CSRNCVA: A MODEL OF CROSS-MEDIA SEMANTIC RETRIEVAL BASED ON NEURAL COMPUTING OF VISUAL AND AUDITORY SENSATIONS

Y. Liu,[*] *K. Cai,*[†] *C. Liu,*[‡] *F. Zheng*[§]

Abstract: Cross-media semantic retrieval (CSR) and cross-modal semantic mapping are key problems of the multimedia search engine. The cognitive function and neural structure for visual and auditory information process are an important reference for the study of brain-inspired CSR. In this paper, we analyze the hierarchy, the functionality and the structure of visual and auditory in the brain. Considering an idea from deep belief network and hierarchical temporal memory, we presented a brain-inspired intelligent model, called cross-media semantic retrieval based on neural computing of visual and auditory sensation (CSRNCVA). Algorithms based on CSRNCVA were developed. It employs belief propagation algorithms of probabilistic graphical model and hierarchical learning. The experiments show that our model and algorithms can be effectively applied to the CSR. This work provides an important significance for brain-inspired cross-media intelligence framework.

Key words: *cross-media cognitive neural computing, cross-media semantic retrieval, deep belief network, hierarchical temporal memory, probabilistic graphical model, hierarchical reinforcement learning*

Received: August 8, 2016 **DOI:** 10.14311/NNW.2018.28.018
Revised and accepted: August 10, 2018

1. Introduction

Cross-media semantic retrieval (CSR) is key technological issues of multimedia search engine. The challenges to the cross-modal semantic computing are the

[*] Yang Liu – Corresponding author; Key Laboratory of Big Data Analysis and Processing of Henan Province, Intelligent Technology and Application Engineering Research Centre of Henan Province, and College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, P. R. China, E-mail: ly.sci.art@gmail.com

[†] Kun Cai; Intelligent Technology and Application Engineering Research Centre of Henan Province, and College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, P. R. China, E-mail: caikun@henu.edu.cn

[‡] Chun Liu; Key Laboratory of Big Data Analysis and Processing of Henan Province, and College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, P. R. China, E-mail: liuchun@henu.edu.cn

[§] Fengbin Zheng – Corresponding author; Engineering Laboratory of Spatial Information of Henan Province, and College of Computer Science and Information Engineering, Henan University, Kaifeng 475004, P. R. China, E-mail: zhengfb@henu.edu.cn

semantic gap and the curse of dimensionality. Multimedia search engine and cross-media intelligent engines utilize CSR technique. CSR is the new information retrieval technique that finds the mono-modal media which are semantic similarity and the multi-modal media which are semantic correlation from the unstructured information of multimedia by cross-media intelligent techniques. Essentially, CSR concerns the multimedia computing issue. Multimedia computing of the main objective is to research the methods and theories of information collection, representation and analysis for vision, hearing, touch, taste, smell and other sensory media. It establishes computational theory and information processing, semantic analysis and target recognition algorithm for text, graphics, images, audio, MIDI, video, animation, and other representation media.

The development of CSR undergoes the three stages: keyword-based text information retrieval, mono-modal media retrieval based on content similarity, and CSR based on semantic correlation. Cross-media analysis and reasoning would play the important role in new-generation artificial intelligence [1]. CSR-related research mostly concerned low-level information described for high-dimensional indexing, high-level information semantic mining, and cross-modal and different dimensions information correlation, as well as relevance feedback based on human-computer interaction for retrieval results performance promotion. Recently, research mostly has focused on deep learning and statistical learning in CSR semantic-based. Multi-modal deep learning methods were proposed in order to achieve cross-modal audio-video classification [2, 3]. Document [4] proposed an unsupervised method called convolutional cross Autoencoder for cross-modality element-level feature learning, which can capture the cross-modality correlations in element samples of social media datasets. The reference [5] proposed a multi-modality fusion framework and a topic recovery approach to effectively detect topics from cross-media data. Reference [6] proposed a modality-dependent cross-media retrieval model, where two couples of projections are learned for different CMR tasks instead of one a couple of projections. To fully understand users' sentiment, literature [7] proposed a cross-media public sentiment analysis system for short text and image of microblog. Document [8] proposed an effective cross-media distance metric learning framework based on sparse feature selection and multi-view matching. Moreover, cross-media active learning algorithm also was used to reduce the effort on labeling images for training [9].

The neurocognitive function and structure are an important reference for the study of neural computing. It also has an important inspiration for multimedia intelligence analysis and information retrieval. However, there is fundamentally different in methods of research and realization between computer science and neurocognitive science, which is due to CSR complexity. So it is an urgent and important research issue that how to use knowledge of neurocognitive science to model and realize efficient algorithms.

We now unveil a series of interlocking innovations in a set of two papers to illuminate models and algorithms of multimedia search engine in the two ways: Cross-modal Semantic Mapping based on Cognitive Computing of Visual and Auditory sensation (CSMCCVA) based on Multimedia Neural Cognitive Computing (MNCC) [10], and Cross-media Semantic Retrieval based on Neural Computing of Visual and Auditory sensation (CSRNCVA) based on Cross-media Cognitive

Neural Computing (CCNC). In order to address the key problem of CSR, neurocognitive function and structure were researching into brain-inspired computing. In this paper, we present a set of algorithms and models of CSRNCVA, which originally sprang from idea of Deep Belief Nets (DBN), Hierarchical Temporal Memory (HTM) and Probabilistic Graphical Model (PGM).

2. Related works

The relevant researches of cognitive science and neuroscience found that multisensory neurons cognition of the environment is through the fusion of multiple sensory organs in the brain. The human brain is one of the most complex systems in nature. Brain-inspired computing is a simulation of the human brain in function and structure. Overall, now brain science fails to achieve the breakthrough in cerebrum advanced functions. No doubts, this causes tremendous challenge to the research of brain-inspired computing in information science. But we believe that it is entirely possible to create CCNC model based on brain-inspired intelligence, if cognitive computing methods [11, 12] based on cognitive framework, and neural computing methods based on neural processing mechanisms were used, such as formal concept analysis for cognitive functions [13], deep learning features for CMR [14], heterogeneous similarity measure with nearest neighbors and cross-media correlation propagation [15]. This will be beneficial for solving the problem of semantics-oriented multimedia computing.

