{C_v}, N_{f_a}^{C_a}\} \tag{26}$$

And the connection between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ through N_{new}^{A} will be created as follows,

$$c_{(f_v,\text{new},f_a)} = \{N_{f_v}^{C_v}, N_{\text{new}}^{A}, N_{f_a}^{C_a}, 1\} \tag{27}$$

After the introspection process, the learning for current input finishes. PCN will go to the next input. Algorithm 1 summarizes the learning process.

Algorithm 1 Perception Coordination Network

Initialize: Set the value of parameter θ.

1: Receive a pair of image (OIDS) and name (ODS) of an object.
2: PSA: Extract features from image and voice. Execute competitive learning for the feature, formula (6) to (8).
3: SAA: Execute concept learning procedure, formula (9) to (14).
4: HAA: Execute unconscious impulse process an introspection process, formula (15) to (27).
5: Waiting for the next input pair and go to Step 1.

Experiments

The concept acquisition and binding among vision, audition, and gustation are conducted. 20 common fruits and foods are used. Because we do not have real taste data, an artificial taste data set is designed. The data format is a 6-dimensional vector which is (sweet, sour, salt, bitter, umami, hot). The value of each attribute is in the range [0, 1]. For example, we design the taste of apple as follows, the values of sweet and sour are uniformly distributed in the range [0.5, 0.6] and [0, 0.1] respectively, other attributes are 0. During the learning experiment, we let PCN learn objects' views, names and tastes. Firstly, an object is put in front of a camera. Then the audition program is started and the name of the object is pronounced by the user. At the same time, the vision program captures the images of the object. When the pronunciation is finished, the audition program is closed. And the current round of learning of the object's name finishes. Then we go to the next round. When all objects' image-name learning are finished, we give the taste data of each object to PCN with the image of the object simultaneously to make PCN learn the object's taste[2].

[2]Algorithm of the gustatory module is similar with that of the visual module, because they both are OIDS channel.

Figure 4: Objects used in the experiment.

We conduct the experiment in two environments, which are (1) Closed environment; and (2) Open-ended environment (for the stability-plasticity dilemma). In the closed environment, object is randomly chosen from the 20 objects in each round of learning. In the open-ended environment, we first let the methods learn 10 objects. In each round of learning, object is randomly chosen. After that, we give the methods the remaining 10 "new" objects in the second learning period. Similarly, in each round of learning, object is also randomly chosen. We conduct the experiment 30 times in both closed and open-ended environment, each time contains 352 rounds of image-name and image-taste learning for 20 objects. Fig. 4 shows examples of the objects. There are some voices with the same syllables but in different order referring to different objects (Order Dependent Stimulus), e.g., "[bō luó]" and "[luó bō]". There are also different voices referring to the same object, e.g., "[píng guǒ]" and "[zhì huì guǒ]".

We compare PCN with the IKR1 system (He, Kojima, and Hasegawa 2007), MMSOM (Jantvik, Gustafsson, and Papliński 2011) and GHF-ART (Meng, Tan, and Xu 2014). The MMSOM deals with visual-audio fusion, however, it is not an incremental method which cannot handle the stability-plasticity dilemma. The IKR1 system and GHF-ART are incremental method. They handle word grounding and multi-modal feature fusion in an online way. Because IKR1 system and MMSOM only give a vision-audition bimodal system, we only let them learn visual and auditory data. GHF-ART does not give a fuzzy operation for auditory information with different dimensions, we let GHF-ART learn visual and gustatory data. For visual, auditory and gustatory module, parameter θ of PCN is set as 1/4 times the 2-norm of the weight vector of the feature neurons, 200 and 0.05 respectively. The parameters of the comparison methods are set following the authors' suggestions.

