doi:10.3772/j.issn.1006-6748.2016.01.013

CSMCCVA: Framework of cross-modal semantic mapping based on cognitive computing of visual and auditory sensations[①]

Liu Yang (刘 扬)[②] [*] [**] [***], Zheng Fengbin [*] [**], Zuo Xianyu [**] [****]
([*] Laboratory of Spatial Information Processing, Henan University, Kaifeng 475004, P. R. China)
([**] College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, P. R. China)
([***] College of Environment and Planning, Henan University, Kaifeng 475004, P. R. China)
([****] Institute of Data and Knowledge Engineering, Henan University, Kaifeng 475004, P. R. China)

Abstract

Cross-modal semantic mapping and cross-media retrieval are key problems of the multimedia search engine. This study analyzes the hierarchy, the functionality, and the structure in the visual and auditory sensations of cognitive system, and establishes a brain-like cross-modal semantic mapping framework based on cognitive computing of visual and auditory sensations. The mechanism of visual-auditory multisensory integration, selective attention in thalamo-cortical, emotional control in limbic system and the memory-enhancing in hippocampal were considered in the framework. Then, the algorithms of cross-modal semantic mapping were given. Experimental results show that the framework can be effectively applied to the cross-modal semantic mapping, and also provides an important significance for brain-like computing of non-von Neumann structure.

Key words: multimedia neural cognitive computing (MNCC), brain-like computing, cross-modal semantic mapping (CSM), selective attention, limbic system, multisensory integration, memory-enhancing mechanism

0 Introduction

Cross-modal semantic mapping (CSM) is the key technological issue of multimedia search engine. The challenges to the cross-modal semantic computing are dimensional. Multimedia search engine is an information retrieval technique with CSM, which can retrieval multimodal media that are similar and correlated in semantics from the network multimedia databases. CSM is a process that addresses how to find semantic objects from the same modal media similar in semantic and different modal media which are correlated in semantic based on the dimension reduction of the media features.

Essentially, CSM concern about multimedia computing (MC) issues. The MC's main objective is to research methods and theory of information collection, information representation and information presentation for vision, hearing, touch, taste, smell and other sensory media. It establishes a general computing theory and application techniques of transmission, processing, content analysis and recognition algorithm for representation media such as text, graphics, images, audio, MIDI, video, animation and so on. The development of CSM undergoes three stages: keyword-based text information retrieval, mono-modal media retrieval based on content similarity, and cross-media retrieval based on semantic correlation. The most popular CSM methods are the text co-occurrence and annotations model, the cross-media correlation graph, the semantic ontology model etc. The most related research focuses on low-level information described for high-dimensional indexing, high-level information semantic mining, and cross-modal and different dimension information correlation, as well as relevance feedback based on human-computer interaction for retrieval results performance promotion. Recently, research focuses on deep learning and statistical learning in CSM semantic-based. The multi-modal deep learning methods proposed to achieve cross-modal audio-video classification in Refs[1,2]. In order to resolve the cross media retrieval task, parallel field alignment for cross-media retrieval (PFAR)

[①] Supported by the National Natural Science Foundation of China (No. 61305042, 61202098), Projects of Center for Remote Sensing Mission Study of China National Space Administration (No. 2012A03A0939) and Science and Technological Research of Key Projects of Education Department of Henan Province of China (No. 13A520071).

[②] To whom correspondence should be addressed. E-mail: xyz3397681@163.com

Received on Feb. 2, 2015

method was introduced in Ref. [3], which integrated a manifold alignment framework from the perspective of vector fields to solve the semantic gap. Bi-directional cross-media semantic representation model (Bi-CM-SRM) was proposed in Ref. [4], which is a general cross-media ranking algorithm to optimize the bi-directional listwise ranking loss with a latent space embedding. A mechanism of cognitive science and neuroscience system structure is an important reference for the study of neural computing. It also greatly inspires multimedia intelligent analysis and information retrieval. However, there is difference essence in methods of research and realization between computer science and neurocognitive science, due to CSM complexity. So it is an urgent and important research issue that how to use knowledge of neurocognitive science to realize efficient models and algorithms.

Now a series of interlocking innovations in a set of two papers is unveiles to illuminate frameworks and algorithms of multimedia search engineer based on multimedia neural cognitive computing (MNCC) in two ways: cross-media semantic retrieval based on neural computing of visual and auditory sensations (CSRNC-VA) and cross-modal semantic mapping based on cognitive computing of visual and auditory sensations (CSMCCVA). In this paper, a set of algorithms and frameworks of CSMCCVA is presented with the function of cerebrum such as neurocognitive mechanism of visual-auditory collaborative, control structures of selective attention of thalamo-cortical, emotional control of limbic system and the memory-enhancing effects of hippocampal, and search brain-like computing based on neurocognitive function and structure. In second paper, a set of algorithms and models of CSRNCVA would be presented which originally sprang from ideal of deep belief network (DBN), hierarchical temporal memory (HTM) and probabilistic graphical model (PGM), and research brain-like computing based on neurocognitive function and structure.

1 Related work

According to related researches of cognitive science and neuroscience, cognitive processes and multisensory neurons of the human brain have cross-modal properties. The human brain is one of the most complex systems in nature, and brain-like computing is a simulation of human brain function and structure. Overall, now the brain science fails to achieve breakthrough in cerebrum advanced functions. No double, this causes tremendous challenge to research artificial brain by computer science. But the authors believe that it is entirely possible to create brain-like computable framework based on MNCC, if cognitive computing (CC) methods[5-9] based on cognitive information processing framework and neural computing (NC) methods[10-12] based on neural information processing mechanisms are used, which will be benefit for solving the problem of MC semantic-based. From the perspective of MNCC, CSM can be treated as the methods of classification and recognition of CC.