The neural computing's main objective is to discover the mechanism of biological nervous systems, to mimic the mechanism of neural network structure. It constructed the computational model and algorithms of the Artificial Neural Network (ANN). Understanding the network structure of white matter communication pathways is essential for unraveling the mysteries of the brain's function, organization, and evolution. Macaque brain white matters of Long-Distance Pathways (LDP) are successfully found and mapped [16]. By using diffusion spectrum imaging, the article noninvasively mapped whole-brain structural connectivity network of human cerebral pathways within 66 cortical regions and 998 regions of interest, and found brain regions within the structural core share high degree, strength, and betweenness centrality, and they constitute the connector hubs that link all major structural modules [17]. Using noninvasive multi-modal neuroimaging techniques, the reference [18] designed a connectivity-based parcellation framework that identifies the subdivisions of the entire human brain with 210 cortical and 36 subcortical subregions, revealing the in vivo connectivity architecture. The cognitive computing framework - TrueNorth [19], and neural processing unit - Darwin [20], the novel modular, non-von Neumann, ultra-low power, and compact architecture was proposed. TrueNorth and consists of a scalable network of neurosynaptic cores, with each core containing neurons, dendrites, synapses, and axons. The Semantic Pointer Architecture Unified Network (SPAUN) was to perform brain-like task by neural simulator of Nengo, which can simulate some brain functions such as copy drawing, image recognition, reinforcement learning, serial working memory, counting, question answering, rapid variable creation and fluid reasoning [21]. In addition, the accelerator for large-scale algorithms of Convolution Neural Networks(CNNs) and other Deep Neural Networks(DNNs) was designed, such as the

DianNao, ShiDianNao, PuDianNao and DaDianNao [22]. The CCNC's main objective is to research the problems of semantic computing for unstructured, massive, multi-modal, multi-temporal and spatial distribution of multimedia information processing, to establish a new generation of the cross-media information processing models and algorithms. The CCNC would mimic in two methods, which is the functional behavior of the cognitive framework at macroscopic level, and physiological mechanisms of the nervous system at the microscopic level. Currently, there are two main aspects, which have attracted attention for brain-inspired computing. The first is to simulate cognitive function based on system behavior, and the second is to research neural mechanisms based on structures of neurons, synapses, or local networks. However, there is still lacking effective methods about how to build an advanced system for complex function with the simple local neural network. The researchers have made unremitting exploration in the mechanisms of brain-inspired computing for a long time. The main research directions include ANN, HTM, DBN, PGM and so on.

With the development of neural computing, some new ANN appears such as neocognitron,spiking neural network, convolutional network, Hierarchical Model and X (HMAX) model, Neural Turing Machines (NTM), recursive neural networks, recurrent neural networks, deep residual networks, Long Short-Term Memory (LSTM), attention-based neural network and memory-based neural network etc. With fast deep learning algorithm was proposed by Hinton, deep learning has achieved unprecedented results in many applications. Microsoft research found that the relative algorithm can reduce the error to 33 % for large-vocabulary speech recognition in Switchboard dataset [23], and Google Labs also found that the accuracy of recognizing object categories increased to 70 % than the current best result in ImageNet dataset [24]. Document [25] presented a single model that yields good results on a number of problems spanning multiple domains, which contain convolutional layers, an attention mechanism, and sparsely-gated layers. Currently, deep learning algorithm has achieved unprecedented results in image category, speech recognition and natural language processing. The reference [26] proposes learning a set of high-level face representations through deep convolutional networks, and the accuracy of their algorithm achieves an impressive accuracy rate of 97.45 % on the LFW benchmark. The reference [27] developed a deep Q-network agent with theory of reinforcement to receiving only the pixels and the game score as inputs, it was able to surpass the performance of all previous algorithms and achieve a level comparable to that of a professional games tester across a set of 49 games. Furthermore, the reference [28] designed algorithms by a new search algorithm that combines Monte Carlo simulation with value and policy networks, and the program AlphaGo achieved a 99.8 % winning rate against other Go programs.

It is often a very difficult task that we have attempted to find a general model since the complexity of the central nervous system. We need to let the neural network imitate not only brain's function but also brain's structure. Hawkins and George have proposed the HTM model based on cortical micro-circuits in 2006. Bayesian belief propagation theory of HTM was proposed [29], and set up cortical micro-circuits mathematical model for design of the brain's cortical functional column. Similar in structure design of Restricted Boltzmann Machine (RBM), HTM also has double layers. The difference is that HTM nodes consider the property of spatial-temporal locality and hierarchical. The idea of HTM is from approximate reasoning and belief propagation algorithm by Pearl's Bayesian belief network, but there are no effective learning algorithms.

In contrast, between Hinton's deep learning algorithm for DBN and Hawkins's cortical algorithm for HTM, both can be classified by unsupervised learning, and can be stacked to build up a feedback hierarchy structure. RBM doesn't utilize the spatial-temporal locality fully, but HTM more modular to use spatial-temporal locality and hierarchical based on belief propagation algorithm of PGM [30]. Reference [31] proposed a data-driven approach for cross-media retrieval based on the probabilistic topic model by automatically learning its underlying semantic vocabulary. The reference [32] presents the concepts computational model by a Bayesian criterion, and achieves human-level performance while outperforming while outperforming recent deep learning approaches. In a few words, both DBN and HTM can be seen as a special case in mathematical formalism.

3. Visual and auditory information integration of cortical semantic classification

3.1 LDP of cerebrum

The human central nervous system has white matter, grey matter, substantia nigra and other tissue. On the one hand, neocortex's function in the grey matter is structurally similar to the processing unit in linear analogue system and gate circuit in nonlinear digital system. On the other hand, LDP in white matter constitutes complex wiring diagram of a neural network for information processing.

Function and structure of the cerebrum are one of the most complex systems in nature, it is generally thought that neocortex of the cerebrum is an important part which processes logical intelligence. The thalamus is switching of selective attention, which processes information from all the senses except smell and sends it on to the cerebral cortex for more analyses. Hippocampus and the limbic system are the controllers of memory and emotions. Now, various methods are used to discover the brain mechanism. Neuroscience used white box and bottom-up methods to research neural information processing mechanism of cortical structures and neural pathways. Cognitive science used methods of the black box and top-down to analyze the function and phenomenon of cognitive. Then built brain model of information processing in theory, computer science implemented mathematical logic operation on finite state machine based on the Turing machine. From MLP model to HTM and DBN, people never stop to explore the use of cognitive processing mechanisms of the nervous system to promote information computing. You can think of the neural system and cognitive function as an isomorphic relationship.

