Perception Coordination Network: A Neuro Framework for Multimodal Concept Acquisition and Binding

You-Lu Xing, Xiao-Feng Shi, Fu-Rao Shen, Jin-Xi Zhao, Jing-Xin Pan, and Ah-Hwee Tan, *Senior Member, IEEE*

***Abstract*— To simulate the concept acquisition and binding of different senses in the brain, a biologically inspired neural network model named perception coordination network (PCN) is proposed. It is a hierarchical structure, which is functionally divided into the primary sensory area (PSA), the primary sensory association area (SAA), and the higher order association area (HAA). The PSA contains feature neurons which respond to many elementary features, e.g., colors, shapes, syllables, and basic flavors. The SAA contains primary concept neurons which combine the elementary features in the PSA to represent unimodal concept of objects, e.g., the image of an apple, the Chinese word "[píng guǒ]" which names the apple, and the taste of the apple. The HAA contains associated neurons which connect the primary concept neurons of several PSA, e.g., connects the image, the taste, and the name of an apple. It means that the associated neurons have a multimodal response mode. Therefore, this area executes multisensory integration. PCN is an online incremental learning system, it is able to continuously acquire and bind multimodality concepts in an online way. The experimental results suggest that PCN is able to handle the multimodal concept acquisition and binding effectively.**

***Index Terms*— Concept acquisition and binding, multimodal learning, online incremental learning, perception coordination network (PCN), unsupervised learning.**

I. Introduction

THE brains of animals and organisms continuously receive signals from multiple modalities via different types of sensory receptors. Also, the nerve impulses generated by these signals are transmitted on the network of the brain to form perceptions; meanwhile, the brain network itself is updated to strengthen, acquire, or bind concepts.

Manuscript received October 21, 2017; revised May 10, 2018 and July 5, 2018; accepted July 25, 2018. Date of publication August 21, 2018; date of current version March 18, 2019. This work was supported in part by the National Science Foundation of China under Grant 61703002 and in part by the Jiangsu NSF of China under Grant BK20171344. *(You-Lu Xing and Xiao-Feng Shi contributed equally to this work.) (Corresponding author: You-Lu Xing.)*

Y.-L. Xing is with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: youluxing@sina.com).

X.-F. Shi, F.-R. Shen, and J.-X. Zhao are with the National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210046, China (e-mail: cell000feng@gmail.com; frshen@nju.edu.cn; jxzhao@nju.edu.cn).

J.-X. Pan is with the Medical School, Nanjing University, Nanjing 210046, China (e-mail: jackiejackpan@126.com).

A.-H. Tan is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: asahtan@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2861680

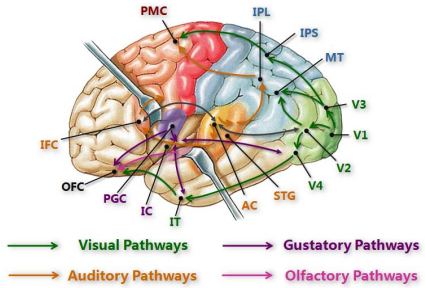


Fig. 1. Visual, auditory, gustatory, and olfactory pathways in the brain. Different sensations interact through the areas where different sensory pathways converge, such as the entorhinal cortex and orbitofrontal cortex. The pathways are summarized and deduced from [2]– [18].

As mentioned in [1], our brain uses different types of sensory information, including vision, touch, and audition, to perceive the external environment. All these different types of sensory information are efficiently merged in the brain to form a coherent and robust percept. Therefore, the coordination among different senses is essential for human cognition. Fig. 1 illustrates 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 sensations become complicated through its pathway. Different sensations interact with each other through the areas where different sensory pathways converge.

Fuster [19] gave an example of sensory association between vision and touch at the cell level. The explanation is based on the Hebbian theory [20], which is summarized as cells that fire together and wire together [21], [22]. For example, as shown in Fig. 2(4) and (5), when a visual and a tactile signal stimulate the network synchronously, a cell assembly will be formed by the facilitated synapses to associate the visual and tactile sense. The example is relatively simple, but very enlightening.

These studies [1]–[18], [19] make us to think about the brain function of multimodal concept acquisition and binding of different senses. This is essential for human cognition. Creating a computational model for such a brain function is indispensable for artificial intelligence. Thus, here we pose the following problem:

Problem. How to build a neuro framework to imitate the concept acquisition and binding of different senses in the brain?

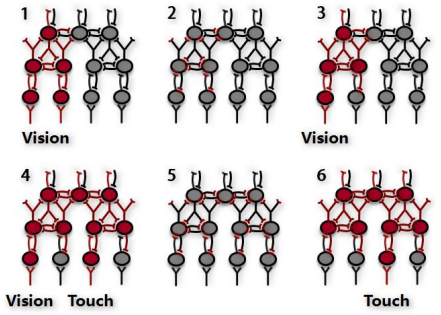


Fig. 2. Sensory association. (1) Two visual inputs coincide in time. (2) Passive long-term memory formed by the facilitated synapses generated during step 1, marked in red. (3) One of the visual inputs activates the subnetwork in step 2. (4) Visual and a tactile input coincide. (5) Bimodal network of long-term memory formed by the facilitated synapses generated during step 4, marked in red. (6) Tactile stimulus activates the bimodal network. The figure is modified from [19].

II. Related Work

A. Some Related Physiological Studies

The brain network can be seen as a hierarchical structure [3], [23], which is functionally modularized and specified, which means that different areas of the brain are usually in charge of different functions. Within the hierarchy, the higher level processes will "synthesize" the functions at lower levels.

For example, in the vision channel, there is a functional segregation of primary visual system, which includes color, depth, movement, and form perception [25]. Also, in the "what" pathway of the visual system, Livingstone and Hubel [25] find that cells in V1 are tuned to some simple visual properties, such as particular orientation, color, and spatial frequency. Beyond V1, cells in V2 respond to not only what V1 cells are tuned to but also intermediate complex shapes [26]. Next, cells in V4 and posterior inferotemporal (IT) cortex are tuned to complex shapes, combinations of a shape and texture, and combinations of a shape and color. Finally, at the top level, anterior IT cells are tuned to particular complex object features [27].

In the audition channel, there are multiple levels of computation and representation mapping acoustic speech inputs onto conceptual and semantic representations [28] in the "sound to meaning" pathway. First, cells in the primary auditory cortex (A1) are organized according to the frequency of sound to which they respond best [29]. Next, in the superior temporal gyrus (STG), Mesgarani *et al.* [30] find that the local region of the STG shows invariant selectivity to single phonemes, e.g., plosive phonemes, sibilant fricatives, different types of vowels, and nasals. Then, cells in the anterior STG are tuned to particular spoken words [12]. Lexical, semantic, and grammatical linkages, which are the least understood, include a much broader network, involving most of the temporal lobe and the inferior frontal lobe [28], e.g., left anterior temporal and temporal-parietal areas respond more strongly to sentences than to randomly ordered lists of words [31].

As mentioned by Simmons *et al.* [11], increasing research indicates that concepts are represented as distributed circuits of property information across the brain's modality-specific areas. Quiroga *et al.* [16] found that single cells in the human medial temporal lobe responded selectively to representations of the same individual across different sensory modalities, including vision and audition, and such neuronal representations could be generated within less than a day or two [17]. We can name this type of cell as a multimodality cell. Simmons *et al.* [11] implies that the multimodality cell is vital for concept binding. In [2], a region of the human insular taste cortex, as defined by its response to a prototypical taste stimulus (sucrose), which is activated by olfactory stimuli, is found. This means that coordination among a different sensory perception works all the time.

B. Some Related Computational Models

Many computational models for sensory integration and multimodal concept acquisition have been proposed during the past years.

Modular neural networks (MNNs) are inspired by the modular structure of the brain, where different modules perform different functions [32]. MNNs are very useful in sensory integration and information fusion. In [33], an MNN for concept acquisition is proposed, where the self-organizing maps (SOMs) [34] are used to build concept prototype and a brain-state-in-a-box [35] is used to associate concept names, i.e., lexicon, to the output of the SOM, i.e., concept. In [36], several neural network modules, which consist of forward and backward parts, are fused by some integrating units. The network is able to learn categories of objects by integrating information from several sensors, such as the acoustics of Japanese vowels and corresponding visual shapes of the mouth.

In [37], a bag of multimodal latent Dirichlet allocation (LDA) is introduced for sensory integration. The bag includes object categories LDA, color categories LDA, and haptic categories LDA. In [38], 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 [39], a bimodal deep belief network (DBN) is trained to learn a shared representation of visual and auditory input. First, a top restricted Boltzmann machine over the pretrained 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 [40]–[42], which learn joint representation between text features and image features.

In [43] and [44], the meanings of words are grounded in visual features by conversations between users and a robot. An initial learning phase is needed which leads to the methods that cannot deal with words grounding in a totally online incremental way.

The methods above do not focus on learning new concepts or new bindings in an online incremental way. However, new concepts and bindings always occur in the real world. Thus, a better learning system should be able to learn new concepts and bindings continuously. Keeping on learning new concepts or bindings without catastrophic forgetting of already learned ones is a very important ability for the learning system. Just as humans are able to learn new objects and their names

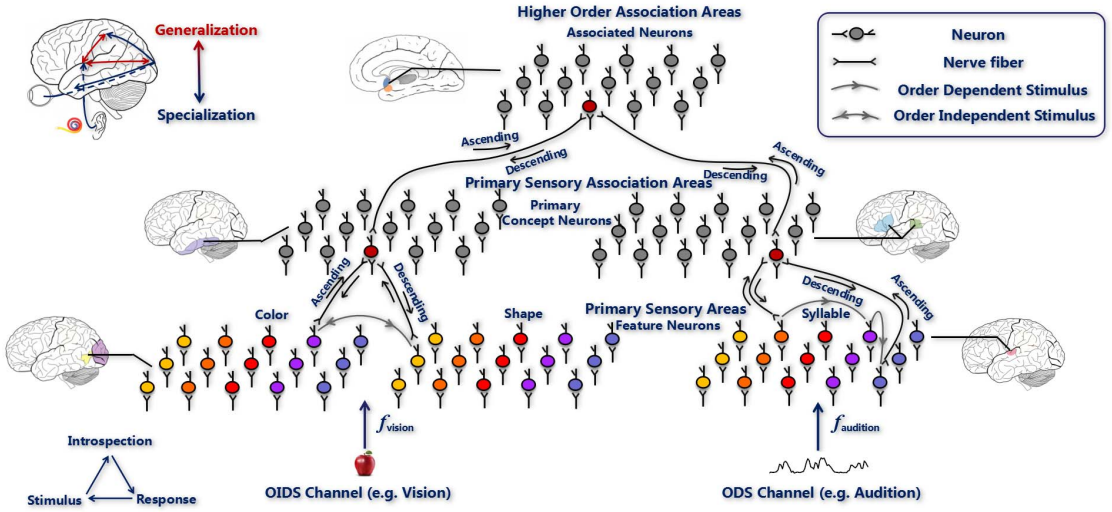


Fig. 3. Neural network modeling of the Perception Coordination Network. The hierarchical structure is inspired by brain's structure, i.e., each area in PCN is corresponding to an area in the brain, which performs some particular functions. Note that the figure only takes vision and audition for example, other sensations can also be involved in the structure.

without forgetting the previously learned ones throughout their lifetimes. Unfortunately, many learning systems suffer the stability-plasticity dilemma [45]. Taking this problem as a target, many online incremental methods for sensory integration are proposed.