The CC's main objective is to explore the brain mental thoughts of cognitive information processing, including sensation, perception, attention, memory, language, thinking, awareness etc., and build brain-like computational framework and algorithm. It needs learning from experience to find relationship from different things, and to implement reasoned, memory and computing from logical principles. A theory for cognitive informatics (CI) based on abstract intelligence, denotational mathematics and algebraic system theory proposed in Ref. [13] which implements computational intelligence by autonomous inferences and perceptions of the brain, and presents a survey on the theoretical framework and architectural techniques of cognitive computing beyond conventional imperative and autonomic computing technologies. According to operationalized vast collections of neuroscience data by leveraging large-scale computing and simulations, the core algorithms of the brain are delivered to gain a deep scientific understanding of how the mind perceives, thinks, and acts. The novel cognitive systems, computing architectures, programming paradigms, practical applications, and intelligent business machines were proposed in Ref. [14].

The MNCC's main objective is to research the problems of semantic computing and dimensionality reduction for unstructured, massive, multi-modal, multi-temporal and spatial distribution of multimedia information processing, to establish a new generation of the multimedia information processing frameworks and algorithms which system behavior of CC is in macroscopic level and physiological mechanisms NC in microscopic level. Currently, there are two main aspects which have attracted much attention in brain-like computing. The first is to simulate cognitive function based on system behavioral, and the second is to research neural mechanisms based on structures of neurons, synapses, or local networks. However, there are still lacks of effective methods about how to build advanced system of complex function with simple local neural networks. The researchers have made unremitting exploration in the mechanisms brain-like computing for a long time. The main research directions include artificial neural network (ANN), HTM, DBN, PGM and so on.

Since McCulloch and Pitts modeled a simple neural network using electrical circuits in 1943, after experiencing setbacks, ANN's research has returned to the track of normal development. Rosenblatt proposed perceptron is an essentially linear classifier in 1958. Multi-layer perceptron (MLP) can solve the problem of non-linear, since using back propagation algorithm it often suffers from local minima when ANN hidden layers number increased, and is difficult to solve the problem of under fitting or over fitting. After that, many feedback neural networks were developed, such as Hopfield and SOM network in 1982, ART network of Grossberg in 1988 etc. For some complex cognitive phenomena such as associative, memory can be used feedback neural networks to simulation and interpretation. Furthermore, some new ANN appears such as neocognitron[15], spiking neural network[16], convolutional network[17], hierarchical model and X (HMAX) model etc. Vapnik put forward support vector machines (SVM) in 1992, which deals with linearly nonseparable problems using kernel tricks. SVM is a special double-layer ANN which has efficient learning algorithm. ANN and SVM are classical methods of statistical machine learning theories. In neuroscience, the structure of a biological neural network (sometimes called a neural pathway) is directed, loop network with characteristics of feedback and temporal. Neural circuit is a functional entity of interconnected neurons that is able to regulate its own spiking activity using a feedback loop. Training methods of ANN can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The essence of ANN learning can be understood as dynamics problems from system perspective, and can also be apprehend as optimization problems from functional aspect. It is needed to simulate not only the cognitive function but also the structure of the nervous system.

Hinton points out that one cannot solve complex problems in machine learning only with simple supervised learning methods. He proposed that one should aim to establish a structure by generative models of neural network, and one should research the ANN affections mechanism of inner representation from external environment. Hinton presents a restricted Boltzmann machine (RBM) of a double layer structure and Bayesian belief networks to configure DBN in 2006. Both cortical structure and cognitive processes are deep architecture, and a fast deep learning algorithm for DBN was proposed in Refs[18,19]. To DBN train, the best results obtained from supervised learning tasks involve an unsupervised learning component, usually in an unsupervised pre-training phase. Deep learning algorithm has achieved unprecedented results in many applications; Microsoft research found that relative error reduced to 33% for large-vocabulary speech recognition in Switchboard dataset[20], and Google labs also found that accuracy of recognizing object categories increased to 70% more than current best result in ImageNet dataset[21].

From the probabilistic perspective, ANN is also a graph model problem. PGM itself has complete theoretical system, and can be used in complex field for uncertain reasoning and information analysis with probabilistic and statistical theories[22,23]. PGM is divided between the undirected graphs (for example Markov random field) and directed graphs (for instance Bayesian network). Each node only connects with other limited nodes in PGM. PGM has properties of locality principle, sparse and small-word. The most common methods of approximate reasoning in PGM are Markov chain Monte Carlo (MCMC) algorithm (such as Gibbs sampling) and belief propagation algorithm[24].

In contrast, between Hinton's deep learning algorithm for DBN with RBM and Hawkins's cortical algorithm for HTM[25], both can be classified by unsupervised learning, and can be stacked to build up feedback hierarchy structure. RBM doesn't fully utilize the spatial-temporal locality, but HTM is even more modular to use spatial-temporal locality and hierarchical based on belief propagation algorithm of PGM. Both DBN and HTM can be seen as a special case in mathematical formalism.