Conjecture 1. By establishing related computational model M, it can build the mapping between the neural structures (or processes) N and cognitive operations (or functions) Ψ.

$$M : \Psi \longleftrightarrow N.$$

According to the macaque brain LDP database of CoCoMac processing [16], a large number of neural pathways and circuits exist in between primary auditory cortex area (A1) and primary visual cortex area (V1), and the number of interneuron is linearly related to the derivative of neural pathways and circuits. The hierarchy of cortical function and structure of neural pathways and circuits can provide significant evidences for the neurocognitive model. Neural circuits are the important material of relevance feedback, stochastic resonance, even recurrence iterative. Studies indicate the central nervous network is a scale-free and small-world complex network. Fig. 1 illustrates the node degree and pathways connection relationship in the eight parts of the cerebral cortex, where the amount of connections denoted by line size, and the amount of nodes and degree in areas denoted by node size. There are the pathways number of all brain regions, but the pathways amount and degree have a significant difference in each area.

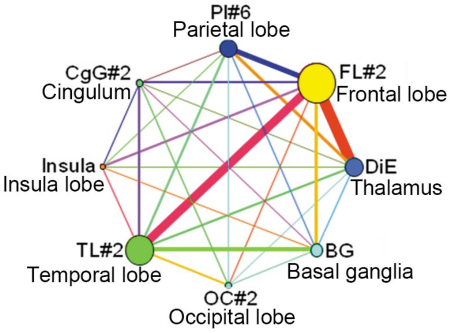

Fig. 1 *The simplify connection diagrams in 8 major areas of the whole-brain.*

3.2 Cerebral cortex model

Cerebral cortex is grouped into three major types of area: paleocortex, olfactory and neocortex. Paleocortex doesn't have clear layers; olfactory has three layers, and neocortex made up of six layers which account for 90 % of the area of cerebral cortex. The neocortex is commonly described as comprising three parts: primary sensory cortex, primary motor cortex and association cortex. The neocortex is the outer layer of the cerebral hemispheres, and made up of six layers (Fig. 2). Each layer of the neocortex has different functions. L4 receives incoming information; L2 and L3 layers make up local neural circuitry to process information; L1 resolve project information of convergence and inhabitation; the last information outputs from L5 and L6.

Most studies suggest that neocortex of vision, audition and association are similar in structure. Cortical columns are a base unit for information processing in neocortex. Cortical columns have the phenomenon of hierarchical processing and the mechanism of lateral inhibition of each other. Micro-columns consist of local circuits in neocortex. Physical stimuli are perceived and programmed to generate nerve impulse by visual-auditory sensory neurons, and function of micro-column is feature detection. Macro-column or super-column consists of micro-columns to process special information and generates some cognitive functions. The spike active probability is propagating among micro-columns. Micro-columns

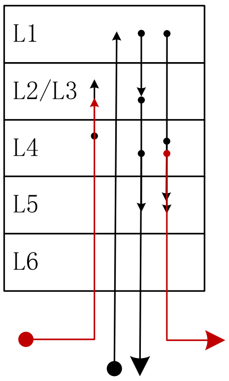

Fig. 2 *The neocortex micro-column structure schematic.*

converge information from lower neighbor micro-columns, and diverge information from upper neighbor micro-columns. At the same time, it also receives feedback information from LDP, and the receives the prediction information from upper neighbor. Fig. 2 illustrates the structure of information propagation and processing in micro-column.

In addition, there are a lot of connections in thalamo-cortical projection system, the thalamic association nuclei which deal with selective attention of sensory information, and the thalamic association nuclei which deal with cortical information interchange. According to neurocognitive mechanism, Fig. 3 illustrates

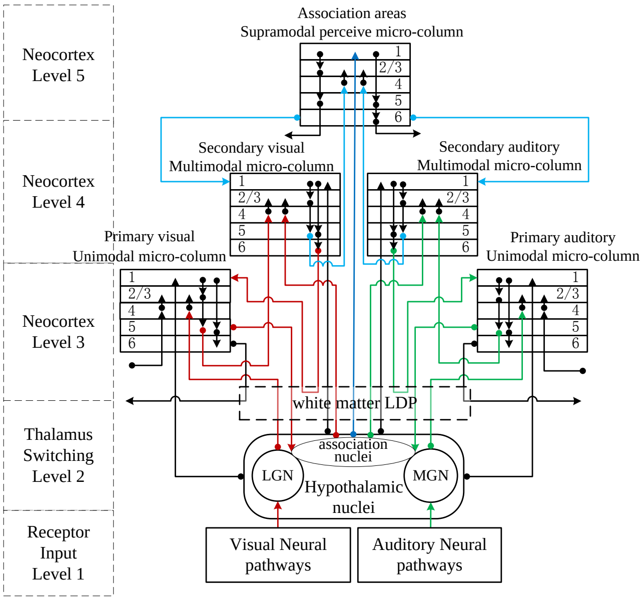

Fig. 3 *Schematic of thalamo-cortical projection system of the visual-auditory neurocognitive collaborative information processing pathway.*

neurocognitive pathway of visual-auditory information collaborative processing in thalamo-cortical projection system which has 5 levels. The thalamus is the center of information switching; Level 2 is visual-auditory feature detector such as brightness, edge, tone and loudness. Level 3 constructs the super-column to mimic primary visual-auditory sensory cortex. Level 4 imitates multi-modal sensory of secondary visual-auditory cortex. Level 5 simulates association cortex.

3.3 Structure temporal-spatial node of micro-column

Conjecture 2. In view, that neocortex has the similar structure. We can make the following hypothesis. Similar mechanisms can be used in all of cerebrum neocortex. The visual-auditory information processing can be described by uniform neocortex framework, and can be applied to learn, inference, prediction, and other issues.