In [46], an incremental knowledge robot 1 (IKR1) for word grounding is proposed, where a self-organizing incremental neural network (SOINN) [47] handles the visual module, and a vector-quantization (VQ) 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 [48], a multimodal self-organizing network is proposed for sensory integration with SOM [34] modules. Positional coordinates of unimodal SOM which receive sensory data are fused by a high-level SOM. Based on the architecture built in [48], a bimodal incremental self-organizing network (BiSON) [49] is developed which can incrementally integrate stimuli in visual and auditory modalities. As an application, a set of Chinese characters and related spoken words are effectively integrated. Binding of mental objects to written and spoken names is also reported in [50] and [51]. Besides, based on the BiSON, online transfer learning between a pair of multimodal integration systems is studied in [52]. In [53], a generalized heterogeneous fusion adaptive resonance theory (GHF-ART) is proposed. It describes a multichannel variant of the ART network which can be used for fusion of multimodal features, such as visual and textual features.

III. Perception Coordination Network

In this section, a biologically inspired neural network model named perception coordination network (PCN) is proposed to handle multimodal concept acquisition and binding between different sensory modules. Taking the visual and auditory channel for example, Fig. 3 presents the neural network modeling of the PCN. It is a hierarchical structure which contains the primary sensory area (PSA), the primary sensory association area (SAA), and the higher order association area (HAA). The PSA contains feature neurons, which process many elementary features, e.g., colors, shapes, and syllables. The primary SAA contains primary concept neurons, which combine the features in the PSA to represent unimodal concept, e.g., the image of an apple combines color and shape and the Chinese word "[píng guǒ]" combines a series of syllables. The higher order association area contains association neurons, which connect several primary sensory association areas just like synaesthesia, e.g., connecting the image of an apple and the word "[píng guǒ]". External input is categorized into Order InDependent Stimulus (OIDS) and Order Dependent Stimulus (ODS). For example, different visual features of an object, such as the color and shape features, belong to the OIDS. Because different orders of color and shape features do not affect the activation of the 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. Nerve impulses in PCN transmit in parallel, descending, and ascending directions. PCN works with a Stimulus-Introspection-Response formula, which means it receives input (stimulus) from users, then makes an introspection by comparing the current input with the learned knowledge, finally gives a response to the users based on the introspective result then waits for a new stimulus. Briefly, the main contributions of PCN as are as follows,

1) Different types of neurons with particular computational models are defined, increasing the interpretability of the hierarchical PCN structure.
2) Through creating of connections between neurons, PCN learns new concepts and bindings quickly without forgetting of already learned concepts and bindings.

In the following, we first give an overview of the network structure. Then, we give the learning process of the PCN. Table I gives the meaning of the notations used in PCN. Note that the dimensions of $\mathbf{M}^{\alpha}$, $\mathbf{M}^{\beta,i}$, $\mathbf{V}_{\alpha_j}^{\beta_i}$, and $\mathbf{M}^{A_i}$ are not fixed.

A. Overview of the Network Structure

As mentioned above, PCN is a modular structure. Each area in the network performs some particular function. In this

TABLE I
MEANING OF NOTATIONS

Notation	Meaning
$N_i^{F_\alpha}$	Feature neuron i in the primary sensory area α
$N_i^{C_\beta}$	Concept neuron i in the primary sensory association area β
N_i^A	Association neuron i in the higher order association area
$\boldsymbol{w}_i$	Weight vector of feature neuron i
σ_i	Activation times of neuron i
$c^h_{(i,j)}$	Horizontal connection between feature neuron i and j
$\tau_{(i,j)}$	Time parameter of the horizontal connection $c^h_{(i,j)}$
$\mathbf{M}^\alpha$	Matrix that stores the horizontal connections of PSA α
$c^v_{(i,j)}$	Vertical connection between feature neuron i and concept neuron j
$\rho_{(i,j)}$	Activity of the vertical connection $c^v_{(i,j)}$
$\mathbf{M}^{\beta,i}$	Matrix that stores the ODS bindings of concept neuron i in SAA β
$\mathbf{V}^{\beta,i,\alpha_j}$	Vector that stores the OIDS bindings of concept neuron i in SAA β to PSA α_j
$c^v_{(m,i,n)}$	Vertical connection between concept neuron m and concept neuron n through association neuron i
$\rho_{(m,i,n)}$	Activity of the vertical connection $c^v_{(m,i,n)}$
$\mathbf{M}^{\beta,A}$	Matrix that stores the connections between HAA and PSA β

section, an overview of each area within the PCN architecture will be given.

1) Primary Sensory Area: The PSA includes feature neurons, which respond to particular features, e.g., color features, shape features, or syllable features. See the bottom layer in Fig. 1.

Feature neurons in area α are stored in set N^{F_α}, as shown in Fig. 1; α can be the color feature area, the shape feature area, or the syllable feature area. $N_i^{F_\alpha}$ is used to denote feature neuron i in area α. $N_i^{F_\alpha} \triangleq \{\boldsymbol{w}_i, \sigma_i\}$, where $\boldsymbol{w}_i$ and σ_i represent the weight vector and the activation times of feature neuron i. The activating domains (ADs) of $N_i^{F_\alpha}$ are defined as follows:

$$\boldsymbol{ADs}\left(N_i^{F_\alpha}\right) \triangleq \{\boldsymbol{x} \mid \boldsymbol{Dis}(\boldsymbol{x}, \boldsymbol{w}_i) \le \theta\} \tag{1}$$

where θ is a parameter which controls the response range of the ADs. $\boldsymbol{x}$ is the feature vector extracted from stimuli received by sensory receptors. $\boldsymbol{Dis}(\boldsymbol{x}, \boldsymbol{w}_i)$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{w}_i$. Different distance functions are applied to different types of features, e.g., we use Euclidean distance for visual features and dynamic time warping (DTW) for auditory features. The details of the distance function used are described in Section III-B.

In this layer, horizontal connections between feature neurons are developed to organize the feature neurons into a feature map. The connection between feature neuron $N_i^{F_\alpha}$ and feature neuron $N_j^{F_\alpha}$ is defined as follows:

$$c^h_{(i,j)} \triangleq \left\{N_i^{F_\alpha}, N_j^{F_\alpha}, \tau_{(i,j)}\right\} \tag{2}$$

where $\tau_{(i,j)}$ is a time parameter which represents the age of the connection. Assuming that there are d feature neurons in area α, then a $d \times d$ dimensional 0-1 matrix $\mathbf{M}^\alpha$ can be used to store the connections, where $\mathbf{M}^\alpha_{i,j} = 0$ means that there is no horizontal connection between feature neurons $N_i^{F_\alpha}$ and $N_j^{F_\alpha}$ and $\mathbf{M}^\alpha_{i,j} = 1$ means that there exists a connection between $N_i^{F_\alpha}$ and $N_j^{F_\alpha}$.

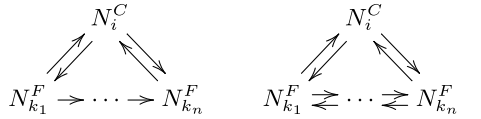


Fig. 4. Two types of the binding structures between the primary concept neurons and the feature neurons.

2) Primary Sensory Association Area: The primary SAA includes primary concept neurons, which combine feature neurons to represent some 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. Concept neurons in SAA β are stored in set N^{C_β}, where β can be the visual association area, auditory association area, or other SAA, see the middle layer in Fig. 1. $N_i^{C_\beta}$ is used to denote concept neuron i in area β.

Vertical connection between concept neuron $N_j^{C_\beta}$ and feature neuron $N_i^{F_\alpha}$ is defined as follows:

$$c^v_{(i,j)} \triangleq \left\{N_i^{F_\alpha}, N_j^{C_\beta}, \rho_{(i,j)}\right\} \tag{3}$$

where $\rho_{(i,j)}$ represents the cumulative times of the connection be activated.

As shown in Fig. 1, the external stimuli are divided into two types, which include Type I. OIDS and Type II. ODS. For example, different visual features, such as color and shape features, belong to the OIDS. Because different activation orders of color and shape features do not affect the activation of the visual concept that they refer to; syllables contained in a voice wave belong to the ODS, because different orders of the same group of syllables may refer to different concepts (words here). Correspondingly, two types of binding structure between the primary concept neurons and the feature neurons are defined as shown in Fig. 4. Then, two different types of the ADs of the concept neuron $N_i^{C_\beta}$ are defined as follows:

$$\boldsymbol{ADs}(N_i^{C_\beta}) \triangleq \begin{cases} \overrightarrow{\left(N_{k_1}^{F_\alpha}, N_{k_2}^{F_\alpha}, \ldots, N_{k_n}^{F_\alpha}\right)}, & \text{ODS} \\ \left(N_{k_1}^{F_{\alpha_1}}, N_{k_2}^{F_{\alpha_2}}, \ldots, N_{k_n}^{F_{\alpha_n}}\right), & \text{OIDS} \end{cases} \tag{4}$$

where feature neurons $N_{k_1}^{F_\alpha}, N_{k_2}^{F_\alpha}, \ldots, N_{k_n}^{F_\alpha}$ and $N_{k_1}^{F_{\alpha_1}}, N_{k_2}^{F_{\alpha_2}}, \ldots, N_{k_n}^{F_{\alpha_n}}$ connect the concept neuron $N_i^{C_\beta}$. Note that the arrow over the vector means that $N_i^{C_\beta}$ can be fired only by the firing of $N_{k_1}^{F_\alpha}, N_{k_2}^{F_\alpha}, \ldots, N_{k_n}^{F_\alpha}$ through the arrow's direction. We can represent the ODS binding of $N_i^{C_\beta}$ with a 0-1 matrix $\mathbf{M}^{\beta,i} = [\mathbf{M}_1^{\beta,i}, \mathbf{M}_2^{\beta,i}, \ldots, \mathbf{M}_n^{\beta,i}]$. $\mathbf{M}_j^{\beta,i}$ $(1 \le j \le n)$ is an d-dimensional column vector, where d is the number of feature neurons in area α. $\mathbf{M}^{\beta,i}_{k_j,j} = 1$ and other elements are 0. To present the OIDS binding of $N_i^{C_\beta}$, a group of column vectors $\{\mathbf{V}^{\beta,i,\alpha_1}, \mathbf{V}^{\beta,i,\alpha_2}, \ldots, \mathbf{V}^{\beta,i,\alpha_n}\}$ can be used. $\mathbf{V}^{\beta,i,\alpha_j}$ $(1 \le j \le n)$ is an d-dimensional column vector, where d is the number of feature neurons in area α_j. $\mathbf{V}^{\beta,i,\alpha_j}_{k_j} = 1$ and other elements are 0.

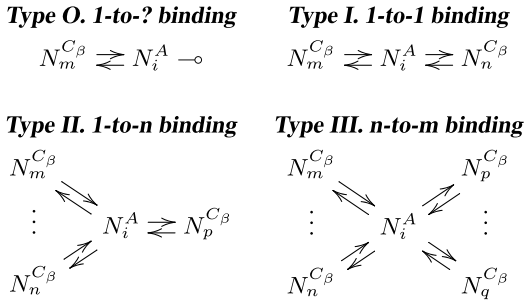


Fig. 5. Four types of the binding structures of the association neurons. Note that the association neurons can bind more than two SAAs.

3) Higher Order Association Area: The HAA includes the association neurons, which connect different SAAs, e.g., connects the image of an object, the object's names, and its taste, see the top layer in Fig. 1. Association neurons are stored in set N^A. N_i^A is used to denote association neuron i. Four types of binding structure will be generated during learning as shown in Fig. 5. The concept neurons in different sides of N_i^A in Fig. 5 come from different SAAs, e.g., the left side can be the primary visual association area and the right side can be the primary auditory association area.