2 Visual-auditory information collaborative of semantic mapping

2.1 Visual-auditory collaborative cognitive mechanisms

The human central nervous system has white matter, grey matter, substantia nigra and other tissue. On the one hand, neocortex's function in grey matter is structurally similar to the processing unit in linear analogue systems and gate circuit in nonlinear digital system. On the other hand, long-distance pathways (LDP) in white matter construct complexes wiring diagrams of neural networks of information processing.

Function and structure of the cerebrum are one of the most complex systems in nature. It is generally thought that neocortex of cerebrum is an important part of processing logical intelligent, thalamus is the switching of selective attention which controls the information pass in and out, and hippocampus and the limbic system are controllers of memory and emotional. Now, a variety of methods is used to explore the brain mechanism. Neuroscience uses white box and bottom-up methods to research neural information processing mechanism of cortical structures and neural pathways. Cognitive science uses methods of black box and top-down to analyzed function and phenomenon of cognitive, then builds brain information processing model in theory. Computer science implements mathematical logic operation on finite state machine based on Turing machine model. From MLP model to HTM and DBN, people never stop exploring the use of cognitive processing mechanisms of the nervous system to promote complex information computing. It is thought that the neural system and cognitive function belong to isomorphic relationship.

Theorem 1 By establishing related computing model M, it can build mapping between the neural structures (or processes) Φ and cognitive operations (or functions):

$$M: \Psi \leftrightarrow \Phi \tag{1}$$

In addition, there are a lot of connections in the thalamo-cortical projection systems. There are three basic types of thalamic nuclei: i) relay nuclei; ii) association nuclei; and iii) nonspecific nuclei. Relay nuclei receive very well defined inputs and project this signal to functionally distinct areas of the cerebral cortex. The association nuclei are the second type of thalamic nuclei and receive most of their input from the cerebral cortex and project back to the cerebral cortex in the association areas where they appear to regulate activity. The third type of thalamic nuclei are the nonspecific nuclei, including many of the intralaminar and midline thalamic nuclei that project quite broadly through the cerebral cortex, may be involved in general functions such as alerting. According to the neurocognitive mechanism, Fig. 1 illustrates neurocognitive pathway of visual-auditory information collaborative processing in thalamo-cortical projection system which has 5 layers. Layer 1 is visual-auditory feature detectors such as brightness, edge, tone and loudness. Thalamus is located at layer 2, which is the center of information switching. The neocortex consists of 3 layers. Layer 3 constructs super-column to mimicking primary visual-auditory sensory cortex. Layer 4 imitates multi-modal sensory of secondary visual-auditory cortex. Layer 5 simulates association cortex.

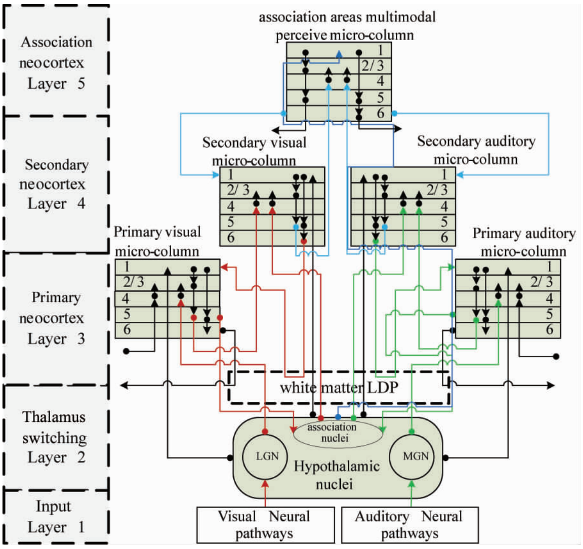

Association neocortex Layer 5
Secondary neocortex Layer 4
Primary neocortex Layer 3
Thalamus switching Layer 2
Input Layer 1
association areas multimodal perceive micro-column
Secondary visual micro-column
Secondary auditory micro-column
Primary visual micro-column
Primary auditory micro-column
1
2/3
4
5
6
white matter LDP
association nuclei
LGN
Hypothalamic nuclei
MGN
Visual Neural pathways
Auditory Neural pathways

Fig. 1 Schematic of thalamo-cortical projection system of the visual-auditory neurocognitive collaborative information processing pathway

According to brain's structure and cognitive processes, Fig. 2 describes the visual-auditory collaborative cognitive information processing schematic. Rectangular nodes stand for hierarchical probability network for memory, inference, and logic control. The hexagonal nodes stand for payoff network for in mood, emotion control and optimize control. There are three main modules in this schematic. The emotion control module consists of limbic lobe, insula lobe, basal ganglia and thalamus. The visual-auditory perception module consists of occipital lobe, temporal lobe and thalamus. The visual-auditory cognitive module consists of frontal lobe, parietal lobe and thalamus.