In order to simplify the model design, at first, we merge 6 nodes micro-column (Fig. 4) to triple nodes micro-column (Fig. 5) where middle layer 4 input information, lower layer (L5 and L6) output information, and upper layer (L1, L2 and L3) process information. In fact, cortex information processing has the temporal and

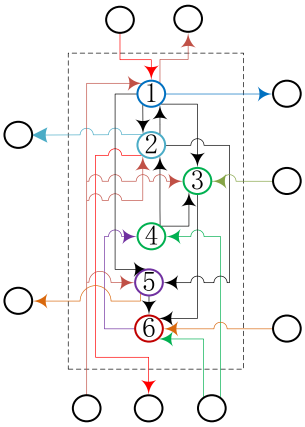

Fig. 4 *The hierarchical structure model of the micro-column of 6 nodes.*

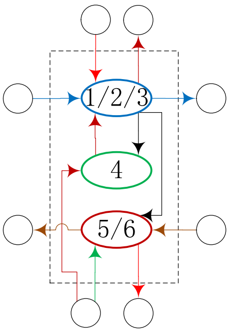

Fig. 5 *The hierarchical structure model of the micro-column of 3 nodes.*

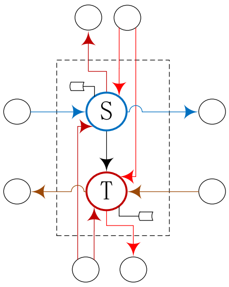

Fig. 6 *The hierarchical structure model of the micro-column of 2 nodes.*

spatial property. Furthermore, we simplify the model with dual nodes structure (Fig. 6). It is noted that this simplification does not lose the advantage of bionics; in fact, HTM node, RBM node and SVM are using double structure. The function of S mimicking from L1 to L4 to memory and process spatial patterns; the function of T mimicking L5 and L6 to memory and process temporal patterns; both nodes of S and T can memorize belief which comes from owner and other nodes.

3.4 Hierarchical network architecture of super-column

According to neurocognitive system hierarchical architecture and temporal-spatial locality, super-column architecture also uses hierarchical, multi-level, and bidirectional mapping structure. Super-column composed by a micro-column with principles of "the same layer collaborative" and "hierarchical processing" (Fig. 7).

According to neurocognitive data, we design auditory process super-column with 4 layers, the micro-column number of each layer is 8, 4, 2, and 1; and design visual process super-column with 3 layers, the micro-column count of each layer is 16, 4, and 1; design visual-auditory collaborative process super-column with 3 layers, the micro-column count of each layer is 3 (Fig. 8).

3.5 Information propagation algorithm of cortical columns

Fig. 9 shows belief flows that the information of cortical columns receives and propagates. λ_{YX} is the belief from children layers and upward. π_{UX} is the belief from parent layers and downward belief from owner. $\mathrm{Bel}(X)$ is node owner certainty information. $\lambda_{\mathrm{XU}}(X)$ is the belief to children layers and upward. $\pi_{\mathrm{XY}}(X)$ is the belief from parent layers and downward belief.

According to PGM algorithm ideas such as belief propagation algorithm of directed graph, sum product algorithm of an undirected graph, and junction tree algorithm etc., Fig. 10 illustrates information the cortical columns reception and belief propagation. The algorithm of cortical columns information propagation as follows:

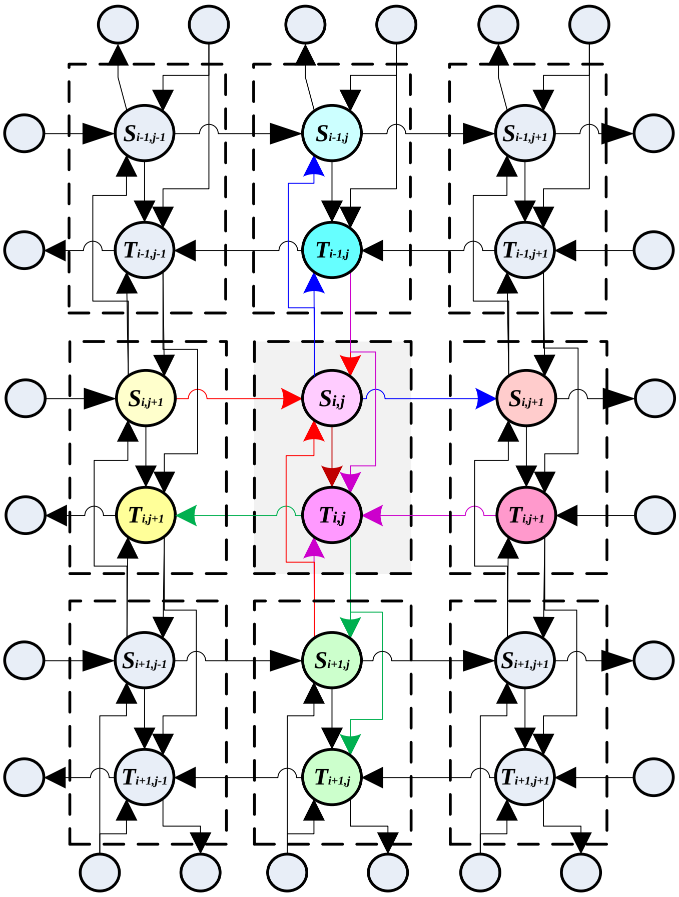

Fig. 7 *The structure model of the super-column.*

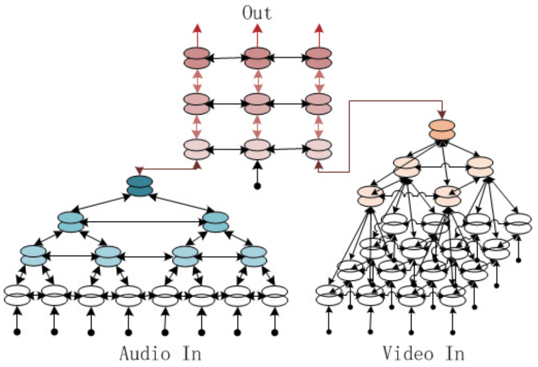

Fig. 8 *Visual-auditory information propagation network architecture of cortical micro-columns.*