The ADs of the association neuron N_i^A is the set of the primary concept neurons that N_i^A binds with

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

Note that the ADs of the association neuron is a set of neurons which means that any primary concept neuron in the set can activate it. Also, the primary concept neurons in the set can come from different SAAs. Thus, the association neuron has a multimodality activation mode.

Vertical connections between a concept neuron $N_m^{C_\beta}$ and another concept neuron $N_n^{C_\beta}$ through N_i^A are defined as follows:

$$c_{(m,i,n)}^{v} \triangleq \left\{N_m^{C_\beta}, N_i^A, N_n^{C_\beta}, \rho_{(m,i,n)}\right\} \tag{6}$$

where $\rho_{(m,i,n)}$ represents the cumulative times of the connection between N_m^C and N_n^C be activated. The connections between the HAA and SAA β can be stored with a 0-1 matrix $\mathbf{M}^{A,\beta}$, where $\mathbf{M}_{m,i}^{A,\beta} = 1$ means that there exists a connection between concept neuron $N_m^{C_\beta}$ in β and association neuron N_i^A and $\mathbf{M}_{m,i}^{A,\beta} = 0$ means that there is no such a connection. Assume that there are l_1 concept neurons in area β and l_2 association neurons, then $\mathbf{M}^{A,\beta}$ is an $l_1 \times l_2$ dimensional matrix. The cumulative times $\rho_{(m,i,n)}$ can be stored in a 3-D matrix $\mathbf{R}$, where $\mathbf{R}_{m,i,n}$ corresponds to $\rho_{(m,i,n)}$.

B. Learning Process

PCN is an online learning method which means that samples are fed into the network sequentially. When a pair of inputs arrive, e.g., a pair of visual input and auditory input, the PSA will extract features from the input data first. Then, competitive learning among feature neurons is conducted, and firing neurons will transmit activation signals to the SAA. Meanwhile, the firing neurons in the PSA will be updated. When the SAA receives the ascending signal from PSA, competition among concept neurons will be executed to activate the concept neurons. The activated signals will be transmitted to the HAA; meanwhile, the connections and neurons in PSA will be updated. When the HAA receives the ascending signal from the SAA, an unconscious impulse process will be triggered, which uses one channel's signal to wake its corresponding concepts in other channels, e.g., using input image to wake its names and tastes. After that, an introspection process is conducted, which aims to check the consistency between the current input pair and the learned knowledge (concepts and bindings) from the past pairs. Finally, the SAA will be updated according to the introspection process. When all the above processes for the current input sample are complete, the PCN will then deal with the next input pair. In the following, we use a pair of vision (OIDS) and audition (ODS) input to describe the method in detail.

1) Primary Sensory Area: Feature extraction is executed in the first step, as the function f_{vision} and f_{audition} shown in Fig. 1.

For the vision features, shape and color features of the objects are used. Normalized Fourier descriptors [54] are used to extract the shape features of the objects. The boundary $\boldsymbol{b}$ of the object $\boldsymbol{o}$ in a picture $\boldsymbol{p}$ is computed first, i.e., $\boldsymbol{b} = f_{\text{boundary}}(\boldsymbol{o})$, where $\boldsymbol{b}$ is a series of 2-D coordinates $b_i = (x_i, y_i)$, $0 \le i \le n-1$, which form the object's outline. Then, a complex format $\boldsymbol{b'}$ of $\boldsymbol{b}$ is obtained, that is, $b'_j = x_j + iy_j$, $i = \sqrt{-1}$. A Fourier descriptors $\boldsymbol{d'}$ of the object's outline is computed by the fast Fourier transform (FFT), i.e., $\boldsymbol{d'} = \boldsymbol{FFT}(\boldsymbol{b'})$, where $\boldsymbol{d'} = (d'_0, d'_1, \ldots, d'_{n-1})$. Finally, the normalized Fourier descriptor $\boldsymbol{d} = (d_1, d_2, \ldots, d_{n-1})$ will be obtained by $d_i = \|d'_i\|/\|d'_1\|, 1 \le i \le n-1$. Vector $(d_2, d_3, \ldots, d_{25})$ is chosen for the final shape feature.

Next, the color histogram is used to extract the color features of the objects. First, the object's image $\boldsymbol{g}$, which means the part of the picture that is surrounded by the boundary $\boldsymbol{b}$, will be extracted. Then, the color histogram $\boldsymbol{h}$ is extracted from image $\boldsymbol{g}$ with the color histogram function (HIST), i.e., $\boldsymbol{h} = \boldsymbol{HIST}(\boldsymbol{g}, \text{num})$, where num is a constant which determines the number of the containers in the histogram. We set num $= 0.05$.

For the audition features, Mel-Frequency Cepstral Coefficients (MFCCs) [55] are chosen for each syllable contained in the input voice wave $\boldsymbol{v}$. To do this, all syllables in $\boldsymbol{v}$ need to be extracted first. The wave $\boldsymbol{v}$ is filtered by a high-pass filter, where the system function is $H(z) = 1 - \mu z^{-1}$, and we set μ as 0.9375. Following this, the amplitude of $\boldsymbol{v}$ is normalized. Next, frame blocking is executed as $\boldsymbol{q} = \boldsymbol{FB}(\boldsymbol{v}, \text{len}, \text{olp})$, where len and olp are constants which represent the length of one frame and the length of the overlap between two adjacent frames. We set len $= 256$ and olp $= 128$. Now, assuming that m frames are gained from $\boldsymbol{v}$, i.e., $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m)$. After this step, the short-time energy E_i and short-time zero crossing rate C_i of each frame $\boldsymbol{q}_i$ are calculated. The frame $\boldsymbol{q}_i$ whose short-time energy E_i is larger than a threshold t_e or short-time zero crossing rate C_i is larger than a threshold t_c will be marked as a candidate frame and stored in set R, i.e., $R =$

$\{\boldsymbol{q}_i | E_i > t_e \vee C_i > t_c\}$. We set $t_e = 0.5$ and $t_c = 100$. Then, the consecutive candidate frames in R are spliced together as the candidate syllables. Meanwhile, the adjacent two candidate syllables with gap no more than one frame length, i.e., $1 \times \text{len}$, will be merged together. Also, candidate syllables whose length are less than two frame length, i.e., $2 \times \text{len}$, will be marked as noise and removed. Finally, the syllables $S = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k\}$ contained in the wave $\boldsymbol{v}$ are obtained and the MFCCs $\boldsymbol{m_i}$ of each syllable $\boldsymbol{s_i}$ in set S are extracted.

After the feature extraction step, the obtained feature vectors will be transmitted to their corresponding PSAs, e.g., shape area, color area, and syllable area. Then, competitive learning among feature neurons will be executed.

For the vision, a winner neuron $N_f^{F_\alpha}$ in each feature area α is found as follows:

$$N_f^{F_\alpha} = \underset{N_i^{F_\alpha} \in N^{F_\alpha}}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_i\|, \quad \text{where} \quad \alpha \in \{b, c\} \tag{7}$$

where $\boldsymbol{x}$ is the feature vector $\boldsymbol{d}$ (normalized Fourier descriptors) or $\boldsymbol{h}$ (color histogram). $\boldsymbol{w}_i$ is the weights of the feature neuron i in area α. $\|\cdot\|$ is the distance function; for vision, Euclidean distance is used, and for audition, DTW is used. The superscribed F_b and F_c represent the shape PSA and color PSA, respectively.

If $\boldsymbol{x}$ belongs to the ADs of neuron $N_f^{F_\alpha}$, i.e.,

$$\|\boldsymbol{x} - \boldsymbol{w}_f\|_2 \leq \frac{1}{4}\|\boldsymbol{w}_f\|_2 \tag{8}$$

the winner neuron N_f^F is activated.[1] The activation time of $N_f^{F_\alpha}$ is incremented by 1 and the weight vector of $N_f^{F_\alpha}$ is moved toward $\boldsymbol{x}$ as follows:

$$\sigma_f = \sigma_f + 1, \quad \boldsymbol{w}_f = \boldsymbol{w}_f + \frac{1}{\sigma_f}(\boldsymbol{x} - \boldsymbol{w}_f). \tag{9}$$

If $\boldsymbol{x}$ does not belong to the ADs of $N_f^{F_\alpha}$, PCN will recognize $\boldsymbol{x}$ as a new feature, and remember the new feature in its memory, i.e., a new neuron $N_{\text{new}}^{F_\alpha}$ will be created for $\boldsymbol{x}$ as follows:

$$N_{\text{new}}^{F_\alpha} = \{\boldsymbol{x}, 1\}, \quad \text{where} \quad \alpha \in \{b, c\}. \tag{10}$$

Then, $N_{\text{new}}^{F_\alpha}$ is activated.

Finally, the activation signals of the shape PSA and color PSA are transmitted to the SAA as follows:

$$\left(N_{f_b}^{F_b}, N_{f_c}^{F_c}\right) \xrightarrow{\text{Signal}} \text{SAA} \tag{11}$$

where $N_{f_b}^{F_b}$ and $N_{f_c}^{F_c}$ represent the activated neurons in the shape PSA and color PSA, respectively. The format of the activation signal of the shape PSA is a 0-1 row vector $\mathbf{P}^b$ and the dimension of $\mathbf{P}^b$ is the number of the neurons in the shape PSA. Meanwhile, $\mathbf{P}_{f_b}^b = 1$ and other elements are 0. The format of the activation signal of the color PSA $\mathbf{P}^c$ is similar.

[1]Here, θ of the AD function formula (1) is defined as 1/4 times the 2-norm of weight vector of the feature neuron.

For the audition, DTW is used to find winner neuron for each syllable $\boldsymbol{s_i}$, i.e.,

$$N_{f_i}^{F_\alpha} = \underset{N_j^{F_\alpha} \in N^{F_\alpha}}{\operatorname{argmin}} \boldsymbol{DTW}(\boldsymbol{m}_i, \boldsymbol{w}_j), \quad \text{where } 1 \leq i \leq k, \ \alpha \in \{s\} \tag{12}$$

where $\boldsymbol{w}_j$ is the MFCCs of syllable neuron j in the syllable PSA and $\boldsymbol{m}_i$ is the MFCCs of the ith syllable $\boldsymbol{s}_i$ in the input voice wave.

If $\boldsymbol{m}_i$ belongs to the ADs of the winner neuron, i.e.,

$$\boldsymbol{DTW}(\boldsymbol{m}_i, \boldsymbol{w}_{f_i}) < 200. \tag{13}$$

Then, $N_{f_i}^{F_s}$ is activated.[2]

If $\boldsymbol{m}_i$ does not belong to the ADs of $N_{f_i}^{F_s}$. A new syllable neuron N_{new}^F will be created for $\boldsymbol{m}_i$ as follows:

$$N_{\text{new}}^{F_s} = \{\boldsymbol{m}_i, 1\}. \tag{14}$$

Then, $N_{\text{new}}^{F_s}$ is activated.

Finally, the activated signals are transmitted to the SAA ascendingly as follows:

$$\overrightarrow{\left(N_{f_1}^{F_s}, N_{f_2}^{F_s}, \ldots, N_{f_k}^{F_s}\right)} \xrightarrow{\text{Signal}} \text{SAA}. \tag{15}$$

Assume that there are d syllable neurons in the syllable PSA, the format of the activated signals is a $k \times d$ dimensional 0-1 matrix $\mathbf{P}^s$, where $\mathbf{P}_{j,f_j}^s = 1$ $(1 \leq j \leq k)$ and other elements are 0.