Learning Results

During the 60 times experiments, PCN learns 71 to 77 shape feature neurons, 33 to 39 color feature neurons, 71 to 77 visual concept neurons; 221 to 228 syllable neurons, 130 to 137 auditory concept neurons; 34 to 36 basic flavor neurons, 52 to 59 gustatory concept neurons; and 60 to 62 associated neurons. An average of 89 questions are asked during learning. About 85 questions are related to different names of the same object, user's answer is positive. The other about 4 questions are caused by visual and auditory erroneous judgements. Thus, user's answer is negative.

Fig. 5(a) gives some examples of the learning results. It can be found that PCN correctly acquires these concepts and properly binds them. For a more detailed observation, Fig. 5(b) and Fig. 5(c) gives two examples of learned structure of the concept apple and pear. The associated neuron connects the concept neurons in visual, auditory, and gustatory areas. The visual concept neuron connects the shape and color feature neurons. The auditory concept neuron connects a series of syllable feature neurons. The gustatory concept neuron connects basic flavor feature neurons. Probabilities of connections between visual concept neuron and each auditory concept neuron through the associated neuron in Fig. 5(b) are both 0.5, because the we say "[píng guǒ]" and "[zhì huì guǒ]" with equal probability during experiments. Other probabilities of the connections are gained in a similar way.

Interestingly, the name and the taste of the objects are linked together automatically through the associated neuron, while the name and taste data were not given to the system simultaneously during learning. It means PCN is able to make sensory channels coordinate automatically. This is an advantage of PCN's network structure.

Testing Results

To test the model learned by PCN, we use one kind of sensory input to recall other two kinds of sensory output, e.g., the visual input recalls the auditory and gustatory output. Because IKR1 and MMSOM learn a vision-audition bimodal result, we only do vision and audition recall each other. GHF-ART learns visual and gustatory data, we do vision and gustation recalling. Testing is conducted after each time of learning. In each time of testing, 528 rounds of recalling (176 rounds for each type of recall) are executed with different data from the learning experiment. Table 1 shows that PCN recalls "memories" with a much higher accuracy than other methods. The accuracy of IKR1 and MMSOM is very unstable, which has a gap about 5% and 15% between two learning environments. We find that MMSOM cannot learn new objects after a period of learning. Thus, the drop of the accuracy mainly dues to the recognition of the latter 10 "new" objects. IKR1 usually "forgets" previously learned objects when the 10 "new" objects come. Thus, the drop of the accuracy mainly dues to the recognition of the previous 10 "old" objects. The accuracy of PCN is much higher and more stable in both environments.

Experiment Summary

Compared with PCN, IKR1, MMSOM and GHF-ART lack a layer with a function of PCN's SAA. It makes them cannot combine features freely, such as assigning different orders of the same syllables to different objects (because they take the word not the syllable as a unit), which is successfully achieved by PCN. This means PCN reuses features in a much better way. The unconscious impulse and introspection process in the HAA make PCN can communicate with users to check whether there is any contradiction, then update the network according to users' answers, while IKR1,