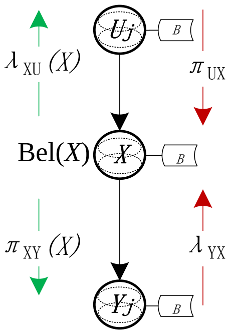

Fig. 9 *The cortical columns reception and belief propagation definition.*

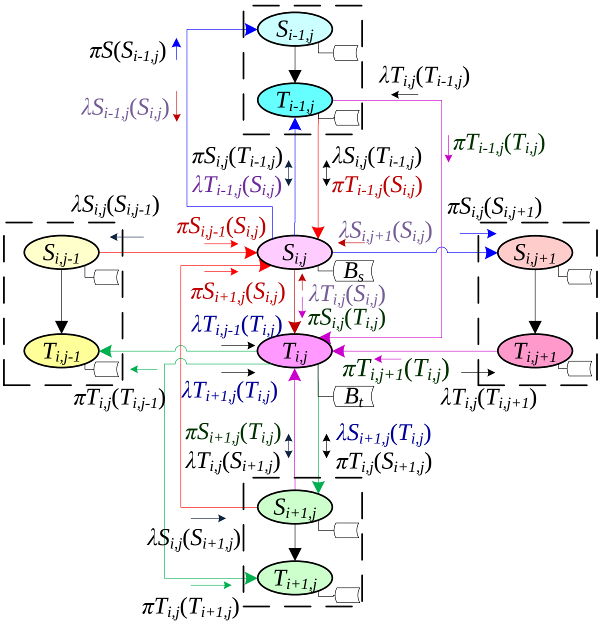

Fig. 10 *Receiving and propagation information computing in the triple-tier structure of cortical columns.*

Algorithm 1 Micro-column information propagation algorithm (MIPA).
MIPA calculates belief of micro-column node.

Step 1: initialization

Receive belief of media temporal-spatial pattern λ_{YX}, π_{UX}, $P(T_{i,j}|u)$, and $P(S_{i,j}|u)$. According to Fig. 8, set numbers of auditory super-column to 4 layers, and a total of 15 micro-columns.

Set numbers of visual super-column to 3 layers, and a total of 9 micro-columns.

Set numbers of collaborative super-column to 3 layers, and a total of 9 micro-columns.

Set $\lambda_{S_{i-1,j}}$, $\lambda_{T_{i-1,j}}$, $\pi_{S_{i-1,j}}$, and $\pi_{T_{i-1,j}}$ equal 0 on the top layers, and let $\lambda_{S_{i+1,j}}$, $\lambda_{T_{i+1,j}}$, $\pi_{S_{i+1,j}}$, and $\pi_{T_{i+1,j}}$ equal 0 on the bottom layers.

Step 2: the spatial micro-column node information which is from input of children and neighboring to be calculated is $\lambda(S)$, as the following equation shows:

$$\lambda(S_{i,j}) = \prod_{y=S_{i-1,j},S_{i,j+1},T_{i-1,j},T_{i,j}} \lambda_y(S_{i,j}).$$

Step 3: the spatial micro-column node information which is from input of parent and neighboring to be calculated is $\pi(S)$, as the following equation shows:

$$\pi(S_{i,j}) = \sum_{u=S_{i,j-1},S_{i+1,j},T_{i-1,j}} P(S_{i,j}|u) \prod_{u=S_{i,j-1},S_{i+1,j},T_{i-1,j}} \pi_u(S_{i,j}).$$

Step 4: calculate information $\mathrm{Bel}(S)$ of spatial micro-column node as follows:

$$\mathrm{Bel}(S_{i,j}) = \alpha\lambda(S_{i,j})\pi(S_{i,j}).$$

Step 5: check whether spatial patterns are existing in the node. Update spatial patterns weight to make it attenuate with time if spatial patterns exists in the node, and the new spatial patterns be inserted if it does not exist on the node.

At the same time, send information $\lambda_{S(u)}$ and $\pi_{S(y)}$ to neighboring nodes. The output information propagation steps are as follows:

Step 6: calculate information $\lambda(u)$ of spatial micro-column node which is sent to parents and neighboring output information as follows:

$$\lambda_{S_{i,j}}(u) = \sum_{y=S_{i-1,j},S_{i,j+1},T_{i-1,j},T_{i,j}} \lambda_y(S_{i,j}) \sum_{v\in U,v\neq u} P(S_{i,j}|v) \prod_{v\in U,v\neq u} \pi_v(S_{i,j}),$$

where $u = \{S_{i,j-1}, S_{i+1,j}, T_{i-1,j}\}$.

Step 7: calculate information $\pi(y)$ of spatial micro-column node which sends to children and neighboring output information as follows:

$$\pi_{S_{i,j}}(y) = \alpha\pi(S_{i,j}) \prod_{k\neq y} \lambda_k(S_{i,j}),$$

where $y = \{S_{i-1,j}, S_{i,j+1}, T_{i-1,j}, T_{i,j}\}$.

Similarly, related information about temporal micro-column node was computed as follows step.

Step 8: temporal micro-column node information which is from input of children and neighboring to be calculated is $\lambda(T)$, as the following equation shows:

$$\lambda(T_{i,j}) = \prod_{y=T_{i,j-1},T_{i+1,j},S_{i+1,j}} \lambda_y(T_{i,j}).$$

Step 9: temporal micro-column node information which is from input of parent and neighboring to be calculated is $\pi(T)$, as the following equation shows:

$$\pi(T_{i,j}) = \sum_{u=S_{i,j},S_{i+1,j},T_{i,j+1},T_{i-1,j}} P(T_{i,j}|u) \prod_{u=S_{i,j},S_{i+1,j},T_{i,j+1},T_{i-1,j}} \pi_u(T_{i,j}).$$

Step 10: calculate information $\mathrm{Bel}(T)$ of temporal micro-column node as follows:

$$\mathrm{Bel}(T_{i,j}) = \beta\lambda(T_{i,j})\pi(T_{i,j}).$$

Step 11: check whether temporal patterns are existing in the node. Update temporal patterns weight to make it attenuate with time if temporal patterns exist in the node, and the new temporal patterns be inserted if temporal patterns do not exist on the node.

At the same time, the node sends information $\lambda T(u)$ and $\pi T(y)$ to neighboring nodes. The output information propagation methods are as follows:

Step 12: calculate information $\lambda(u)$ of temporal micro-column node which sent to the parent and neighboring output information as follows:

$$\lambda_{T_{i,j}}(u) = \sum_{y=T_{i,j-1},T_{i+1,j},S_{i+1,j}} \lambda_y(T_{i,j}) \sum_{v\in U,v\neq u} P(T_{i,j}|v) \prod_{v\in U,v\neq u} \pi_v(T_{i,j}),$$

where $u = \{S_{i,j}, S_{i+1,j}, T_{i,j+1}, T_{i-1,j}\}$.