Next, self-organizing [34] among feature neurons is conducted according to the competitive Hebbian learning rule [56]. If the winner neuron $N_f^{F_\alpha}$ is activated when an input sample $\boldsymbol{x}$ comes, the following condition will be checked for all other neurons:

$$\|\boldsymbol{x} - \boldsymbol{w}_i\| \leq \delta \cdot \theta, \quad \text{where } N_i^{F_\alpha} \in N^{F_\alpha} \setminus \left\{N_f^{F_\alpha}\right\}, \ \alpha \in \{a, b, s\} \tag{16}$$

where δ is a constant to control the activation behavior of $N_i^{F_\alpha}$ during self-organizing processing; usually, it is set larger than or equal to 1. We set $\delta = 3$ in this paper.

If (16) is satisfied, a horizontal connection will be established between $N_f^{F_\alpha}$ and $N_i^{F_\alpha}$ if no connection exists between them, which means thet $\mathbf{M}_{i,j}^{\alpha}$ is set to 1. Then, time parameter $\tau_{(f,i)}$ will be set to 0 to represent that the connection is newly built, i.e., $c_{(f,i)}^h = \{N_f^{F_\alpha}, N_i^{F_\alpha}, 0\}$. If there is already a connection between them, the time parameter $\tau_{(f,i)}$ will be set to 0 to renew the connection.

For the dynamically changing environment, neurons change their weights slowly during learning. Neurons that are connected with each other at an early stage may not have similar weights at an advanced stage. The connections which associate such neurons should be weakened or removed. Thus, if the winner neuron $N_f^{F_\alpha}$ is activated and updated, the connections emanating from it will be weakened by increasing their age parameter, i.e., $\tau_{(f,j)} = \tau_{(f,j)} + 1$, where $N_j^{F_\alpha} \in S_f^{F_\alpha}$, $S_f^{F_\alpha}$ represents the neighbor neuron set of $N_f^{F_\alpha}$, which means that neurons in the set connect to $N_f^{F_\alpha}$ directly. Connections $c_{i,j}^h$

[2]Here, θ of the AD function formula (1) is defined as 200.

whose time parameter is larger than a predefined threshold t_τ will be removed by setting $\mathbf{M}^a_{i,j} = 0$ and t_τ is set as 50 here.

2) Primary Sensory Association Area: When the SAA receives an ascending signal $\boldsymbol{x} = (N^{F_b}_{f_b}, N^{F_c}_{f_c})$ or $\boldsymbol{x} = \overrightarrow{(N^{F_s}_{f_1}, N^{F_s}_{f_2}, \ldots, N^{F_s}_{f_k})}$ from PSA, the signal is checked whether equal to any concept neuron's ADs, i.e., to find the solution of the following equation:

$$\boldsymbol{x} = \boldsymbol{ADs}\big(N^{C_\beta}_i\big), \quad \text{where } N^{C_\beta}_i \in N^{C_\beta},\ \beta \in \{a, v\} \tag{17}$$

where N^{C_a} and N^{C_v} represent the audition SAA and vision SAA, respectively.

For the audition (ODS situation), (17) means to find a concept neuron that satisfies the following condition:

$$\mathbf{P}^s_{j,\cdot} \cdot \mathbf{M}^{a,i}_{\cdot,j} = 1, \quad \text{where } 1 \le j \le k,\ N^{C_a}_i \in N^{C_a} \tag{18}$$

where $\mathbf{P}^s_{j,\cdot}$ is the row vector of $\mathbf{P}^s$ and $\mathbf{M}^{a,i}_{\cdot,j}$ is the column vector of $\mathbf{M}^{a,i}$.

If a concept neuron $N^{C_a}_{f_a}$ is found in (18), it means that the current input signal $\boldsymbol{x}$ has been encountered before, which is stored by neuron $N^{C_a}_{f_a}$. Then, $N^{C_a}_{f_a}$ is activated and the activation times of $N^{C_a}_{f_a}$ is increased by 1, i.e., $\sigma_{f_a} = \sigma_{f_a} + 1$. The activities of the connections between $N^{F_s}_{f_1}, N^{F_s}_{f_2}, \ldots, N^{F_s}_{f_k}$ and $N^{C_a}_{f_a}$ are also increased by 1, i.e., $\rho_{(f_i, f_a)} = \rho_{(f_i, f_a)} + 1, 1 \le i \le k$.

If no concept neuron is found, it means that $\boldsymbol{x}$ has not been encountered before. The SAA will remember this new concept, and a new neuron $N^{C_a}_{\text{new}}$ will be created for $\boldsymbol{x}$ as follows:

$$\boldsymbol{ADs}\big(N^{C_a}_{\text{new}}\big) = \boldsymbol{x}, \quad \sigma_{\text{new}} = 1. \tag{19}$$

Then, $N^{C_a}_{\text{new}}$ is activated. In the following, we denote $N^{C_a}_{\text{new}}$ as $N^{C_a}_{f_a}$. New connections between $N^{F_s}_{f_1}, N^{F_s}_{f_2}, \ldots, N^{F_s}_{f_k}$ and $N^{C_a}_{f_a}$ will then be created as follows:

$$c^v_{(f_j, f_a)} = \big\{N^{F_s}_{f_j}, N^{C_a}_{f_a}, 1\big\}, \quad 1 \le j \le k. \tag{20}$$

To realize (20), a new connection matrix $\mathbf{M}^{a,f_a}$ is created, where $\mathbf{M}^{a,f_a}_{f_j,j} = 1,\ 1 \le j \le k$.

For the vision (OIDS situation), (17) means to find a concept neuron that satisfies the following conditions:

$$\mathbf{P}^b \cdot \mathbf{V}^{v,i,b} = 1, \quad \text{where } N^{C_v}_i \in N^{C_v} \tag{21}$$

$$\mathbf{P}^c \cdot \mathbf{V}^{v,i,c} = 1, \quad \text{where } N^{C_v}_i \in N^{C_v}. \tag{22}$$

If a concept neuron $N^{C_v}_{f_v}$ is found by (21) and (22), $N^{C_v}_{f_v}$ is activated. Then, σ_{f_v} is increased by 1, i.e., $\sigma_{f_v} = \sigma_{f_v} + 1$, the activity of the connections between $N^{C_v}_{f_v}$ and $N^{F_b}_{f_b}$, $N^{F_c}_{f_c}$, i.e., $\rho_{(f_b, f_v)}$ and $\rho_{(f_c, f_v)}$, is also increased by 1.

If the shape $N^{F_b}_{f_b}$ falls in the ADs of some concept neuron $N^{C_v}_{f_v}$ but the color $N^{F_c}_{f_c}$ does not fall in the ADs of $N^{C_v}_{f_v}$, PCN will create a connection between $N^{C_v}_{f_v}$ and $N^{F_c}_{f_c}$ to bind different colors with the same shape, which we call a shape-centered neural coding strategy as shown in Fig. 6.

If no concept neuron is found by (21) and (22), meanwhile, the shape $N^{F_b}_{f_b}$ does not fall in any concept neuron's ADs, a new neuron $N^{C_v}_{\text{new}}$ will be created as follows:

$$\boldsymbol{ADs}\big(N^{C_v}_{\text{new}}\big) = \boldsymbol{x}, \quad \sigma_{\text{new}} = 1. \tag{23}$$

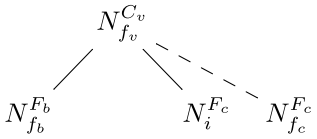


Fig. 6. Primary visual concept neuron connects novel color $N^{F_c}_{f_c}$ to associate it with shape $N^{F_b}_{f_b}$. $N^{F_c}_i$ is the existing color of $N^{F_b}_{f_b}$.

Then, $N^{C_v}_{\text{new}}$ is activated. We denote $N^{C_v}_{\text{new}}$ as $N^{C_v}_{f_v}$. Meanwhile, new connections between $N^{F_b}_{f_b}$, $N^{F_c}_{f_c}$, and $N^{C_v}_{f_v}$ will also be created

$$\begin{aligned} c^v_{(f_b, f_v)} &= \big\{N^{F_b}_{f_b}, N^{C_v}_{f_v}, 1\big\} \\ c^v_{(f_c, f_v)} &= \big\{N^{F_c}_{f_c}, N^{C_v}_{f_v}, 1\big\}. \end{aligned} \tag{24}$$

To realized (24), two vectors $\mathbf{V}^{v,f_v,b}$ and $\mathbf{V}^{v,f_v,c}$ will be created, where $\mathbf{V}^{v,f_v,b}_{f_b} = 1$, $\mathbf{V}^{v,f_v,c}_{f_c} = 1$, and other elements are 0.

Finally, the activated signals in the SAA are transmitted to the HAA

$$\big(N^{C_v}_{f_v}, N^{C_a}_{f_a}\big) \xrightarrow{\text{Signal}} \text{HAA}. \tag{25}$$

The format of the activation signal of the visual SAA is a 0-1 row vector $\mathbf{P}^v$, and the dimension of $\mathbf{P}^v$ is the number of the concept neurons in the visual SAA. $\mathbf{P}^v_{f_v} = 1$ and other elements are 0. The format of the activation signal of the auditory SAA $\mathbf{P}^a$ is similar.

3) Higher Order Association Area: The procedure of the HAA includes two processes which are the unconscious impulse process and the introspection process.

a) Unconscious impulse process: When the HAA receives an ascending signal $\boldsymbol{x} = (N^{C_v}_{f_v}, N^{C_a}_{f_a})$ from SAA, an unconscious impulse process will be triggered. First, taking the visual sense as the start point, to find the association neuron which connects neuron $N^{C_v}_{f_v}$, i.e., to find the solution of the following equation:

$$N^{C_v}_{f_v} \in \boldsymbol{ADs}\big(N^A_i\big), \quad N^A_i \in N^A. \tag{26}$$

Assuming that a solution set N^A_{vf} is found, then association neurons in set N^A_{vf} are activated, and the primary auditory concept neurons connecting to association neurons in N^A_{vf} will be unconsciously activated, which are

$$N^{C_a}_u = \big\{N^{C_a}_j \big| N^{C_a}_j \in \boldsymbol{ADs}\big(N^A_{vf}\big)\big\} \tag{27}$$

where set $N^{C_a}_u$ is used to represent these primary auditory concept neurons. Assume that there are l_1 visual concept neurons, l_2 auditory concept neurons, and l_3 association neurons, then the connection matrix $\mathbf{M}^{A,v}$ is an $l_1 \times l_3$ matrix and $\mathbf{M}^{A,a}$ is an $l_2 \times l_3$ matrix. Auditory concept neurons $N^{C_a}_j$ in set $N^{C_a}_u$ then satisfy the following condition:

$$\mathbf{P}^v \cdot \mathbf{M}^{A,v} \cdot {\mathbf{M}^{A,a}_{j,\cdot}}^\top = 1, \quad \text{where } 1 \le j \le l_2 \tag{28}$$

where $\mathbf{M}^{A,a}_{j,\cdot}$ is the jth row vector of $\mathbf{M}^{A,a}$.

By now, the bottom–up and top–down stimulation flow launched by a visual stimulus $\boldsymbol{V}$ can be summarized as shown in Fig. 7.

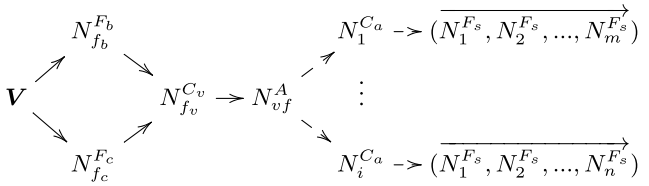


Fig. 7. Visual stimulus $\boldsymbol{V}$ fires its target neurons in auditory area through bottom–up and top–down flows.