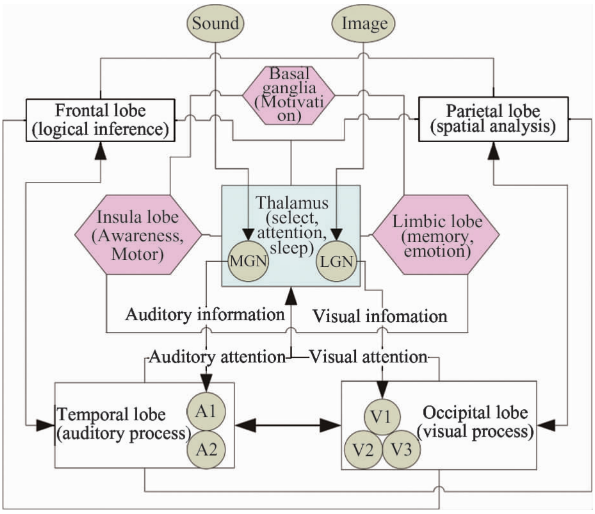

Sound
Image
Basal ganglia (Motivati on)
Frontal lobe (logical inference)
Parietal lobe (spatial analysis)
Insula lobe (Awareness, Motor)
Thalamus (select, attention, sleep)
MGN
LGN
Limbic lobe (memory, emotion)
Auditory information
Visual infomation
Auditory attention
Visual attention
Temporal lobe (auditory process)
A1
A2
V1
V2
V3
Occipital lobe (visual process)

Fig. 2 Visual-auditory collaborative cognitive information processing process schematic

2.2 CSM framework

Fig. 3 shows the structure of CSM framework. The principles of hierarchical reinforcement and incremental training are used to the framework of semantic mapping. CSM framework can be decomposed visual-auditory multisensory integration, attention control in thalamus, emotional control in limbic system and the memory mechanism in hippocampal. For the convenience of description, some concepts and processing are defined as follows:

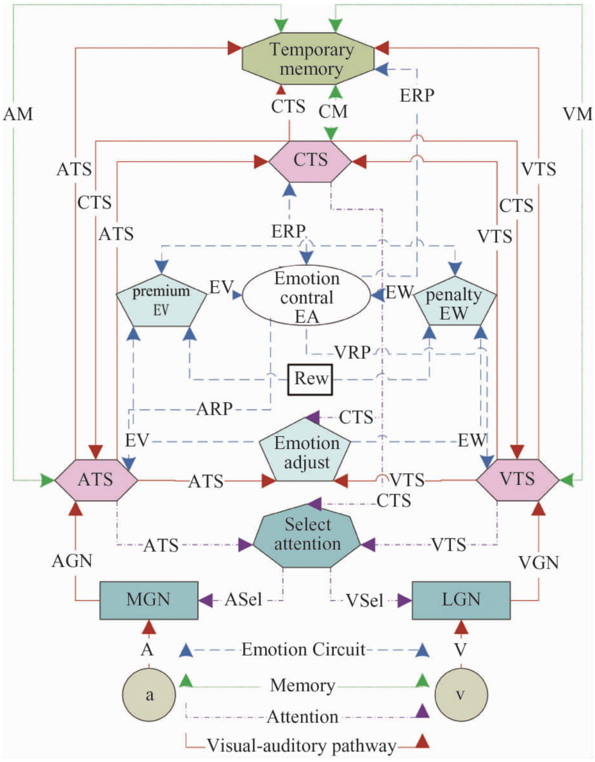


Fig. 3 Visual-auditory cross-modal semantic mapping framework based on MNCC

Definition 1 Semantic of visual and auditory media $\boldsymbol{v}_t$ and $\boldsymbol{a}_t$ are denoted by normalized raw data $\boldsymbol{V}$ and $\boldsymbol{A}$. Video V is defined as empty by space-frequency distribution according to temporal sampling, and $\boldsymbol{A}$ is defined as time-frequency distribution by short-time Fourier transform, where $\boldsymbol{a}_t$ is a one-dimensional array and $\boldsymbol{v}_t$ a two-dimensional matrix.

$$\begin{cases} \boldsymbol{V} = \mathrm{FFT}(\boldsymbol{v}_t) \\ \boldsymbol{A} = \mathrm{STFT}(\boldsymbol{a}_t) \end{cases} \tag{2}$$

Definition 2 Visual temporal-spatial patterns probability VGN of visual media $\boldsymbol{V}$ and auditory temporal-spatial patterns probability AGN of auditory media $\boldsymbol{A}$ can be defined as

$$\begin{cases} \mathrm{AGN}(t,s) = P(\boldsymbol{A})(1 + \mathrm{ASel}(t,s)) \\ \mathrm{VGN}(t,s) = P(\boldsymbol{V})(1 + \mathrm{VSel}(t,s)) \end{cases} \tag{3}$$

while ***VSel*** and ***ASel*** see also Definition 3.

Definition 3 Both visuals objects attention values $VSel$ and auditory objects attention value $ASel$ are described by selection values of thalamus which can be calculated by temporal-spatial compactness of neighboring media. The more salient the objects, the greater probability by thalamus selected. Both $VSel$ and $ASel$ can be defined as

$$\begin{cases} \mathrm{ASel}(t,s) = \left(1 - \dfrac{(t-at)^2 + (s-as)^2}{CTS + VTS + ATS}\right) e^{-\frac{(t-at)2+(s-as)2}{CTS+VTS+ATS}} ATS \\ \mathrm{VSel}(t,s) = \left(1 - \dfrac{(t-vt)^2 + (s-vs)^2}{CTS + VTS + ATS}\right) e^{-\frac{(t-vt)2+(s-vs)2}{CTS+VTS+ATS}} VTS \end{cases} \tag{4}$$

where vt, vs, as, at are spatial and temporal parameters of visual and auditory respectively. VTS, ATS and CTS see also Definition 4.