Step 13: calculate information $\pi(y)$ of temporal micro-column node which sent children and neighboring output information as follows:

$$\pi_{T_{i,j}}(y) = \alpha\pi(T_{i,j}) \prod_{k\neq y} \lambda_k(T_{i,j}),$$

where $y = \{T_{i,j-1}, T_{i+1,j}, S_{i+1,j}\}$.

Step 14: calculate information $\mathrm{Bel}(N)$ of temporal-spatial micro-column node as follows:

$$\mathrm{Bel}(N_{i,j}) = \gamma\left(\mathrm{Bel}(S_{i,j}) + \mathrm{Bel}(T_{i,j})\right).$$

Return: temporal-spatial pattern belief of micro-column $\mathrm{Bel}(T_{i,j})$ and $\mathrm{Bel}(S_{i,j})$, and output temporal-spatial pattern belief $\lambda_{\mathrm{XU}}(X)$ and $\pi_{\mathrm{XY}}(X)$.

3.6 CSR algorithm of auditory-visual information

Algorithm 2 Cross-media semantic retrieval algorithm (CSRA). CSRA solves CSR processing from multi-media by cross-modal semantic mapping algorithm (CSMA) algorithm [10].

Step 1: initialization auditory media A, visual media V, and cross-modal semantic mapping (CSM) parameters [10].

It includes visual temporal-spatial patterns probability VGN, auditory temporal-spatial patterns probability AGN, auditory cortex belief ATS, visual cortex belief VTS, concept of visual-auditory integration belief CTS, visual object emotional control value VRP, auditory object emotional control values ARP, visual-auditory object emotional control value ERP, visual object temporal memory value VM, auditory object temporal memory value AM, visual-auditory object temporal memory value CM, and so on.

Step 2: **if** input media is audio A **then**

Similar audio A', correlation video V', and correlation concept of visual-auditory integration CTS$'$ can be calculated as follows:

Step 2.1: auditory cortex belief ATS was calculated by audio A.

```
    Step 2.2: search similar audio $A'$ which meets the condition that auditory object temporal memory value AM $\approx 0$.
    Step 2.3: set visual cortex belief VTS $\approx 0$, and calculate CTS of visual-auditory integration concept to find correlation concept CTS$'$.
    Step 2.4: according to the concept of visual-auditory integration CTS, correlation video $V'$ is found which meets the condition that visual object temporal memory value VM $\approx 0$.
end if
Step 3: if input media is video $V$ then
    Similar video $V'$, correlation audio $A'$ , and correlation concept of visual-auditory integration CTS$'$ can be calculated as follows:
    Step 3.1: visual cortex belief VTS was calculated by video $V$.
    Step 3.2: search similar video $V'$ which meets the condition that visual object temporal memory value VM $\approx 0$.
    Step 3.3: set auditory cortex belief ATS $\approx 0$, and calculate CTS of visual-auditory integration concept to find correlation concept CTS$'$.
    Step 3.4: according to the concept of visual-auditory integration CTS, correlation audio $A'$ is found, which meets the condition that visual object temporal memory value AM $\approx 0$.
Step 4: if input is concept of visual-auditory integration CTS then
    Similar concept of visual-auditory integration CTS$'$, correlation audio $A'$ and video $V'$ can be calculated as follows:
    Step 4.1: find the similar concept of visual-auditory integration CTS$'$ which meets the condition that CM $\approx 0$ for the concept of visual-auditory integration CTS.
    Step 4.2: search correlation video $V'$ which meets the condition that visual object temporal memory value VM $\approx 0$.
    Step 4.3: search correlation audio $A'$ which meets the condition that auditory object temporal memory value AM $\approx 0$.
end if
Return: similar media or correlation media (such as auditory media $A'$, visual media $V'$, and concept of visual-auditory integration CTS$'$, and so on.)
```

4. Experiments and analysis

4.1 Test data

Due to the nature of audio and video information have a lot of noises, in order to take the quantitative and qualitative analysis and evaluation to the modal, we adopt 26 letters in the English alphabet as training media, that are the concept 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. 11 shows all media of training and testing: spectral distribution of English letters fonts (*I* and *III* rows in Fig. 11), spectral distribution of English letters pronunciation (*II* and *IV* rows in Fig. 11); training media time-frequency distribution of all speech (V row and α column in Fig. 11); test media time-frequency distribution of all speech with Gaussian white noise (V row and β column in Fig. 11); train media space-frequency distribution of all fonts (V row and γ column in Fig. 11); test media space-frequency distribution of all fonts with Gaussian white noise (V row and σ column in Fig. 11).

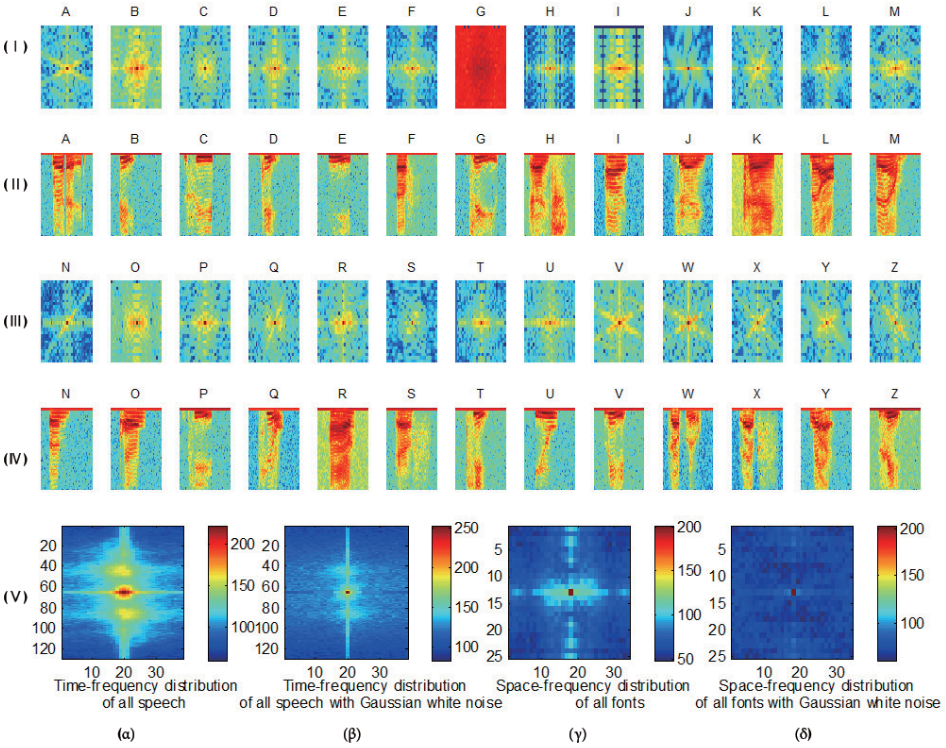

Fig. 11 *Time-frequency and space-frequency distribution of the training media.*

All media of training and testing were transformed features with 333 dimensions after processing. The image features have frequency-domain with 25 dimensions and time-domain with 50 dimensions, and audio features have frequency-domain with 111 dimensions and time-domain with 222 dimensions.

4.2 CSRNCVA result analysis

Fig. 12 shows the simulation result that information propagation characteristic of the visual-auditory cortical columns neurocognitive by Algorithm 1 (MIPA). Let six inputs of cortical columns denoted by sets 0, 0.5, 1, Fig. 12 (a) shows the distribution of input information. When setting eight I/O parameters ($P(T_{i,j}|u), P(S_{i,j}|u)$) values are 1, Fig. 12 (c) shows the distribution of output information. Fig. 12 (b) is the simulation result of temporal-spatial information Bel(T) and Bel(S) of cortical columns for different I/O information. Based on the simulation results, we can find that cortical column can generate responsive only to specific input patterns.

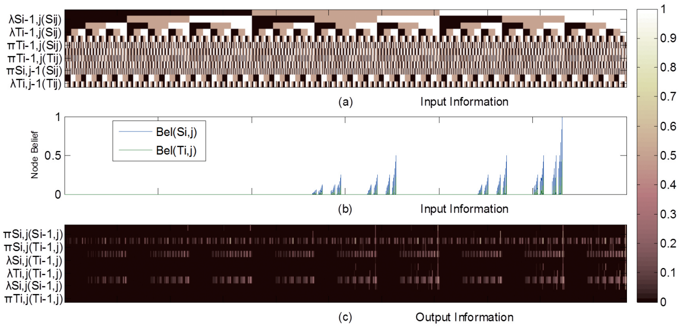

Fig. 12 *Characteristics of information propagation of the visual-auditory cortical columns neurocognitive.*

To validate the model's effectiveness, the MIR Flickr-25000 dataset was used in our experiments [33]. The dataset consists of 1 million images retrieved from the social photography website Flickr along with their user assigned tags. Among the 1 million images, 25,000 have been annotated for 38 topics including object categories. It is high-quality dataset for multimedia information semantic retrieval. Without loss of generality and efficiency, 80 samples were randomly selected for each topic as training data in 36 topics. Then 40 samples were randomly selected for each topic as test data. Image features were extracted by pyramid histogram of words based on SIFT, Gist features, MPEG-7 descriptors, total 3857 dimensions. Similar simulation experiment, training and testing image were transformed frequency-domain with 25 dimensions and time-domain (context of text and image by sequence and attention) with 50 dimensions, a total of 75 dimensions.

Tab. I shows the results of the mean average precision (MAP) and the Top-20 precision (Precision@20) by CSRNCVA model, and with methods of multimodal LDA, SVM, DBN, DBM [2, 3, 33] , the results which can be seen in CSRNCVA model of CSR Precision@20 is slightly better than the other models, but MAP slightly lower than the multi-modal DBN and DBM.

Models	Multi-modal LDA	Multi-modal SVM	Multi-modal DBN	Multi-modal DBM	CSRNCVA modal
MAP	0.499	0.476	0.591	0.607	0.587
Precision@20	0.764	0.758	0.868	0.865	0.888

Tab. I *CMR results of mean average precision (MAP) and Precision@20 obtained by different models.*

The experiments results show that the MIPA algorithm has the function of cross-media feature extraction in our CSRNCVA model. CSRA algorithm can effectively compute semantic similarity measure of homogeneous media and cross-media semantic correlation of heterogeneous media.

5. Conclusion

In this paper, we presented a CSRNCVA model with mechanism neurocognitive visual-auditory cortex of the central nervous system. Then we give the CSR algorithms which take into account the idea of DBN, HTM, PGM and hierarchical learning theory. Simulation results show that CSRNCVA model is robust and effective. We only do a preliminary exploration with CCNC. The most model's parameters only learn from physiological data, which due to neurocognitive mechanisms of the brain's complexity. Looking into the future, the models need to be combined with a new theory of deep learning, probability learning and modern neuroscience findings improve the relevancy algorithms. The model improvement would be discussed in another article [34]. Furthermore, cross-media semantic search engine based on CCNC would be built for target classification and recognition of high resolution remote-sensing image [35].

Acknowledgement

This work is supported by National Natural Science Foundation of China (61305042), Projects of Center for Remote Sensing Mission Study of China National Space Administration (2012A03A0939), Key Research and Promotion Projects of Henan Province (182102310724), and Science and Technological Research of Key Projects of Henan Province (172102110006).

References

[1] PENG Y.X., ZHU W.W., et al. Cross-media analysis and reasoning: advances and directions. *Frontiers of Information Technology and Electronic Engineering*, 2017, 18(1), pp. 44–57, doi: 10.1631/FITEE.1601787.

[2] NGIAM J., KHOSLA A., et al. Multimodal deep learning. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Bellevue, ACM, 2011, pp. 689–696.

[3] 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, pp. 2231–2239.

[4] GUO Q., JIA J., SHEN G.Y., et al. Learning robust uniform features for cross-media social data by using cross autoencoders. *Knowledge-Based Systems*, 2016, 102, pp. 64–75, doi: 10.1016/j.knosys.2016.03.028.

[5] CHU L., ZHANG Y., et al. Effective multimodality fusion framework for cross-media topic detection. *IEEE Transactions on Circuits and Systems for Video Technology*. 2016, 26(3), pp. 556–569, doi: 10.1109/TCSVT.2014.2347551.

[6] WEI Y.C., ZHAO Y., ZHU Z.F., et al. Modality-dependent cross-media retrieval. *ACM Transactions on Intelligent Systems and Technology*, 2016, 7(4), pp. 57–57, doi: 10.1145/2775109.

[7] CAO D.L., JI R.R., LIN D.Z., et al. A cross-media public sentiment analysis system for microblog. *Multimedia Systems*, 2016, 22(4), pp. 479–486, doi: 10.1007/s00530-014-0407-8.

[8] ZHANG H., GAO X.Y., WU P., et al. A cross-media distance metric learning framework based on multi-view correlation mining and matching. *World Wide Web-Internet and Web Information Systems*, 2016, 19(2), pp. 181–197, doi: 10.1007/s11280-015-0342-4.

[9] YAN Y., NIE F.P., LI W., et al. Image Classification by Cross-Media Active Learning With Privileged Information. *IEEE Transactions on Multimedia*, 2016, 18(12), pp. 2494–2502, doi: 10.1109/TMM.2016.2602938.

[10] LIU Y., ZHENG F.B., ZUO X. CSMCCVA: Framework of cross-modal semantic mapping based on cognitive computing of visual and auditory sensations. *High Technology Letters*, 2016, 22(1), pp. 90–98, doi: 10.3772/j.issn.1006-6748.2016.01.013.

[11] WANG Y. On cognitive informatics. *Brain and Mind*. 2003, 4, pp. 151–167, doi: 10.1023/A:1025401527570.

[12] MODHA D.S. ANANTHANARAYANAN R. et al. Cognitive computing. *Cognitive computing*. 2011, 54(8), pp. 62–71, doi: 10.1145/1978542.1978559.

[13] KUMAR C.A., ISHWARYA M.S., CHU K.L. Formal concept analysis approach to cognitive functionalities of bidirectional associative memory. *Biologically Inspired Cognitive Architectures*. 2015, 12, pp. 20–33, doi: 10.1016/j.bica.2015.04.003.

[14] SHANG X., ZHANG H., CHUA T.S. Deep Learning Generic Features for Cross-Media Retrieval. In: Tian Q., Sebe N., Qi GJ., Huet B., Hong R., Liu X., ed. *MultiMedia Modeling. Springer, Cham.*, MMM 2016, 9516, pp. 239–304, doi: 10.1007/978-3-319-27671-7_22.

[15] ZHAI X., PENG Y., XIAO J. Cross-media retrieval by intra-media and inter-media correlation mining. *Multimedia Systems*. 2013,19, pp. 395, doi: 10.1007/s00530-012-0297-6.

[16] 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), pp. 13485–13490, doi: 10.1073/pnas.1008054107.

[17] HAGMANN P., CAMMOUN L., et al. Mapping the structural core of human cerebral cortex. *PLOS Biology*. 2008,6(7), pp. 1479–1493, doi: 10.1371/journal.pbio.0060159.

[18] FAN L., LI H., et al. The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral Cortex*. 2016, 26(8), pp. 3508–3526, doi: 10.1093/cercor/bhw157.

[19] PREISSL R., WONG T.M., et al. Compass: a scalable simulator for an architecture for cognitive computing. In: *High Performance Computing, Networking, Storage and Analysis (SC), 2012 International Conference for*, Salt Lake City, UT, 2012, pp. 1–11, doi: 10.1109/SC.2012.34.

[20] SHEN J., MA D., et al. Darwin: a Neuromorphic Hardware Co-Processor based on Spiking Neural Networks. *Science China Information Sciences* . 2016, 59(2), pp. 1–5, doi: 10.1007/s11432-015-5511-7.

[21] ELIASMITH C., STEWART T.C., et al. A large-scale model of the functioning brain. *Science*. 2012, 338(6111), pp. 1202–1205, doi: 10.1126/science.1225266.

[22] CHEN Y., LUO T., et al. DaDianNao: A Machine-Learning Supercomputer. In: *47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2015, pp. 609–622, doi: 10.1109/MICRO.2014.58.

[23] DAHL G.E., YU D., 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), pp. 30–42, doi: 10.1109/TASL.2011.2134090.

[24] LE Q.V., RANZATO M.A., et al. Building high-level features using large scale unsupervised learning. In: *The 29th International Conference on Machine Learning (ICML)*, Edinburgh, 2012, pp. 81–88.

[25] KAISER L., GOMEZ A.N., SHAZEER N., et al. One Model To Learn Them All. arXiv:1706.05137. 2017.

[26] SUN Y., WANG X.G., TANG X.O. Deep Learning Face Representation from Predicting 10,000 Classes. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, Columbus, OH, USA, pp. 1891–1898, doi: 10.1109/CVPR.2014.244.

[27] MNIH V., KAVUKCUOGLU K., et al. Human-level control through deep reinforcement learning. *Nature* . 2015, 518(7540), pp. 529–533, doi: 10.1038/nature14236.

[28] SILVER D., HUANG A., et al. Mastering the game of Go with deep neural networks and tree search. *Nature* . 2016, 529(7587), pp. 484–489, doi: 10.1038/nature16961.

[29] GEORGE D., HAWKINS J. Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*. 2009, 5(10), pp. 1–26, doi: 10.1371/journal.pcbi.1000532.

[30] BLEI D.M. Probabilistic topic models. *Communications of the ACM*. 2012, 55(4), pp. 77–84, doi: 10.1145/2133806.2133826.

[31] HABIBIAN A., MENSINK T., et al. Discovering Semantic Vocabularies for Cross-Media Retrieval. In: *ICMR'15: Proceedings of the 2015 ACM International Conference on Multimedia Retrieval*, 2015, pp. 131–138, doi: 10.1145/2671188.2749403.

[32] LAKE B., SALAKHUTDINOV R., TENENBAUM J., Human-level concept learning through probabilistic program induction. *Science* . 2015, 350(6266), pp. 1332–1338, doi: 10.1126/science.aab3050.

[33] HUISKES M.J., LEW M.S. The MIR Flickr retrieval evaluation. In: *ACM International Conference on Multimedia Information Retrieval (MIR'08)*, SVancouver, Canada, 2008. Available from: http://press.liacs.nl/mirflickr.

[34] LIU Y. Research on the brain-inspired cross-modal neural cognitive computing framework. 2018, arXiv:1805.01385 [cs.NE].

[35] LIU Y., ZHENG F.B. Object-oriented and multi-scale target classification and recognition based on hierarchical ensemble learning. *Computers and Electrical Engineering*. 2017, 62(2017), pp. 538–554, doi: 10.1016/j.compeleceng.2016.12.026.