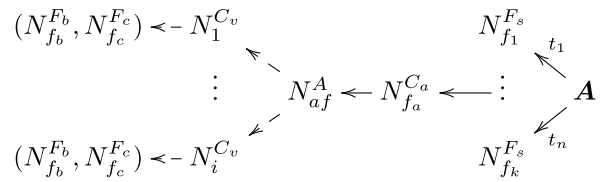


Fig. 8. Auditory stimulus $\boldsymbol{A}$ fires its target neurons in visual area through bottom–up and top–down flow.

Next, taking the auditory sense as the start point, do the unconscious impulse process using $N_{f_a}^{C_a}$. First, finding the association neuron which connects $N_{f_a}^{C_a}$, i.e., solve the following equation:

$$N_{f_a}^{C_a} \in \boldsymbol{ADs}\left(N_i^A\right), \quad N_i^A \in N^A. \tag{29}$$

Assuming that a solution set N_{af}^A is found, then association neurons in set N_{af}^A are activated, and the primary visual concept neurons connecting to association neurons in N_{af}^A are unconsciously activated, which are

$$N_u^{C_v} = \left\{N_j^{C_v} \middle| N_j^{C_v} \in \boldsymbol{ADs}\left(N_{af}^A\right)\right\} \tag{30}$$

where set $N_u^{C_v}$ is used to represent these primary visual concept neurons. Similarly, visual concept neurons $N_j^{C_v}$ in set $N_u^{C_v}$ should satisfy the following condition:

$$\mathbf{P}^a \cdot \mathbf{M}^{A,a} \cdot \mathbf{M}_{j,\cdot}^{A,v\top} = 1, \quad \text{where } 1 \le j \le l_1 \tag{31}$$

where $\mathbf{M}_{j,\cdot}^{A,v}$ is the jth row vector of $\mathbf{M}^{A,v}$.

The bottom–up and top–down stimulation flows launched by an auditory stimulus $\boldsymbol{A}$ can be summarized as shown in Fig. 8.

b) Introspection process[3]*:* After the unconscious impulse process, an introspection process will be executed. The process is divided into four conditions.

(a) If some association neuron N_{af}^A is found through (29) and no association neuron is found through (26), i.e., $N_{vf}^A = \varnothing \wedge N_{af}^A \neq \varnothing$. This means that the view of the current input of the object is new to PCN, but the voice is met by PCN which is used to call some other views. The current input should look like the views symbolized by the primary concept neurons in set $N_u^{C_v}$ according to the current input voice. Therefore, a contradiction emerges in the network and PCN will ask the user a question: "The current input name $N_{f_a}^{C_a}$ can also represent the current input view?" An answer γ from users is needed to help make a judgment.

If the user gives a positive answer, e.g., $\gamma = 1$, it means that the current input view is also called $N_{f_a}^{C_a}$, then concept neuron $N_{f_v}^{C_v}$ is added to the ADs of each association neuron in N_{af}^A, i.e.,

$$\boldsymbol{ADs}\left(N_{af}^A\right) = \boldsymbol{ADs}\left(N_{af}^A\right) \cup N_{f_v}^{C_v}. \tag{32}$$

The connection between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ will be initialized as follows:

$$c_{(f_v,af,f_a)}^v = \left\{N_{f_v}^{C_v}, N_{af}^A, N_{f_a}^{C_a}, 1\right\}. \tag{33}$$

Formulas (32) and (33) can be realized by setting $\mathbf{M}_{f_v,af}^{A,v} = 1$ and $\mathbf{R}_{f_v,af,f_a} = 1$; here, we use af to represent the index of the association neuron in N_{af}^A.

If the user gives a negative answer, e.g., $\gamma = 0$, it means that the current input view is not called $N_{f_a}^{C_a}$. The view $N_{f_v}^{C_v}$ will be stored as a Type O binding with a new association neuron, i.e.,

$$\boldsymbol{ADs}\left(N_{\text{new}}^A\right) = \left\{N_{f_v}^{C_v}\right\}, \quad \sigma_{\text{new}} = 1. \tag{34}$$

To realize (34), we set $\mathbf{M}_{f_v,\text{new}}^{A,v} = 1$. Note that the dimension of $\mathbf{M}^{A,v}$ is increased here.

(b) If some association neuron N_{vf}^A is found through (26) and no association neuron is found through (29), i.e., $N_{vf}^A \neq \varnothing \wedge N_{af}^A = \varnothing$, it means that the current input voice is new to PCN, but the view of the current object is met before and recognized by PCN. The object should be called with the names symbolized by the primary auditory concept neurons in set $N_u^{C_a}$. Then, PCN will ask the user a question, "This object is called $N_u^{C_a}$ previously. Is it also called $N_{f_a}^{C_a}$?" And an answer γ is needed.

If the user gives a positive answer, e.g., $\gamma = 1$, it means that the object is also called $N_{f_a}^{C_a}$. Concept neuron $N_{f_a}^{C_a}$ is added to the ADs of each association neuron in N_{vf}^A, i.e.,

$$\boldsymbol{ADs}\left(N_{vf}^A\right) = \boldsymbol{ADs}\left(N_{vf}^A\right) \cup N_{f_a}^{C_a}. \tag{35}$$

The connection between this new combination will be initialized as follows:

$$c_{(f_v,vf,f_a)}^v = \left\{N_{f_v}^{C_v}, N_{vf}^A, N_{f_a}^{C_a}, 1\right\}. \tag{36}$$

Formulas (35) and (36) can be realized by setting $\mathbf{M}_{f_a,vf}^{A,a} = 1$ and $\mathbf{R}_{f_v,vf,f_a} = 1$; here, we use vf to represent the index of the association neuron in N_{af}^A.

If the user gives a negative answer, e.g., $\gamma = 0$, it means that 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. Of course, one can also store neuron $N_{f_a}^{C_a}$ in the network as an ungrounded association; this means a different situation that you heard this thing, but never saw it. To realize this, we set $\mathbf{M}_{f_a,\text{new}}^{A,a} = 1$.

(c) If some association neurons N_{vf}^A and N_{af}^A are found through (26) and (29), i.e., $N_{vf}^A \neq \varnothing \wedge N_{af}^A \neq \varnothing$, it means that

[3]When the input pair includes auditory input, the introspection process is needed, otherwise, it can be omitted, for example, the vision-gustation input pair, because such a pair does not reflect definition of a concept with others. Although contradictions between input and learned knowledge can happen, we cannot avoid it by rejecting the input pair through conversing with users.

PCN has seen the image and heard the voice. The coherence should be checked first. If

$$N_{vf}^{A} \cap N_{af}^{A} \neq \varnothing \tag{37}$$

it means that $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ activate some association neurons in common. The current input pair of image and voice is consistent with some previous pairs. The activity of the connections between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ through the activated association neurons in common will be increased by 1 to strengthen the association, i.e., $\mathbf{R}_{f_v,i,f_a} = \mathbf{R}_{f_v,i,f_a} + 1$, where $N_i^A \in N_{vf}^A \cap N_{af}^A$.

If $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ activate different association neurons, i.e.,

$$N_{vf}^{A} \cap N_{af}^{A} = \varnothing \tag{38}$$

it means that the current combination between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ is inconsistent with some previous pairs. PCN will ask a question, "The current input pair is inconsistent with some previous pairs, is it an expected combination?" An answer γ is needed to help make a judgment.

If the user gives a positive answer, e.g., $\gamma = 1$, it means that the current view and name recognized by PCN is an expected combination. Then, neurons $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ are added to the ADs of each association neuron in set N_{af}^A and N_{vf}^A, respectively, i.e.,

$$\begin{aligned} \boldsymbol{ADs}\left(N_{vf}^{A}\right) &= \boldsymbol{ADs}\left(N_{vf}^{A}\right) \cup N_{f_a}^{C_a} \\ \boldsymbol{ADs}\left(N_{af}^{A}\right) &= \boldsymbol{ADs}\left(N_{af}^{A}\right) \cup N_{f_v}^{C_v} \end{aligned} \tag{39}$$

meanwhile, connections between the new combination will be initialized as follows:

$$\begin{aligned} c_{(f_v,vf,f_a)}^{v} &= \left\{N_{f_v}^{C_v}, N_{vf}^{A}, N_{f_a}^{C_a}, 1\right\} \\ c_{(f_v,af,f_a)}^{v} &= \left\{N_{f_v}^{C_v}, N_{af}^{A}, N_{f_a}^{C_a}, 1\right\}. \end{aligned} \tag{40}$$

Similarly, (39) and (40) can be realized by setting $\mathbf{M}_{f_v,af}^{A,v} = 1$, $\mathbf{M}_{f_a,vf}^{A,a} = 1$, $\mathbf{R}_{f_v,vf,f_a} = 1$, and $\mathbf{R}_{f_v,af,f_a} = 1$.

If the user gives a negative answer, e.g., $\gamma = 0$, it means that the combination is not an expected combination. Then, no operation will be done to the network.

(d) If no association neurons are found through (26) and (29), i.e., $N_{vf}^A = \varnothing \wedge N_{af}^A = \varnothing$. This means that $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ are new to PCN. The HAA will remember this new combination by adding a new association neuron N_{new}^A as follows:

$$\boldsymbol{ADs}\left(N_{\text{new}}^{A}\right) = \left\{N_{f_v}^{C_v}, N_{f_a}^{C_a}\right\}, \quad \sigma_{\text{new}} = 1. \tag{41}$$

A connection between $N_{f_v}^{C_v}$ and $N_{f_a}^{C_a}$ through the new association neuron N_{new}^A will then be created for this combination, i.e.,

$$c_{(f_v,\text{new},f_a)}^{v} = \left\{N_{f_v}^{C_v}, N_{\text{new}}^{A}, N_{f_a}^{C_a}, 1\right\}. \tag{42}$$

To realize (41) and (42), we expand matrix $\mathbf{M}^{A,v}$, $\mathbf{M}^{A,a}$, $\mathbf{R}$, and set $\mathbf{M}_{f_v,\text{new}}^{A,v} = 1$, $\mathbf{M}_{f_a,\text{new}}^{A,a} = 1$, and $\mathbf{R}_{f_v,\text{new},f_a} = 1$.

Now, the whole learning procedure for a pair of visual and auditory input is finished. PCN will go to the next input pair.[4] As a summary, the algorithm of PCN is given in Algorithm 1.

[4]The code is available at https://github.com/cloudlee711/PCN

Algorithm 1 Perception Coordination Network

Initialize: Set the value θ, δ, t_τ and the parameters in feature extraction step.

1: Receive a pair of image (OIDS) and name (ODS) of an object.
2: PSA: Execute feature extraction. For the vision, extract normalized Fourier descriptor and color histogram. For the audition, extract Mel-Frequency Cepstral Coefficients of each syllable contained in the stimulus.
3: PSA: Execute competitive learning. For the vision, use the procedure from formula (7) to (11). For the audition, use the procedure from formula (12) to formula (15). Execute self-organizing among feature neurons.
4: SAA: Execute concept learning procedure. For the audition, use the procedure from formula (18) to (20). For the vision, use the procedure of formula (21) to formula (24). Transmit activating signals to HAA using formula (25)
5: HAA: Execute the unconscious impulse process from formula (26) to formula (31).
6: HAA: Execute the introspection process from formula (32) to formula (42).
7: Waiting for the next input pair and go to step 1.

Fig. 9. Examples of the objects and the pronunciations of their Chinese names.