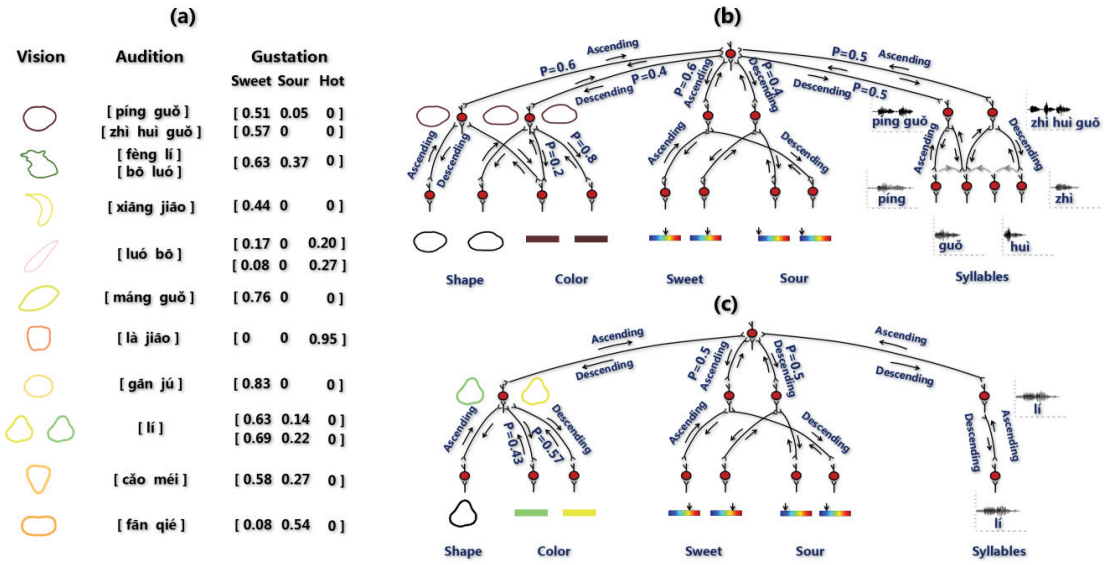

Figure 5: Learning results. (a) Ten examples of the associations of learned visual, auditory, and gustatory concepts. (b) and (c) Example of the network structure of the concept apple and pear including view, taste and name. Icons next to the neurons represent the objects to which the neurons maximally respond.

Table 1: Statistical results of the testing experiments (mean+std). N/A represents the testing condition is Not Applicable for the method. Significant decrease of the accuracy is marked by ↓. The best results are in bold. V: Vision, A: Audition, G: Gustation.

		V recalls Other	A recalls Other	G recalls Other
Closed environment	IKR1	69.10±4.17% ↓	63.48±3.52% ↓	N/A
	GHF-ART	81.92±3.16%	N/A	85.11±2.17%
	MMSON	76.31±4.25%	65.41±5.27%	N/A
	PCN	**83.98±3.02%**	**86.24±3.30%**	**90.35±2.74%**
Open-ended environment	IKR1	75.71±4.22%	70.57±5.25%	N/A
	GHF-ART	82.07±3.18%	N/A	86.40±2.01%
	MMSON	63.64±4.04% ↓	50.11±3.53% ↓	N/A
	PCN	**84.83±3.26%**	**88.07±2.78%**	**92.09±2.16%**

MMSOM and GHF-ART are lack of such a function in their association areas. Through building new connections and neurons, PCN can handle the stability-plasticity dilemma very well, while IKR1 and MMSOM cannot. Overall, the structure of PCN is much more flexible and has a much stronger plasticity than other methods.

Conclusion

We propose a Perception Coordination Network (PCN) for online multi-modal concept acquisition and binding. Three types of neurons, which are in charge of different functions, are modeled and embed to a hierarchical structure network. Meanwhile, PCN is able to learn new concepts and bindings through creating new connections between neurons. Experimental results in both closed environment and open-ended environment demonstrate that PCN works effectively.

In the future, horizontal connections in each area will be developed. The framework will be combined with other features to make a more robust perception. Finally, as Simon's analysis that the hierarchic system is helpful to evolution (Simon 1962), we will enable PCN a sense augment ability to solve the perception evolution problem (Xing, Shen, and Zhao 2013; 2015; 2016; Xing 2016).

Acknowledgments

This work is supported by the 973 Program of China (2015CB351705) and the National Science Foundation of China under Grant Nos. (61703002, 61472002, 61373130, 61375064, 61373001). The code can be downloaded at https://cs.nju.edu.cn/rinc/publish/code/PCN.rar

References

Araki, T.; Nakamura, T.; Nagai, T.; Funakoshi, K.; Nakano, M.; and Iwahashi, N. 2011. Autonomous acquisition of multimodal information for online object concept formation by robots. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1540–1547.

Carpenter, G. A., and Grossberg, S. 1988. The art of adaptive pattern recognition by a self-organizing neural network. *IEEE Computer* 21(3):77–88.

Chan, A. M.; Dykstra, A. R.; Jayaram, V.; Leonard, M. K.; Travis, K. E.; Gygi, B.; Baker, J. M.; Eskandar, E.; Hochberg, L. R.; Halgren, E.; and Cash, S. S. 2014. Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex* 24(10):2679–2693.

de Araujo, I. E., and Simon, S. A. 2009. The gustatory cortex and multisensory integration. *International Journal of Obesity* 33:S34–S43.

Fuster, J. M. 1997. Network memory. *TRENDS in Neurosciences* 20(10):451–459.

Gersho, A., and Gray, R. M. 1992. *Vector Quantization and Signal Compression.* Boston: MA: Kluwer.

He, X.; Kojima, R.; and Hasegawa, O. 2007. Developmental word grounding through a growing neural network with a humanoid robot. *IEEE Transactions on System, Man, and Cybernetics – Part B: Cybernetics* 37(2):451–462.

Hegdé, J., and Essen, D. C. V. 2000. Selectivity for complex shapes in primate visual area v2. *Journal of Neuroscience* 20(5):1–6.

Jantvik, T.; Gustafsson, L.; and Papliński, A. P. 2011. A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration. *Neural Computation* 23(23):2101–2139.

Kobatake, E., and Tanaka, K. 1994. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology* 71(3):856–867.

Livingstone, M., and Hubel, D. 1988. Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science* 240(4853):740–749.

Meng, L.; Tan, A.-H.; and Xu, D. 2014. Semi-supervised heterogeneous fusion for multimedia data co-clustering. *IEEE Transactions on Knowledge and Data Engineering* 26(9):2293–2306.

Mesgarani, N.; Cheung, C.; Johnson, K.; and Chang, E. F. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science* 343(6174):1006–1010.

Nakamura, T.; Nagai, T.; and Iwahashi, N. 2011. Bag of multimodal lda models for concept formation. In *Proceedings of IEEE International Conference on Robotics and Automation*, 6233–6238.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of International Conference on Machine Learning*, 689–696.

Parde, N.; Hair, A.; Papakostas, M.; Tsiakas, K.; Dagioglou, M.; Karkaletsis, V.; and Nielsen, R. D. 2015. Grounding the meaning of words through vision and interactive gameplay. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1895–1901.

Quiroga, R. Q. 2012. Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience* 13(8):587–597.

Schyns, P. G. 1991. A modular neural network model of concept acquisition. *Cognitive Science* 15(4):461–508.

Shen, F., and Hasegawa, O. 2006. An incremental network for online unsupervised classification and topology learning. *Neural Networks* 19(1):90–106.

Simon, H. A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106(6):467–482.

Srivastava, N., and Salakhutdinov, R. 2014. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research* 15:2949–2980.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. J. 2016. Learning multi-modal grounded linguistic semantics by playing "i spy". In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3477–3483.

Xing, Y.-L.; Shen, F.-R.; and Zhao, J.-X. 2013. A perception evolution network for unsupervised fast incremental learning. In *Proceedings of the International Joint Conference on Neural Networks*, 1–8.

Xing, Y.-L.; Shen, F.-R.; and Zhao, J.-X. 2015. Perception evolution network: Adapting to the emergence of new sensory receptor. In *Proceedings of the International Conference on Artificial Intelligence*, 3967–3973.

Xing, Y.-L.; Shen, F.-R.; and Zhao, J.-X. 2016. Perception evolution network based on cognition deepening model – adapting to the emergence of new sensory receptor. *IEEE Transactions on Neural Networks and Learning Systems* 27(3):607–620.

Xing, Y.-L. 2016. Perception, coordination and evolution – intelligence lies in structure and movement. *Ph.D. Dissertation*.

Yamauchi, K.; Oota, M.; and Ishii, N. 1999. A self-supervised learning system for pattern recognition by sensory integration. *Neural Networks* 12(10):1347–1358.