Definition 4 Both auditory cortex belief ATS, visual cortex belief VTS and concept of visual-auditory integration belief CTS are used to describe certainty factor of cortical column for media objects. It can be defined as

$$ATS = ARP \sum_{N = AGN, CTS, AM} \mathrm{Bel}(N) \tag{5}$$

$$VTS = VRP \sum_{N = VGN, CTS, VM} \mathrm{Bel}(N) \tag{6}$$

$$CTS = \mathrm{ERP}(\mathrm{Bel}(CM) + ATS + VTS) \tag{7}$$

where Bel is the belief of cortical column. For VRP, ARP and ERP also see Definition 5. For meaning of VM, AM, and CM also see Definition 6.

Definition 5 Visual object emotional control value VRP, auditory object emotional control values ARP and visual-auditory object emotional control value ERP are applied to describe emotion control for reward of amygdala and orbitofrontal cortex in limbic system. They can be defined as

$$r_t = x_t EV_t - \frac{EW_t}{EV_t} + EA_m \tag{8}$$

At time t, let values of x be ATS, VTS and CTS, then values of r will be ARP, VRP and ERP in sequence. They influence and control each other, and compose emotion circuit of amygdala and orbitofrontal cortex in the limbic system; moreover, they can also influence hippocampal to temporary memory information[26,27], where EV is premium value, EW is penalty value, and EA is emotion value. They can be defined as

$$\begin{cases} EA_m = \mathrm{Max}(x_t)\mathrm{Max}(AGN, VGN) \\ EV_t = EV_{t-1} + \alpha(x_{t-1}\mathrm{Max}(0, Rew - x_{t-1}EV_{t-1})) \\ EW_t = EW_{t-1} + \beta\left(x_{t-1}\left(x_{t-1}EV_{t-1} - \dfrac{EW_{t-1}}{EV_{t-1}} - Rew\right)\right) \end{cases} \tag{9}$$

where Rew is reward value, α and β are coefficients.

Definition 6 Research findings related to the long-term memory, the presynaptic and postsynaptic activity of the hippocampus vary with the phenomenon of long term potentiation (LTP) and long term depression (LTD). According to the spike-timing dependent plasticity (STDP) theory, visual objects temporal memory value *VM*, auditory object temporal memory value *AM* and visual-auditory object temporal memory value *CM* in hippocampus are defined as

$$Mem = \begin{cases} -x(1-ERP^2)e^{\left(-\frac{\Delta t}{5}\right)} & \Delta t \geqslant 0 \\ x(1-ERP^2)e^{\left(\frac{\Delta t}{10}\right)} & \Delta t < 0 \end{cases} \tag{10}$$

where Δt is the difference between the current time and the start time of memory object x, when values of *Mem* are *AM*, *VM* and *CM*, then values of x are *ATS*, *VTS* and *CTS* in sequence.

2.3 CSM algorithm of auditory-visual information

The cross-modal semantic mapping algorithm (CSMA) transforms media **A** and **V** into a set of CSM parameters. CSMA includes two steps: pre-learning algorithm in waking state (PLAW) and precisely adjust algorithm in sleeping state (PAAS).

Algorithm 1 Cross-media semantic mapping algorithm (CSMA)

Input: media **A** and **V**.
Output: CSM parameters
Procedure:

(1) Call PLAW to calculate and get initial CSM parameters such as *AGN*, *VGN*, *ATS*, *VTS*, *CTS*, *ARP*, *VRP*, *ERP*, *AM*, *VM* and *CM*;

(2) Call PAAS to modify and optimize CSM parameters.

PLAW mimic cognitive function is controlled by emotion and memory under the waking state. It is unsupervised training and using bottom-up and to preprocess input information **A** and **V** step by step, and generate a set of CSM initial parameters. The PLAW can be expressed as follows:

Algorithm 2 Pre-learning algorithm in waking state (PLAW)

Input: media **A** and **V**.
Output: CSM initial parameters.
Procedure:

(1) This step is information preprocess. According to Eq. (3), calculate visual temporal-spatial patterns probability *VGN*, auditory temporal-spatial patterns probability *AGN* from visual media **V** and auditory media **A**;

(2) According to Equation 4, calculate visual objects attention values *VSel* and objects attention value *ASel* by thalamus;

(3) Set *CTS*, *ARP*, *VRP*, *AM*, and *VM* initials equal 0. Call micro-column information propagation algorithm (MIPA) to calculate auditory cortex belief *ATS* and visual cortex belief *VTS*;

(4) Let *CM* initials equal 0, According to Equation 7, calculate concepts of visual-auditory integration belief *CTS*;

(5) if *ATS*, *VTS* or *CTS* values are similar with the owner storage pattern values in node of cortical column, then set *AM*, *VM* or *CM* equal 0; otherwise, calculate *AM*, *VM* and *CM* according to Equation 10 by *ATS*, *VTS* and *CTS*, then feedback information to input node cortical columns.

PAAS mimics cognitive function of sleeping state, when thalamus closes the input information **A** and **V** input parameters have *AM*, *VM*, *CM* in framework temporary memory and other CSM initial parameters, and output *VRP*, *ARP* and *ERP*, and *CSM* parameters after the optimization. It is supervised training, and top-down adjusts and optimizes internal parameters with hierarchical reinforcement learning strategies under the memory and emotional control, where reward function of the limbic system is designed by the "principle of lowest energy" E and "maximizing benefit" M of the system. That is, rewarding successes and punishing failure. PAAS process as follows:

Algorithm 3 Precisely adjust algorithm in sleeping state (PAAS)

Input: AM, VM, CM in framework temporary memory and other CSM initial parameters.
Output: VRP, ARP and ERP, and CSM parameters after the optimization.
Procedure:

(1) Set current time $t = 0$, calculate *AM*, *VM* and *CM* to mimicking hippocampal memory according to Equation 10. Then let hippocampal feedback number is n when time is t.

(2) When values for x are *ATS*, *VTS* and *CTS*, then let values for π equal *ARP*, *ERP* and *VRP* in sequence. Belief of cortical columns nodes x can be calculate as follows:

$$p(x) = (1 + e^{\Delta E_x / T})^{-1} \tag{11}$$

where T is "temperature" parameter of belief. ΔE_x is "energy" parameter of belief. The information propagation and adjust strategies as follows:

$$\pi : x \rightarrow \Delta E_x \tag{12}$$

(3) According to framework performance, calculate r_t by Eq. (8) to get *VRP*, *ARP* and *ERP*.

(4) Calculate benefit of *ATS*, *VTS* and *CTS* as follows:

$$M(x_k) \leftarrow M(x_k) + \alpha[r_{k+1} + \gamma M(x_{k+1}) - M(x_k)] \tag{13}$$

where values for x equal *ATS*, *VTS* and *CTS* in sequence; $\gamma$is discounting coefficient, $\gamma \in (0,1]$. r_t is receive rewards and penalties form $x_t$to x_{t+1}, andα is learning rate.

(5) Calculate framework whole "energy"

$$\Delta E_x = \sum_j p(x_j) x_j \tag{14}$$

Search and choose the best optimal decisionπ^*:

$$\pi^* = \arg_\pi \max \frac{M(x)}{\Delta E_x} \tag{15}$$

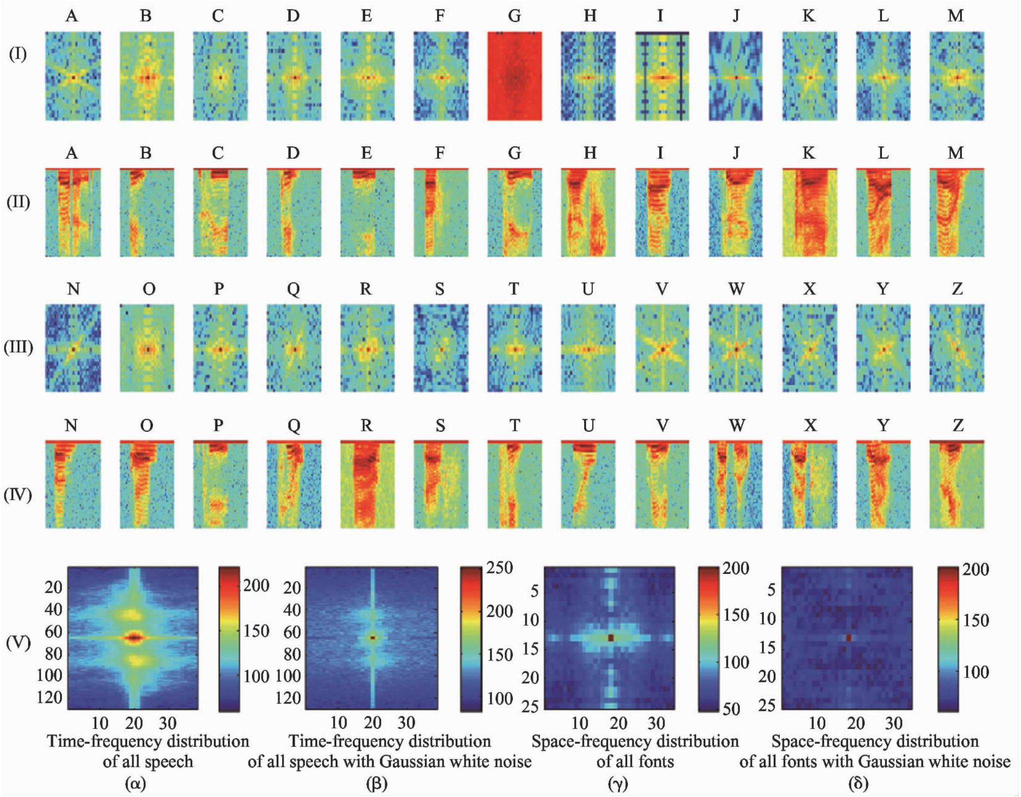
$\forall x \in ATS, CTS, VTS$ and $\forall \pi \in ARP, ERP, VRP$

If $t < n$ and framework don't meet the condition, then let $t = t + 1$, return to step (2), until framework meet the condition or t reach feedback count n of hippocampal at time t.

3 Experiment and simulation

3.1 Experimental data

Since to the nature of audio and video information has a lot of uncertainty noises, in order to take the quantitative and qualitative analysis and evaluation to the modal, 26 letters in the English alphabet are adopted as train media, that are concepts of 26 English letters, pronunciation of Microsoft TTS Anna of 26 English letters, and image of Chinese Kai font of 26 English uppercase letters. Fig. 4 shows all train and test media: spectral distribution of English letters fonts (Ⅰ and Ⅲ rows in Fig. 4), spectral distribution of English letters pronunciation (Ⅱ and Ⅳ rows in Fig. 4); train media time-frequency distribution of all speech (Ⅴ row and α column in Fig. 4); test media time-frequency distribution of all speech with Gaussian white noise (Ⅴ row and β column in Fig. 4); train media space-frequency distribution of all fonts (Ⅴ row and γ column in Fig. 4); test media space-frequency distribution of all fonts with Gaussian white noise (Ⅴ row and δ column in Fig. 4). All of the training and testing media were transformed totaling 333 dimensions after processing. Where image frequency-domain is 25 dimensions and time-domain is 50 dimensions, audio frequency-domain is 111 dimensions and time-domain is 222 dimensions.



Time-frequency distribution of all speech (α)

Time-frequency distribution of all speech with Gaussian white noise (β)

Space-frequency distribution of all fonts (γ)

Space-frequency distribution of all fonts with Gaussian white noise (δ)

Fig. 4 Time-frequency and space-frequency distribution of the training media

3.2 CSMCCVA simulation

(1) Simulation and computing for selection attention of thalamo-cortical projection system

Due to the temporal-spatial characteristics of visual-auditory neural spiking activity (for instance duration of vision is 0.25s, and duration of auditory is 0.1s), temporal-spatial parameters are set $vs = 7$pixel, $vt = 250$ms, $as = 5$Hz and $at = 100$ms. Fig. 5 shows the simulation result for selection attention of thalamo-cortical projection systems according to Eq. (4). When a temporal-spatial media object is selected by the framework, the neighboring objects at the near time and space will be inhibited, which is similar to the selection attention mechanism of temporal-spatial compactness of cerebrum.

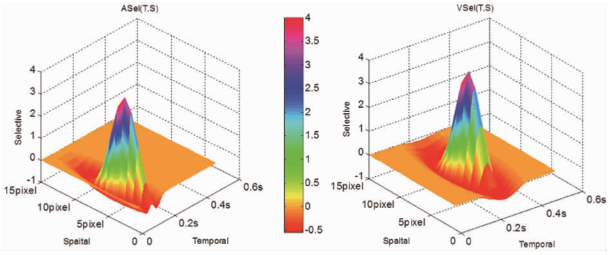


Fig. 5 Simulation and calculate for selective attention of thalamo-cortical projection system

(2) Simulation and computing for emotional controls of limbic system

To validate the emotional controls of framework to learning, let values forα, β, *AGN*, *VGN*, *ATS* and *VTS* equal 1, set *Rew* = 2, *t* = 20 and *CTS* = 1 ~ 5. Fig. 6 illustrates the *ERP* decaying with time according to Eq. (8). The figure shows that visual-auditory integration belief *CTS* is found to increase with visual-auditory object emotional control value *ERP*, and then it can adjust learning and memory ability of framework.

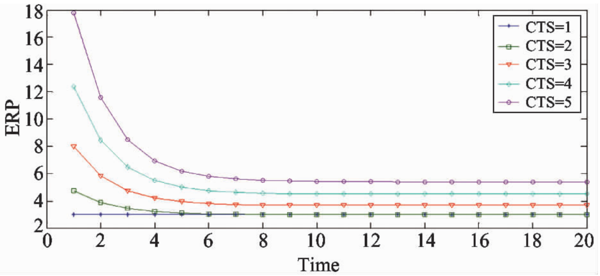


Fig. 6 Emotional control of the limbic system computing and simulation

(3) Simulation and computing for memory controls of hippocampus

Fig. 7 illustrates the memory control simulation results of hippocampus according to Eq. (10). According to the STDP theory, pre-memory can interfere to the near learning memory, the interfere increases with the time interval decreases when it's waking state. In addition, the memory is exponential decay with time when it's in sleeping state. This shows that the framework meets the basic law of the memory of the cerebrum.

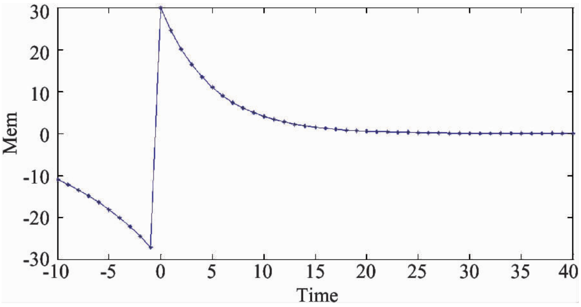


Fig. 7 The memory control computing and simulation of hippocampus

3.3 CSM result and analysis

To verify the CSM performance of the framework, Fig. 8 illustrates CSM result of CSMVACC in 26 letters test dataset in English alphabet. Gaussian noise has a great influence on CSMCCVA from Fig. 8, the accuracy of the framework tends to over 90% of the stable region when SNR is 30dB. The same low SNR noise has more influence power to auditory media CSM than visual media CSM. Performance of cross-modal CSM of auditory and visual media are better than CSM of monomodal media in the same high SNR background noise, which matches also the cognition law of the cerebrum.

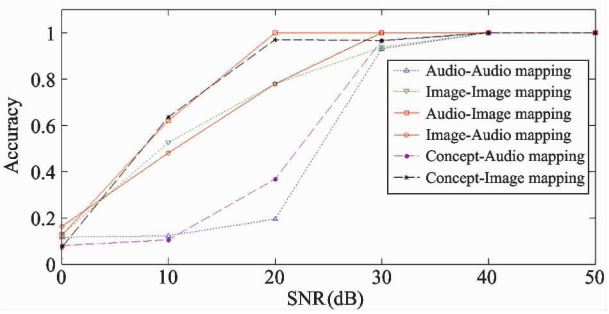


Fig. 8 Gaussian noise effects to CSM performance of the CSMCCVA

4 Conclusion

In this study, framework of CSMCCVA with mechanisms of central nervous systems is presented such as neurocognitive visual-auditory multi-sensory integration, selective attention in thalamo-cortical, emotional control in limbic system and the memory-enhancing effects of hippocampal. Then the semantic mapping algorithms is given. Simulation results show that this framework is robust and effective to CSM in experiments of text, speech and script. Only a preliminary exploration is done with MNCC, the framework's parameters only learn from physiological data, which is due to neurocognitive mechanisms of the brain complexity. Looking for the future, it is necessary to combine with deep learning theory, probability theory and modern cognitive findings to improve the relevant algorithms, and build cross-modal semantic search engine based on MNCC.

References

[1] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, USA, 2011. 689-696

[2] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. In: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe: Curran Associates, 2012. 2231-2239

[3] Mao X, Lin B, Cai D, et al. Parallel field alignment for cross media retrieval. In: Proceedings of the 21st ACM International Conference on Multimedia (MM'13), Barcelona, Spain, 2013. 21-25

[4] Wu F, Lu X, Zhang Z, et al. Cross-media semantic representation via bi-directional learning to rank. In: Proceedings of the 21st ACM International Conference on Multimedia (MM' 13), Barcelona, Spain, 2013. 877-886

[5] Preissl R, Wong TM, Datta P, et al. Compass: A scalable simulator for an architecture for cognitive computing. In: Proceedings of the High Performance Computing, Networking, Storage and Analysis (SC), Almaden, San Jose, USA, 2012. 1-11

[6] Shaw B, Cox A, Besterman P, Minyard J, et al. Cognitive computing commercialization: boundary objects for communication. In: Proceedings of the 3rd International Conference on Integration of Design, Engineering, and Management for Innovation, Porto, Portugal, 2013. 4-6

[7] Esser S K, Andreopoulos A, Appuswamy R, et al. Cognitive computing systems: algorithms and applications for networks of neurosynaptic cores. In: Proceedings of the International Joint Conference on Neural Networks, Dallas, USA, 2013. 1-10

[8] Amir A, Datta P, Cassidy A S, et al. Cognitive computing programming paradigm: a corelet language for composing networks of neurosynaptic cores. In: Proceedings of the International Joint Conference on Neural Networks, Dallas, USA, 2013. 1-10

[9] Cassidy A S, Merolla P, Arthur J V, et al. Cognitive computing building block: a versatile and efficient digital neuron model for neurosynaptic cores. In: Proceedings of the International Joint Conference on Neural Networks, Dallas, USA, 2013. 1-10

[10] Modha D S, Singh R. Network architecture of the long-distance pathways in the macaque brain. *Proceedings of the National Academy of Sciences*, 2010, 107 (30): 13485-13490

[11] Hagmann P, Cammoun L, Gigandet X, et al. Mapping the structural core of human cerebral cortex. *PLOS Biology*, 2008, 6(7):1479-1493

[12] Eliasmith C, Stewart T C, Choo X, et al. A large-scale model of the functioning brain. *Science*, 2012, 338 (6111):1202-1205

[13] Wang Y. On cognitive informatics. *Brain and Mind*, 2003, 4(2):151-167

[14] Modha D S, Ananthanarayanan R, Esser S K, et al. Cognitive computing. *Communications of the ACM*, 2011, 54 (8):62-71

[15] Fukushima K. Neocognitron trained with winner-kill-loser rule. *Neural Networks*, 2010, 23(7): 926-938

[16] Sterne P. Information recall using relative spike timing in a spiking neural network. *Neural Computation*, 2012, 24 (8):2053-2077

[17] Turaga S C, Murray J F, Jain V, et al. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Computation*, 2010, 22(2): 511-538

[18] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, 18(7):1527-1554

[19] Arel I, Rose D C, Karnowski T P. Deep machine learning-a new frontier in artificial intelligence research. *Computational Intelligence Magazine*, *IEEE*, 2010, 5(4): 13-18

[20] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2012, 20(1):30-42

[21] Le Q V, Ranzato M A, Monga R, et al. Building high-level features using large scale unsupervised learning. In: Proceedings of the 29th International Conference on Machine Learning (ICML), 2012, Edinburgh, UK, 2012. 81-88

[22] Blei D M. Probabilistic topic models. *Communications of the ACM*, 2012, 55(4): 77-84

[23] Liu L, Zhu F, Zhang L, et al. A probabilistic graphical model for topic and preference discovery on social media. *Neurocomputing*. 2012, 95:78-88

[24] Sun S L. A review of deterministic approximate inference techniques for Bayesian machine learning. *Neural Computing and Applications*. 2013. 23:2039-2050

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

[26] Ziemke T, Lowe R. On the role of emotion in embodied cognitive architectures: from organisms to robots. *Cognitive Computation*. 2009. 1(1):104-117

[27] Laurent P A. A neural mechanism for reward discounting: insights from modeling hippocampal-striatal interactions. *Cognitive Computation*. 2013. 5(1):152-160

Liu Yang, born in 1971, He is currently a Ph. D. candidate of Henan University. He is an associate professor and master student supervisor of college of computer science and information engineering, Henan University. He received M. S. degrees from Henan University in 2009. He received his B. S. degrees in Changchun University of science and technology (now it merged into faculty of science, Jilin University) in 1996. His research interests include multimedia neural cognitive computing, temporal and spatial information high-performance computing.