IV. Experiments

A. Experimental Details

The concept acquisition and binding among vision, audition, and gustation are conducted. Twenty objects are used. Fig. 9 shows the examples of the objects and the pronunciations of their Chinese names. There are voices with the same syllables but in different orders referring different objects (ODS), e.g., "[bō luó]" and "[luó bō]." There are also different voices referring the same object, e.g., "[píng guǒ]" and "[zhì huì guǒ]." And objects with different colors but similar shapes are included. Because we do not have real taste data, an artificial taste data set is designed. The taste data are a 6-D vector (sweet, sour, salt, bitter, umami, and hot). The value of each attribute is in the range of [0, 1]. For example, we design the taste of apple as follows: the value of sweet is uniformly distributed in the range [0.5, 0.6], the value of sour is uniformly distributed in the range [0, 0.1], and other attributes are 0.

We design a task which aims to teach the learning system object's name and taste in an online way, just like the learning task in [46] that teaches the learning system object's color, shape, and name properties. During the learning experiment, we first let PCN learn the view-name concept acquisition and binding. At each round of learning, an object is put in

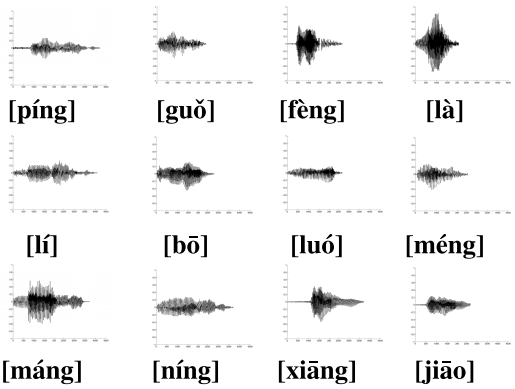


Fig. 10. Several examples of the pronunciation of the syllables. Double click on the icon to listen to the pronunciation.

front of a camera. Then, we start the audition program and pronounce the name of the object in Chinese. At the same time, the vision program captures the images of the object. When the pronunciation is finished, the audition program is closed. One round of learning is finished. Then, we go to the next round. To verify the robustness of PCN, we rotate the object in each round. To test the flexibility, we also call the same object in different names if it has.

When all objects' view-name learning is finished, we let PCN learn the view-taste concept acquisition and binding. At each round of learning, the taste data of an object are given to PCN when the vision program captures the images of the object. Algorithm of gustatory module is similar to that of visual module, because they both receive OIDS. And introspection process is not needed between gustation and vision.

We conduct the experiment in two environments which are: 1) the *Closed environment* and 2) the *Open-ended environment* (for the stability-plasticity dilemma [45]). In a closed environment, object is randomly chosen from the 20 objects in each round of learning. In an open-ended environment, we first let the methods learn 10 objects. In each round of learning, an 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, each time containing 352 rounds of learning for 20 objects, in both closed and open-ended environments to check the statistical properties of the learning results.

Meanwhile, for the auditory channel, we give two settings for the syllable learning. *(Setting 1)*: the incremental learning strategy as described in Section III-B1, the procedure from (12) to (14). *(Setting 2)*: a group of syllable neurons are constructed beforehand. First, we collect voices which are the names of the objects used in the experiments. Then, the voices are processed by the audition program, including syllable segmentation and syllable feature extraction (see Section III-B1). Finally, we get a group of syllable features which will be used in the learning experiment. The syllable group includes many Chinese syllables, see examples in Fig. 10, which constitute the syllable PSA of the PCN. During learning, PCN combines them to form words and does not add new syllables to the syllable group. We give these two settings to compare learning results based on incremental learned syllable neurons with learning results based on a predefined syllable neurons.

The parameters of PCN are set as follows: to extract the object's boundary, a Gaussian filter with [Hsize $= 15$, $\sigma = 9$] is used to smooth the image. Then, the image is converted to a binary image with a gray threshold of 0.7. The dimension of the normalized Fourier descriptors for the object's boundary is 23. For the color histogram, the container size is 0.05, where color value is first normalized into interval [0, 1]. For the syllable extraction and MFCC feature of each syllable, frame size is 256 and frame shift is 128. The pre-emphasis filter coefficient is 0.9375. The threshold of the short-time energy and short-time zero crossing are 0.5 and 100.

We compare PCN with the IKR1 system [46], MMSOM [48], BiSON [49], and GHF-ART [53], which aim to handle word grounding and multimodal feature fusion. Among the four methods, IKR1 system, BiSON, and GHF-ART are online learning methods. Because IKR1 system, MMSOM, and BiSON only give a vision-audition bimodal learning algorithm, we let them learn visual and auditory data in the experiments. GHF-ART does not give a fuzzy operation for auditory information with different dimensions; we let GHF-ART learn visual and gustatory data. The parameters of the IKR1 system are set as follows, $a_{\max} = 200$, $\lambda = 200$, and $c = 0.1$. The grid size of the SOM used in the MMSOM is 36×36 and 22 000 epochs of training are applied. For the BiSON, the number of neuronal units proportional to the number of stimuli is 16. In a closed environment, we train BiSON using all objects in the beginning. In an open-ended environment, we train BiSON using 10 objects in the beginning then add one object from the left 10 objects at each incremental learning step. The 22 000 epochs of training in total are applied. The parameters of the GHF-ART are $\alpha = 0.01$, $\beta = 0.6$, and $\rho = 0.95$. The comparison methods are reconstructed according to the referenced publications, and parameters are tuned following the suggestions in the publications.

B. Learning Results

For setting 1 which means that syllable neurons are incremental learned, for 30 times experiments in the closed environment and 30 times experiments in the open-ended environment, 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, 52 to 59 gustatory concept neurons, and 60 to 62 association neurons. An average of 89 questions are asked by PCN during learning. In about 85 questions on new words with similar views, the answer by user is positive. The other about four questions are caused by visual and auditory erroneous judgments. Thus, they get negative answers.

For setting 2 which means that syllable neurons are constructed beforehand, for 30 times experiments in both closed environment and open-ended environment, PCN gets 21 to 23 auditory concept neurons and 19 to 21 association neurons; other neurons are similar with setting one. An average of 62 questions is asked by the robot during learning. In about 56 questions on different names of the same object,

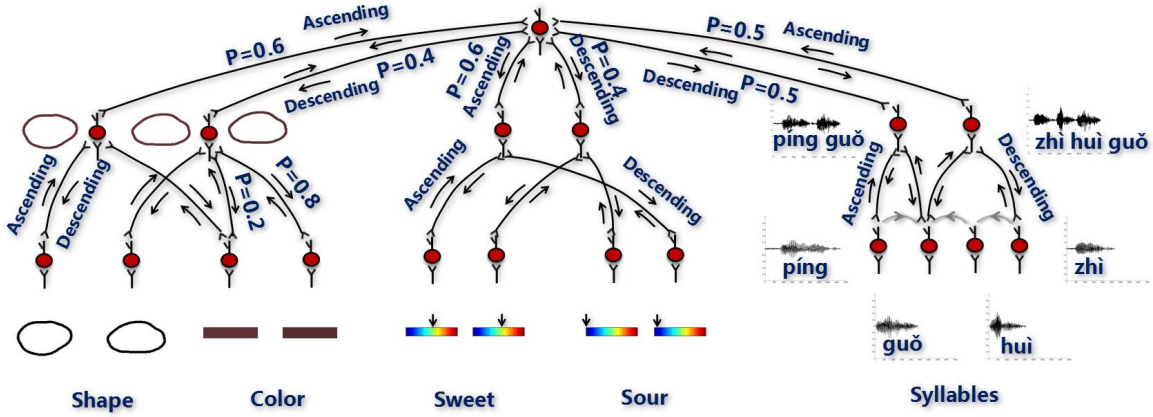


Fig. 11. Example of the network structure of the concept "[píng guǒ]" or "[zhì huì guǒ]" (apple) including view, taste, and name. The probability of the connection between visual concept neuron and each auditory concept neuron (or word neuron) through the association neuron is 0.5, which is gained by the normalization of the activity parameter of these connections. Because the user says "[píng guǒ]" and "[zhì huì guǒ]" with equal probability in the experiment. The probability of the connection between visual concept neuron and each gustatory concept neuron is gained in the similar way. Icons next to the neurons represent the objects to which the neurons maximally respond.

TABLE II

EXAMPLES OF THE ASSOCIATIONS OF VISION, AUDITION, AND GUSTATION (ONE EXAMPLE PER OBJECT). THE VECTORS IN THE GUSTATION COLUMN ARE [SWEET, SOUR, SALT, BITTER, UMAMI, AND HOT]. DOUBLE CLICK ON THE ICON TO LISTEN TO THE PRONUNCIATION

Vision	Audition	Gustation	Vision	Audition	Gustation
	[píng guǒ] [zhì huì guǒ]	[0.58 0.06 0 0 0 0]		[tǔ dòu]	[0.04 0 0 0 0]
	[xiāng jiāo]	[0.44 0 0 0 0 0]		[huā shēng]	[0.08 0 0.03 0 0 0] [0.04 0 0.06 0 0 0]
	[bō luó] [fèng lí]	[0.63 0.37 0 0 0 0]		[bǎn lì]	[0.25 0 0 0 0 0]
	[luó bō]	[0.17 0 0 0 0 0.20] [0.08 0 0 0 0 0.27]		[yáng cōng]	[0 0 0 0 0 0.47]
	[shí liu]	[0.63 0.27 0 0.08 0 0]		[níng méng]	[0 0.90 0 0 0 0] [0.10 0.80 0 0 0 0]
	[máng guǒ]	[0.76 0.01 0 0 0 0]		[jī̄ dàn]	[0 0 0 0 0.51 0]
	[là jiāo]	[0 0 0 0 0 0.95]		[shèng nǚ guǒ]	[0.34 0.28 0 0 0 0]
	[cǎo méi]	[0.58 0.27 0 0 0 0]		[gān jú]	[0.82 0.05 0 0 0 0]
	[fān qié]	[0.08 0.54 0 0 0 0]		[huáng lí]	[0.69 0.10 0 0 0 0] [0.62 0.03 0 0 0 0]
	[lí]	[0.67 0.22 0 0 0 0]		[lí]	[0.69 0.15 0 0 0 0]

the answers are positive. The other six questions are caused by visual and auditory erroneous judgments. Thus, they get negative answers.

It can be found that there are more auditory concept neurons for words in the setting 1 experiments. This is because the incremental learning generates more syllables and the syllables combine to generate more words. As s result, the probability of treating an input word as a new one is larger in setting 1 than that in setting 2. Then, the association neurons and questions in setting 2 are more than those in setting 1.

Table II gives some examples (one example per object) of the associations of vision, audition, and gustation. It can be found that PCN correctly acquires these concepts and properly binds them.

For a more detailed observation, Fig. 11 shows the structure of the concept apple in PCN. The association neuron connects the concept neurons in visual, auditory, and gustatory area. The visual concept neuron connects the shape and color feature neurons. An example of the shape-centered neural coding strategy is given, that is, a visual concept neuron connects

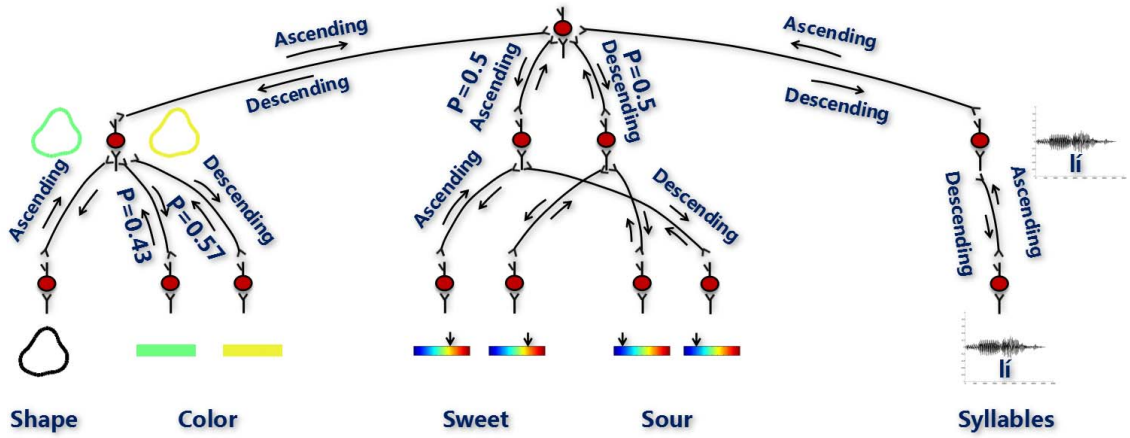


Fig. 12. Example of the network structure of the concept "[lí]" (pear) including view, taste, and name. This situation is the object with different colors but the same shape. In the network, the primary visual concept neuron connects different colors to associate them with their common shape. Icons next to the neurons represent the objects to which the neurons maximally respond.

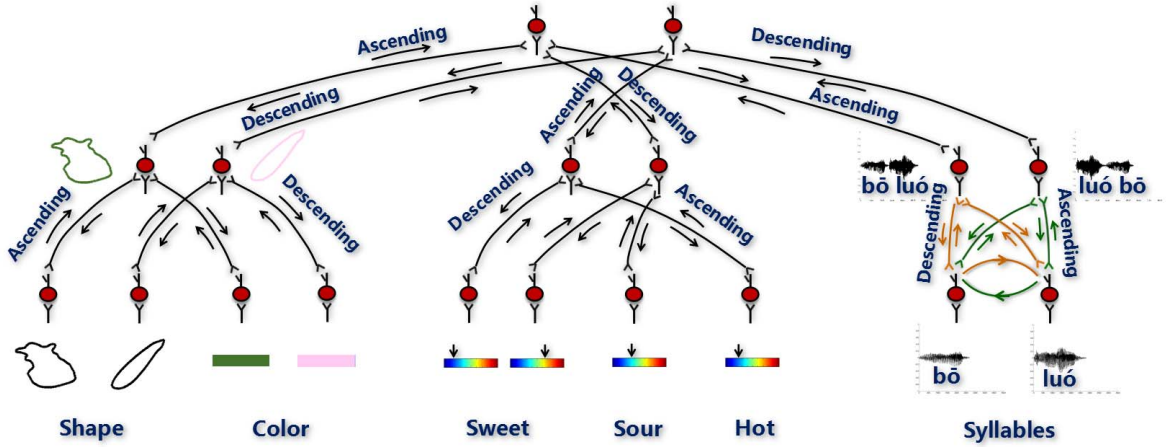


Fig. 13. Example of the network structure of the concept "[bō luó]" (pineapple) and "[luó bō]" (turnip) including view, taste, and name. The auditory input is the ODS, i.e., voices with the same syllables but in different orders referring different objects, see the left (vision) part and right (audition) part of the network. Icons next to the neurons represent the objects to which the neurons maximally respond.

one shape feature neuron and two color feature neurons. The auditory concept neuron (or word neuron) connects several syllable feature neurons. The gustatory concept neuron is a combination of six kinds of flavor neurons. The probability of the connection between visual concept neuron and each auditory concept neuron through the association neuron is 0.5, which is gained by the normalization of the activity parameter ρ of these connections, because we say "[píng guǒ]" and "[zhì huì guǒ]" with equal probability. Every time an apple is shown to the system and the word "[píng guǒ]" is spoken simultaneously; the activity of the connection between the visual concept neuron apple and the auditory concept neuron "[píng guǒ]" through their association neuron is increased by 1. It is a similar situation for the word "[zhì huì guǒ]." The probability of the connection between visual concept neuron and each gustatory concept neuron through the association neuron is acquired in a similar way.

Fig. 12 shows the structure of the concept pear in PCN. In the visual area (left part), the primary visual concept neuron connects different colors to associate them with their common shape. It means that the shape and color features can combine with each other freely through such structure.

Fig. 13 shows the structure of the concept pineapple and turnip in PCN. In the auditory area (right part), different activation orders of two syllable neurons, "[bō]" and "[luó]," fire different auditory concept neurons, "[bō luó]" and "[luó bō]." Then, the two auditory concept neurons fire their corresponding visual concept neurons and gustatory concept neurons through different association neurons.

Fig. 14 shows an example of the self-organizing result of 33 color feature neurons. As shown in Fig. 14, similar colors are connected together. Because the colors of learning objects do not occupy the whole space, six disjunct color clusters are obtained at the ending of the learning. Other features including syllable feature are organized similarly by their particular distance metric. Fig. 15 shows the syllables that connected to the syllable "luó." It can be found that the pronunciations of these syllables are very similar to "luó."

Interestingly, the name and the taste of the object are linked together automatically through the association neuron, whereas the name and taste data were not given to the system simultaneously during learning. It can be said that system with well-defined physical structure is able to make subsystems coordinate naturally.

TABLE III

STATISTICAL RESULTS OF THE TESTING EXPERIMENTS (mean + std). N/A REPRESENTS THE TESTING CONDITION THAT IS NOT APPLICABLE FOR THE METHOD. PCN^1 AND PCN^2 REPRESENT THE RESULTS OF SETTING 1 AND SETTING 2, RESPECTIVELY. SIGNIFICANT DECREASE OF THE ACCURACY IS MARKED BY ↓. THE BEST AND RUNNER-UP 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
	BiSON	76.52±4.31%	66.73±4.95% ↓	N/A
	PCN^2	**84.55±3.63%**	**84.88±4.16%**	**90.64±2.87%**
	PCN^1	**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
	BiSON	80.78±3.96%	77.87±3.69%	N/A
	PCN^2	**84.60±4.98%**	**86.87±4.16%**	**91.85±2.74%**
	PCN^1	**84.83±3.26%**	**88.07±2.78%**	**92.09±2.16%**

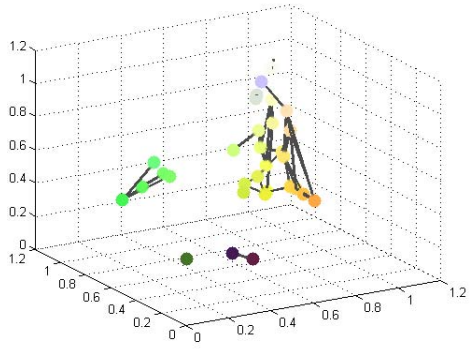


Fig. 14. Self-organizing of 33 color features learned by PCN. The coordinate of each point is the RGB value of the color which is transformed from the color histogram of the feature neuron. The features are organized by the Euclidean distance. Similar colors under the Euclidean distance metric are connected together.

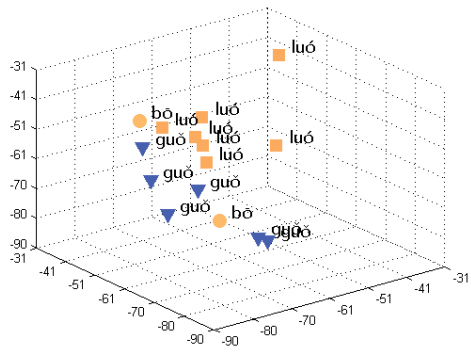


Fig. 15. Syllables that connected to the syllable "luó." To visualize the syllable neurons, we use principal component analysis to reduce the weights (MFCC features) of each syllable neuron to a 3-D vector. The features are organized by the DTW distance. The pronunciations of these syllables are very similar to "luó."

C. Testing Results

To test the model learned by PCN, we use one kind of sensory input to recall other two kinds of sensory output. That is, vision input recalls audition and gustation output, audition input recalls vision and gustation output, and gustation input recalls vision and audition output. Because IKR1 and MMSOM learn a vision-audition bimodal model, we only do vision recalls audition and audition recalls vision in the testing experiments. Also, GHF-ART learns visual and gustatory data; we do vision recalls gustation and gustation recalls vision. Testing is conducted after each time of learning experiment, i.e., totally 60 times of testing, 30 times for closed environment, and 30 times for open-ended environment.

During each time of testing, 528 rounds of recalling (176 rounds for each type of recall) are executed. To test the trained model, the incremental learning function of PCN, IKR1, and GHF-ART is disabled. During testing, if the fired syllables group, assume $N_{f_1}^{F_s}, N_{f_2}^{F_s}, \ldots, N_{f_k}^{F_s}$, by the input voice cannot activate a word neuron, we replace syllable $N_{f_i}^{F_s}$ $(1 \leq i \leq k)$ with syllables that connect to it to try to activate a word neuron again. If all replacements fail to activate a word neuron, we return no result and mark it as a mistake.

Table III shows the testing results. PCN^1 and PCN^2 represent the results of setting 1 and setting 2, respectively. Significant decrease of the accuracy in one environment is marked by ↓ in Table III. The results show 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 is mainly due to the recognition of the latter 10 "new" objects. IKR1 usually "forgets" previously learned objects when samples from a larger number of classes come together. The accuracy of PCN is much higher and more stable in both environments. Meanwhile, from Table III, it can also be found that the testing results of the visual and gustatory channels based on the incremental learned syllable neurons are comparable with the learning results based on the predefined syllable neurons, and for auditory channel, the testing results based on the incremental learned syllable neurons are better.

Next, to test the antinoise capability of the PCN, we add salt and pepper noise to testing images when doing the vision input recalls audition and gustation output experiments. Fig. 16 shows that the noise does not have a strong influence when the noise density is smaller than 0.2, where the accuracy is about 75%.

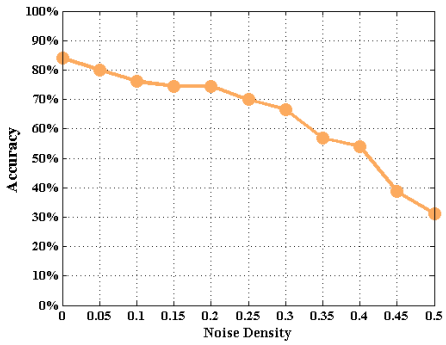


Fig. 16. Result of the recalling experiment with noisy images. The abscissa indicates the noise density of the salt and pepper noise added to testing images. The ordinate indicates the accuracy of the recalling.

V. Conclusion

In this paper, we propose a Perception Coordination Network (PCN) for online incremental multimodal concept acquisition and binding. The hierarchical and modularized structure of PCN is inspired by brain's structure. Different computational models are designed for different types of neurons including the feature neurons, the concept neurons and the association neurons. Meanwhile, new connections between neurons can be created for new concept bindings.

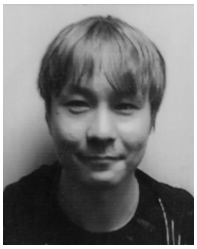
In the future, the framework can be further enhanced. For example, the horizontal connections in the SAA and HAA can be developed; more features can be included to make a more robust perception; the audition ability, which is limited to noun words acquisition currently, should be improved.

Finally, motivated by Simon's analysis [57] that the hierarchical system is helpful to evolution, in the next, we will show PCN has a good perception extendibility to solve the perception evolution problem [58]–[61].

References

[1] M. O. Ernst and H. H. Bülthoff, "Merging the senses into a robust percept," *Trends Cogn. Sci.*, vol. 8, no. 4, pp. 162–169, 2004.
[2] I. E. de Araujo and S. A. Simon, "The gustatory cortex and multisensory integration," *Int. J. Obesity*, vol. 33, pp. S34–S43, Jun. 2009.
[3] D. S. Bassett *et al.*, "Hierarchical organization of human cortical networks in health and schizophrenia," *J. Neurosci.*, vol. 28, no. 37, pp. 9239–9248, 2008.
[4] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, Mar. 2009.
[5] Z. Chen *et al.*, "Structural connectivity between visual cortex and auditory cortex in healthy adults: A diffusion tensor imaging study," *J. Southern Med. Univ.*, vol. 33, no. 3, pp. 338–341, 2013.
[6] L. Cohen, G. Rothschild, and A. Mizrahi, "Multisensory integration of natural odors and sounds in the auditory cortex," *Neuron*, vol. 72, no. 2, pp. 357–369, 2011.
[7] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Rev. Neurosci.*, vol. 14, no. 5, pp. 350–363, 2013.
[8] J. A. Gottfried, "Central mechanisms of odour object perception," *Nature Rev. Neurosci.*, vol. 11, vol. 9, pp. 628–641, 2010.
[9] J. B. Jadauji *et al.*, "Modulation of olfactory perception by visual cortex stimulation," *J. Neurosci.*, vol. 32, no. 9, pp. 3095–3100, 2012.
[10] H.-J. Park and K. Friston, "Structural and functional brain networks: From connections to cognition," *Science*, vol. 342, pp. 579–587, Nov. 2013.
[11] W. K. Simmons, A. Martin, and L. W. Barsalou, "Pictures of appetizing foods activate gustatory cortices for taste and reward," *Cerebral Cortex*, vol. 15, no. 10, pp. 1602–1608, 2005.
[12] A. M. Chan *et al.*, "Speech-specific tuning of neurons in human superior temporal gyrus," *Cerebral Cortex*, vol. 24, no. 10, pp. 2679–2693, 2014.
[13] S. Yaxley, E. T. Rolls, and Z. J. Sienkiewicz, "Gustatory responses of single neurons in the insula of the macaque monkey," *J. Neurophysiol.*, vol. 63, no. 4, pp. 689–700, 1990.
[14] E. T. Rolls, "Neural integration of taste, smell, oral texture, and visual modalities," in *Handbook of Olfaction and Gustation*, R. L. Doty, Ed., 3rd ed. Hoboken, NJ, USA: Wiley, 2015, pp. 1027–1047.
[15] J. X. Maier, M. Wachowiak, and D. B. Katz, "Chemosensory convergence on primary olfactory cortex," *J. Neurosci.*, vol. 32, no. 48, pp. 17037–17047, 2012.
[16] R. Q. Quiroga *et al.*, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 23, pp. 1102–1107, 2005.
[17] R. Q. Quiroga *et al.*, "Explicit encoding of multimodal percepts by single neurons in the human brain," *Current Biol.*, vol. 19, no. 15, pp. 1308–1313, 2009.
[18] R. Q. Quiroga, "Concept cells: The building blocks of declarative memory functions," *Nature Rev. Neurosci.*, vol. 13, no. 8, pp. 587–597, 2012.
[19] J. M. Fuster, "Network memory," *Trends Neurosci.*, vol. 20, no. 10, pp. 451–459, 1997.
[20] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. New York, NY, USA: Wiley, 1949.
[21] S. Löwel and W. Singer, "Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity," *Science*, vol. 255, no. 5041, pp. 209–212, 1992.
[22] C. J. Shatz, "The developing brain," *Sci. Amer.*, vol. 267, no. 3, pp. 60–67, 1992.
[23] M.-M. Mesulam, "From sensation to cognition," *Brain*, vol. 121, no. 6, pp. 1013–1052, 1998.
[24] J. J. Nassi and E. M. Callaway, "Parallel processing strategies of the primate visual system," *Nature Rev. Neurosci.*, vol. 10, no. 5, pp. 360–372, 2009.
[25] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: Anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.
[26] J. Hegdé and D. C. Van Essen, "Selectivity for complex shapes in primate visual area V2," *J. Neurosci.*, vol. 20, no. 5, p. RC61, 2000.
[27] E. Kobatake and K. Tanaka, "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex," *J. Neurophysiol.*, vol. 71, no. 3, pp. 856–867, 1994.
[28] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Rev. Neurosci.*, vol. 8, no. 5, pp. 393–402, 2007.
[29] G. L. Romani, S. J. Williamson, and L. Kaufman, "Tonotopic organization of the human auditory cortex," *Science*, vol. 216, no. 4552, pp. 1339–1340, 1982.
[30] N. Mesgarani *et al.*, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.
[31] C. Humphries *et al.*, "Syntactic and semantic modulation of neural activity during auditory sentence comprehension," *J. Cogn. Neurosci.*, vol. 18, no. 4, pp. 665–679, Apr. 2006.
[32] G. Auda and M. Kamel, "Modular neural networks: A survey," *Int. J. Neural Syst.*, vol. 9, no. 2, pp. 129–151, 1999.
[33] P. G. Schyns, "A modular neural network model of concept acquisition," *Cognit. Sci.*, vol. 15, no. 4, pp. 461–508, 1991.
[34] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.
[35] J. A. Anderson *et al.*, "Distinctive features, categorical perception, and probability learning: Some applications of a neural model," *Psychol. Rev.*, vol. 84, no. 5, pp. 413–451, 1977.
[36] K. Yamauchi, M. Oota, and N. Ishii, "A self-supervised learning system for pattern recognition by sensory integration," *Neural Netw.*, vol. 12, no. 10, pp. 1347–1358, 1999.
[37] T. Nakamura, T. Nagai, and N. Iwahashi, "Bag of multimodal LDA models for concept formation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Shanghai, China, May 2011, pp. 6233–6238.
[38] T. Araki *et al.*, "Autonomous acquisition of multimodal information for online object concept formation by robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Francisco, CA, USA, Sep. 2011, pp. 1540–1547.
[39] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, Washington, DC, USA, 2011, pp. 689–696.
[40] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, CA, USA, 2012, pp. 2222–2230.
[41] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 2141–2149.

[42] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, Oct. 2014.
[43] N. Parde *et al.*, "Grounding the meaning of words through vision and interactive gameplay," in *Proc. Int. Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 1895–1901.
[44] J. Thomason *et al.*, "Learning multi-modal grounded linguistic semantics by playing 'I spy,"' in *Proc. Int. Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 3477–3483.
[45] G. A. Carpenter and S. Grossberg, "The ART of adaptive pattern recognition by a self-organizing neural network," *Computer*, vol. 21, no. 3, pp. 77–88, Mar. 1988.
[46] X. He, R. Kojima, and O. Hasegawa, "Developmental word grounding through a growing neural network with a humanoid robot," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 2, pp. 451–462, Apr. 2007.
[47] S. Furao and O. Hasegawa, "An incremental network for on-line unsupervised classification and topology learning," *Neural Netw.*, vol. 19, no. 1, pp. 90–106, Jan. 2006.
[48] T. Jantvik, L. Gustafsson, and A. P. Papliński, "A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration," *Neural Comput.*, vol. 23, vol. 8, pp. 2101–2139, 2011.
[49] A. P. Papliński and W. M. Mount, "Bimodal incremental self-organizing network (BiSON) with application to learning chinese characters," in *Proc. Int. Conf. Neural Inf. Process.*, Daegu, Korea, Nov. 2013, pp. 121–128.
[50] A. P. Papliński, L. Gustafsson, and W. M. Mount, "A model of binding concepts to spoken names," *Austral. J. Intell. Inf. Process. Syst.*, vol. 11, no. 2, pp. 1–5, 2010.
[51] A. P. Papliński, L. Gustafsson, and W. M. Mount, "A recurrent multimodal network for binding written words and sensory-based semantics into concepts," in *Proc. Int. Conf. Neural Inf. Process.*, Shanghai, China, 2011, pp. 413–422.
[52] A. P. Papliński and W. M. Mount, "Transferring knowledge between learning systems," in *Proc. Int. Conf. Intell. Syst. Design Appl.*, Okinawa, Japan, Nov. 2014, pp. 119–123.
[53] L. Meng, A.-H. Tan, and D. Xu, "Semi-supervised heterogeneous fusion for multimedia data co-clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2293–2306, Sep. 2014.
[54] E. Persoon and K.-S. Fu, "Shape discrimination using Fourier descriptors," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, no. 3, pp. 170–179, Mar. 1977.
[55] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[56] T. Martinetz and K. Schulten, "Topology representing networks," *Neural Netw.*, vol. 7, no. 3, pp. 507–522, 1994.
[57] H. A. Simon, "The architecture of complexity" *Proc. Amer. Philos. Soc.*, vol. 106, no. 6, pp. 467–482, 1962.
[58] Y. Xing, F. Shen, and J. Zhao, "A perception evolution network for unsupervised fast incremental learning," in *Proc. Int. Joint Conf. Neural Netw.*, Dallas, TX, USA, Aug. 2013, pp. 297–304.
[59] Y. Xing, F. Shen, and J. Zhao, "Perception evolution network: Adapting to the emergence of new sensory receptor," in *Proc. Int. Conf. Artif. Intell.*, Buenos Aires, Argentina, Jul. 2015, pp. 3967–3973.
[60] Y. Xing, F. Shen, and J. Zhao, "Perception evolution network based on cognition deepening model—Adapting to the emergence of new sensory receptor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 607–620, Mar. 2016.
[61] Y. Xing, "Perception, coordination and evolution—Intelligence lies in structure and movement," Ph.D. dissertation, Dept. Comput. Sci. Technol., Nanjing Univ., Nanjing, China, 2016.



You-Lu Xing received the B.S. degree in computer science and technology from the East China University of Science and Technology, Shanghai, China, in 2009, and the M.S. degree in software engineering and the Ph.D. degree in computer software and theory from Nanjing University, Nanjing, China, in 2011 and 2016, respectively.

He is currently a Lecturer of computer science and technology with Anhui University, Hefei, China. His current research interests include artificial intelligence, robotics, and physiology.

Xiao-Feng Shi received the B.S. and M.S. degrees from Nanjing University, Nanjing, China, in 2014 and 2017, respectively. He is currently pursuing the Ph.D. degree with the University of California Santa Cruz, Santa Cruz, CA, USA.

His current research interests include sensor network, neural networks, and computer vision.

Fu-Rao Shen received the B.Sc. and M.Sc. degrees in mathematics from Nanjing University, Nanjing, China, in 1995 and 1998, respectively, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006.

He is currently a Full Professor of computer science and technology with Nanjing University. His current research interests include neural computing and robotic intelligence.

Jin-Xi Zhao received the Ph.D. degree in mathematics from Nanjing University, Nanjing, China, in 1987.

Since 1977, he has been with Nanjing University, where he was a Lecturer and an Associate Professor, in 1986 and 1988, respectively, and has been a Full Professor of computer science and technology since 1999. His current research interests include the intelligent computing, numerical linear algebra, and the theory and algorithm in computation of matrix, ill-conditioned system.

Jing-Xin Pan received the Master of Medicine degree from the Medical School, Nanjing University, Nanjing, China, in 2016.

He was enrolled in a 7-year MD/MS Program with Nanjing University. His current research interests include translational medicine, public health, anthropology, and various interdisciplinary research fields.

Ah-Hwee Tan (SM'04) received the B.Sc. (Hons.) and M.Sc. degrees in computer and information science from the National University of Singapore, Singapore, in 1989 and 1991, respectively, and the Ph.D. degree in cognitive and neural systems from Boston University, Boston, MA, USA, in 1994.

He was a Research Manager with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, spearheading the Text Mining and Intelligent Agents Research Programs. He is currently a Professor of computer science and the Associate Chair (Research) with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His current research interests include cognitive and neural systems, brain inspired intelligent agents, machine learning, knowledge discovery, and text mining.

Dr. Tan is an Editorial Board Member of the IEEE ACCESS and an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS.