SCIENCE CHINA

xx xxxx Vol. xx No. x: 1–15
doi:

A Distributed Computational Cognitive Model for Object Recognition

Yong-Jin Liu[1]*, Qiu-Fang Fu[2], Ye Liu[2] & Xiaolan Fu[2]

[1] *Tsinghua National Laboratory for Information Science and Technology,*
Department of Computer Science and Technology, Tsinghua University,
Beijing 100084*, China,*
[2] *State Key Lab of Brain and Cognitive Science,*
Institute of Psychology, Chinese Academy of Sciences,
Beijing 100101*, China*

Received Month Day, Year; accepted Month Day, Year

Abstract Based on cognitive functionalities in human vision processing, we propose a computational cognitive model for object recognition with detailed algorithmic descriptions. The contribution of this paper is of two folds. Firstly, we present a systematic review on psychological and neurophysiological studies, which provide collective evidence for a distributed representation of 3D objects in the human brain. Secondly, we present a computational model which simulates the distributed mechanism of object vision pathway. Experimental results show that the presented computational cognitive model outperforms five representative 3D object recognition algorithms in computer science research.

Keywords distributed cognition, computational model, object recognition, human vision system

1 Introduction

Object recognition is one of fundamental tasks in computer vision. Many recognition algorithms have been proposed in computer science area (e.g., [1, 2, 3, 4, 5]). While the CPU processing speed can now be reached at 10^9 Hz, the human brain has a limited speed of 100 Hz at which the neurons process their input. However, compared to state-of-the-art computational algorithms for object recognition, human brain has a distinct advantage in object recognition, i.e., human being can accurately recognize one from unlimited variety of objects within a fraction of a second, even if the object is partially occluded or contaminated by noises. Thus, it is much desired to explore computational cognitive models of how human brain recognizes objects, in both areas of computer vision and cognitive computation.

Objects in real world space project color natural images on the retina in the human vision system which has normal visual acuity and normal color vision. The information of stimuli are transformed to the visual cortex, in which a two-stream hypothesis is widely accepted [6]. The dorsal stream from V1 to intraparietal areas solves the problem of where the object is located. The ventral stream goes through V2 and V4 to inferior temporal areas solves the problem of what the object is. According to this hypothesis, temporal cortex is involved in object recognition task. Based on the cognitive mechanism in the human vision, in this paper we present a computational cognitive model for object recognition.

*Corresponding author (email: liuyongjin@tsinghua.edu.cn)

To specify the task of object recognition, we make distinctions between object perception and object recognition. In literature, object perception may have different meanings and in this study we take a narrow scope: Object perception concerns how the shape of objects are perceived by controlling stimuli presented to the sense organs. Object recognition, in addition to seeing an object, concerns about seeing an object as something that has been seen before. So object recognition involves memory and learning. Based on this perspective, we study object recognition in a process of perception, memory, learning and judgment[1] [7]. There are many factors affecting the object recognition, including size, illumination, viewpoint, orientation and so on. We will study these factors and organize them into a cognitive model after a systemic review of psychological and neurophysiological research, with collective evidences.

Based on the summarized cognitive model, we present a computational implementation of the cognitive model (also called computational cognitive model in this paper) for object recognition. A computational model of human cognition can be defined in several different senses. In this paper, we present a model which is quantitative in a programmable way. By utilizing cognitive functionalities in the human brain, our model is distinct from existing computational algorithms in computer science area. We further use McGill 3D shape benchmark [8] to demonstrate that the presented computational cognitive model outperforms five representative computer algorithms for object recognition. We make the following two contributions in this paper:

- We present a systemic review on psychological and neurophysiological studies with converging evidences, to uncover cognitive and neural mechanisms of object recognition in the human brain.
- Based on these cognitive mechanisms, we present a computational cognitive model for object recognition. The presented model utilizes distributed local features which are defined as activation patterns and are similarity-invariant. By utilizing a learning process characterized by a Markov chain model, the features are clustered into abstract representations stored in memory traces, which form the partial representations of the object.

2 Object representation and recognition in human vision

In this section, we summarize psychological and neurophysiological findings for object representation and recognition in the human vision. The summarization follows the cognitive process [7] including the mental process of perception, memory, learning[2] and judgment. Based on these findings, we develop a computational cognitive model in Section 3.

2.1 Perception

Perception as objects or features. The existence of neurological disorder prosopagnosia (or face blindness) and category-specific deficits in brain-damaged patients, shows that there are particular areas in human visual cortex for particular categories of stimuli. For examples, the fusiform face area (FFA) devotes to faces [9, 10], the parahippocampal place area (PPA) and the retrosplenial cortex (RSC) devote to scenes depicting places [11, 12, 13], and lateral occipital complex (LOC) devotes to a variety of objects [14], etc. Since the ventral temporal cortex has a limited number of areas, it is unlikely that each distinct category of objects has a corresponding area. Thus, it is possible that only a few biological relevant objects such as faces have responsible cortical areas emergiong through long-term evolution, while the representation of other unfamiliar objects are more widely distributed based on primitive features [15, 16].

2D-view or 3D-part features. Both 2D view-dependent features and 3D structural primitives have been used for distributed object recognition. One representative 3D-part-based approach to object perception is recognition-by-components (RBC) of geon structural description [17]. The fundamental assumption of the RBC theory is that 36 geons of volumetric components (generalized cones) are generally invariant

[1] Here, recognition is the specific judgment task we focus on.

[2] In the PMJ model [7] that is originally proposed in psychology domain, the part M includes both memory and learning. Since in information science domain, learning methods attract particular attention; thus we explicitly state memory and learning as two separated phases.

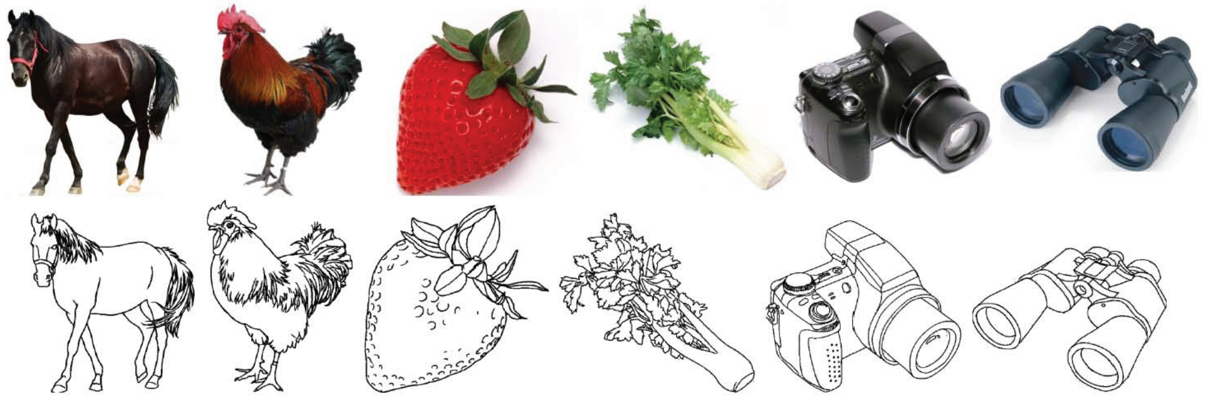

Figure 1: Color photographs and corresponding line drawings of single objects with a clear background. 120 color photographs are selected according to basic level categories: 60 natural objects (30 animals, 30 fruits and vegetables) and 60 man-made objects (30 tools, 30 electronic devices). All line drawings are produced by tracing contours in the color photographs using a simple graphical user interface, by three paid artists at the Academy of Arts & Design, Tsinghua University, China.

over viewing positions and are sufficient for diverse object representations. I.e., each object is represented by a unique relationship of a small set of volumetric building blocks called geons. The main weakness of the RBC theory is that no neurophysiological evidence exists to support the representation of volumetric building blocks. Conversely, several psychophysical studies show that the recognition is view-dependent [18], but not view-independent as hypothesized by the RBC theory. That is, given an object, some views are generally easier to recognize than others. This leads to a feature-based, multiple-view approach [19] to object representation and recognition: in this theory, each view is represented by a collection of small picture elements which are tolerant to slight deformation. Many researches demonstrate this view dependence and specificity property [18, 20, 21].

Color photographs vs. line drawings. Concerning 2D image features in a distributed representation, there could be many factors, such as color, texture, luminance, contrast and spatial frequencies. Line drawing, which is a set of sparse, simple two-dimensional featured lines without hatching lines or stippling for shading/tone effects, has been used to examine what information is applied for low-level features. Line drawings are a convention of art that even untrained infants and children can easily recognize them. Lines in line drawings include not only those edges that can be detected by object silhouette, intensity contrast and color gradients, but also some perceptually important lines that currently can only be captured by artists in an ambiguous way (Figure 1). Several pieces of neurophysiological evidence [15, 22, 23] supported that when the visual stimuli were presented as short as 120 ms, line drawings may trigger neural activities that are similar to the neural activities of color photographs and are sufficient for object and scene recognition. These findings [15, 22, 24] also suggest that line drawings appear to include definitive information sufficient for object recognition. In Section 3, we will present a visual circular feature representation to describe such definitive information. There is a long-standing controversy for two opposite surface-based (in terms of color, brightness, texture, etc.) versus edge-based (in terms of lines) representations of visual recognition [25]. Color is a fundamental factor of human perception [26] and some previous work also showed that color do help in computer program for object recognition [27]. A recent ERP study [28] showed that people need more time to extract information from color photographs than line drawings and people pay more attention to detect features and evaluate internal representation from color photographs than line-drawings. In this study, we take the hypothesis that color may affect the accuracy of recognition in long time, but line features are more critical for fast object category specification.

Single object vs. clustered scene. Much work on cognition was restricted to isolated objects with a clear background. Natural scenes can be regarded as a composition of objects. Detecting objects that are

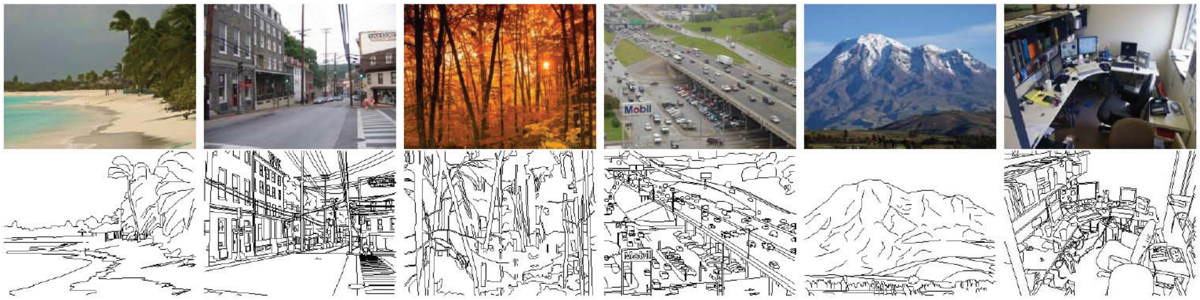

Figure 2: Color photographs and corresponding line drawings of six scene categories, including beaches, city streets, forests, highways, mountains and offices. Each image category has 80 color photographs. All line drawings are produced by tracing contours in the color photographs using a simple graphical user interface, by trained artists. Data courtesy of Professor Dirk B. Walther at Department of Psychology, The Ohio State University, USA.

consistent with some particular scenes contributes to a fast recognition of scenes [29]. However, image belonging to the same scene category such as forests may exhibit quite different color, luminance and individual objects but human can still make subtle distinctions between heterogeneous sets of images and recognize natural scenes as brief as 100-250 ms [30, 31]. By collecting fMRI data of ten participants viewing color photographs and line drawings (Figure 2) of six categories, i.e., beaches, city streets, forests, highways, mountains and offices, it was found [23] that simple line drawings are sufficient for a fast scene recognition. On the other hand, the gist of a scene was also considered to provide visual context for object recognition [32]. These results demonstrate that the distinctive features in natural scenes, as well as single objects, are not 3D-part-based, but may be 2D image features inherent in line drawings.

2.2 Memory

Memory units. What differentiates object recognition from object perception is the involvement of memory and learning. Corresponding to a distributed object representation in the visual cortex, a distributed model of memory was proposed in [33]. In this model, the processing system consists of a collection of simple processing units. Units are interconnected to each other and are organized into modules. Two opinions exist about what have with the units. One is to use specific exemplars (enumeration of specific experiences) in memory units for object category [34] and the other is to use prototype (abstract representation) of a category in memory units [35]. Enumeration of specific experiences works well in many studies. However, if every event or object is stored using an extremely rich representation, an almost unlimited storage is required. Thus we take abstraction (or prototype) representations of objects in memory units.

Memory traces. The memory trace of a particular pattern of activation is the change of weights in the interconnection of units. Neural system research maps the term 'unit' to an individual neuron and there are 10^{11} neurons in human brain connected by 10^{15} synapses [36, 37]. The weight change then corresponds to changes in synapses (a mode of plasticity), including either efficacy changes of existing synapses (weight changes) or structure changes (wiring changes) by addition or subtraction of synapses between neurons. In the presented study, we unify weight and wiring changes by making a complete graph of all units: assigning a nonzero (or zero) weight to an edge corresponds to weight (or wiring) changes. In the distributed memory model [33], what store in the memory are the interconnection strengths (weights) and memory traces are the changes in the strengths. Mental states are patterns of activation over units and when a part of a know pattern is input, its corresponding change in the subpattern can be used for the retrieval/recovery of rest subpattern. Each memory trace is distributed over different connections and each connection is contributed to different memory traces.

Short-term and long-term memory. The memory consolidation theory [37] supports that there are at least two stages of memory, short-term and long-term memory, which act independently in parallel. Short-term (or working) memories of new learned information are created almost instantly and they are easily disrupted by the learning of other information; this shows that the new memories are in a fragile state. The treatment of memory enhancement by long-term training shortly after original learning demonstrates that the memories are consolidated over time. The relatively slow consolidation from short-term (seconds to hours) to long-term (hours to months) has been demonstrated to be a biological mechanism of using adaptive functions to modulate memory strength [38].

2.3 Learning

Learning can evokes both weight and wiring changes. Weight changes are provided by modulating synaptic efficacy and wiring changes are provided by synapse formation and axonic and dendritic outgrowth and retraction in the adult brain [39]. The cortex is sparsely connected [36], meaning that neural circuit only has a small fraction of all possible connections between neurons. Learning-based cortical rewiring increases the storage capacity of neural circuit and long-term memory corresponds to the stability of synapse connection.

In terms of object recognition, 3D objects are likely stored in the brain as 2D views and each view is represented by a collection of distributed small picture elements, rather than 2D templates or 3D primitives [40]. The effect of learning through long-term training has been studied by investigating how continued exposure of an object may affect the representation of that object. It has been shown that sufficient exposure to particular stimuli causes the representation of that stimuli enhanced [41]. Furthermore, exposure experience of an object class also increases the number of neurons that are selectively discriminative for that object class [42].

Temporal correlations in appearance of object views also affect the learning of object representation [40]. When we see an object continuously from different viewing directions, the discrete image sequence for identifying that object is subject to transformation due to changes of viewing directions. The temporal order in the image sequence forms an integral part of the object representation. Several neurophysiological studies have demonstrated the influence of order in temporal sequence [43, 44].

2.4 Judgment

Single-cell-level studies [45] showed that cells in inferotemporal cortex of the monkey brain selectively respond to various object features and columns with cells that respond to similar features tend to cluster together. However, it is still controversial how individual neurons are organized at a large scale for object recognition. Some psychophysical evidences showed that faces are represented and recognized holistically [46, 47]. On the other hand, there is also evidence that a distributed feature representation can better explain the human data of other recognition tasks [15, 16, 22].

To reveal a pattern of large-scale spatial organization in object vision pathway, functional magnetic resonance imaging (fMRI) studies were preformed to measure patterns of response in the ventral temporal cortex. Pictures of faces, cats, five categories of man-made objects (houses, chairs, scissors, shoes and bottles) and nonsense images were used as stimuli and correlations between patterns of response were used as similarity indices [15]. The results showed that for each stimulus category, there is a distinct pattern of response and more importantly, the distinctiveness of response to a given category was not only limited to the cortex regions that responded maximally to that categories (i.e., the primary region), but also were identified by the patterns of non-maximal responses in the secondary regions[3]. For example, the region maximally responsive to houses is preferentially recruited to augment the perception of chairs and vice versa [16]. We take the hypothesis in [40] that a mixed type of holistic and distributed feature representations exist for all objects. For familiar objects like faces, holistic representation is dominated, while for unfamiliar objects, distributed feature representation is dominated: to use which one of two representations actually is determined by a learning process, i.e., familiar (or unfamiliar) objects were learned for a long (or short) process.

[3] They are the regions respond maximally to another categories.

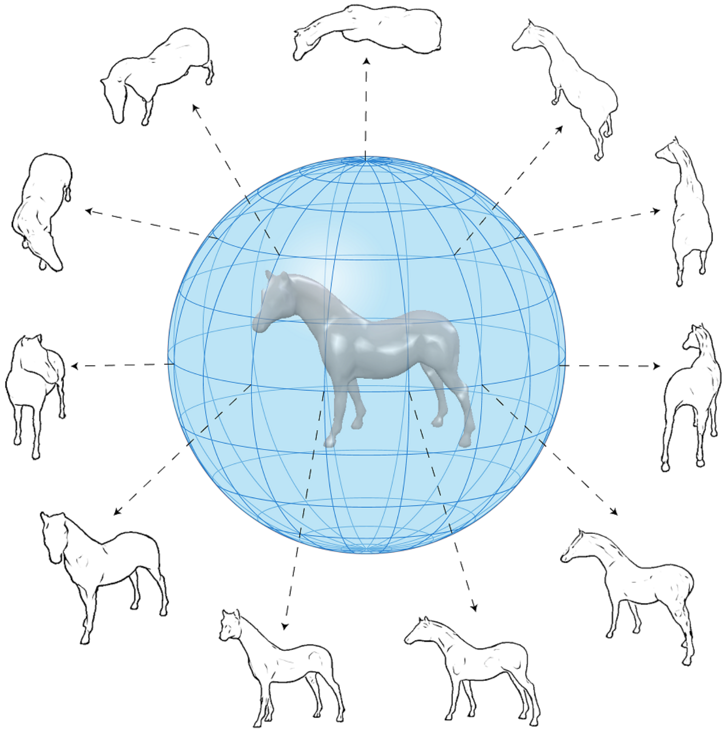

Figure 3: Viewpoint distribution over a spherical domain using an interval of 5 degrees in both longitude and latitude. For each viewpoint, a line drawing is generated using the CLD method [50]. Parts of this figure was previously published in [51] and is republished with permission by Springer.

3 A distributed computational cognitive model

Below we summarize the cognitive rules as indicated in Section 2, which serve as the guidelines for the cognitive model developed in this section.

- Rule 1. Human brain uses a pictorial representation for 3D objects.
- Rule 2. Either color photographs or line drawings trigger similar brain activation in a fast natural scene categorization task and we choose line drawings since they are much succinct in fast information encoding.
- Rule 3. The pictorial representation is encoded in a local, abstract form.
- Rule 4. The local and abstract forms are not static, but dynamically changed during a learning process.
- Rule 5. Object recognition is based on the partial match of abstract forms.

There are several types of cognitive models in the literature. One is a variety of symbolic models proposed in artificial intelligence that only provide a general description of the flow of the information process (e.g., [48]). This type of models need not rigorously matching the human data. In contrast, there are types of rigorous models of human cognitive process, in relation to human data in a quantitatively way; however, there are still arguments against these rigorous models [49]. Here, we state a cognitive model in the former type, i.e., the human cognitive process of object recognition is characterized by the following information processing flow:

- Perception. Human views 3D objects and triggers stimuli to visual cortex similar to the stimuli when viewing an appropriate line drawing.

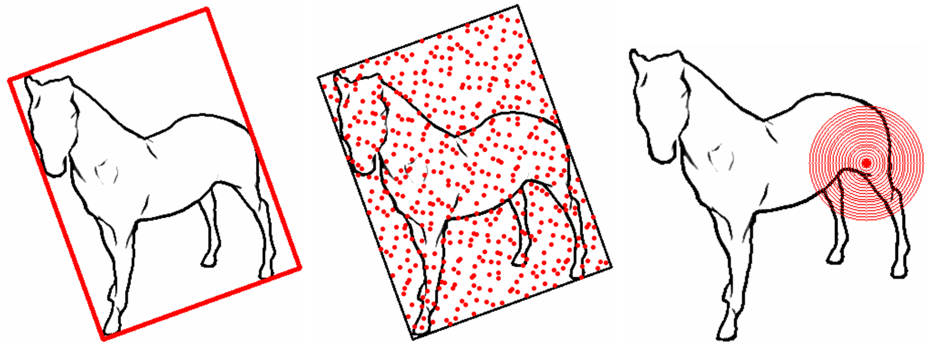

Figure 4: A local feature representation of line drawings. This figure was previously published in a preliminary conference paper [51] and is republished with permission by Springer.

- Memory. Object recognition is to identify an observed object from a set of known labels. The labels of known objects are stored in terms of local abstract forms extracted from line drawings.
- Learning. The known labels or local abstract forms in memory are not static but dynamically changed during a learning process when more 3D objects are perceived.
- Judgement. When a person views a novel 3D object, he/she converts the perceived stimuli into some similar local abstract forms stored in memory. By comparing the local abstract forms in an integrated way, the identity of the perceived object is determined.

We emphasize that our cognitive model is distributed, since individual local abstract forms act independently of each other and the whole model is a collection of independent local abstract forms that co-devote to the identification and recognition of the perceived object as a single coherent system. Meanwhile, these local abstract forms share a similar representation and communicate with each other during the learning process.

Below we present the computational implementation of this distributed cognitive model.

3.1 Perception

Given a viewpoint and viewing direction, a 3D object projects an image on the retina. To simulate this process, we place the 3D object model by coinciding the center of gravity with the center of a sphere which bounds the 3D object (Figure 3). A dense sampling of viewpoints are applied on the spherical surface in which the sampling density is 5 degrees in both longitude and latitude. For each viewpoint, a standard Lambertian light model is applied to generate a shading image of that object and this shading image is further converted into a line drawing using the CLD algorithm [50].

If we represent each line drawing by a 320×320 pixels, all line drawings sampling from the surrounding sphere is clearly a 2-manifold point cloud[4] embedded in a very high-dimensional (more than 100k) feature space. Many dimensional reduction methods can be applied here and in this study we apply a complexity-dependent clustering method (subsection 3.2) to obtain a small yet effective representative set of line drawings for each 3D object.

3.2 Memory

Given a pictorial representation P in terms of line drawings, let $B(P)$ be the minimum-area bounding box of all black pixels in P (Figure 4 left). The Halton's quasi-random point sequence is applied to uniformly sample n_p points in $B(P)$ (Figure 4 middle). At each sample point, a circular histogram is established

[4] This 2-manifold property come from the fact that each point of 320×320 dimensions is sampled in a two-dimension spherical domain and due to the viewpoint continuity, each two adjacent viewpoints devote to two adjacent points in the space of 320×320 dimensions.

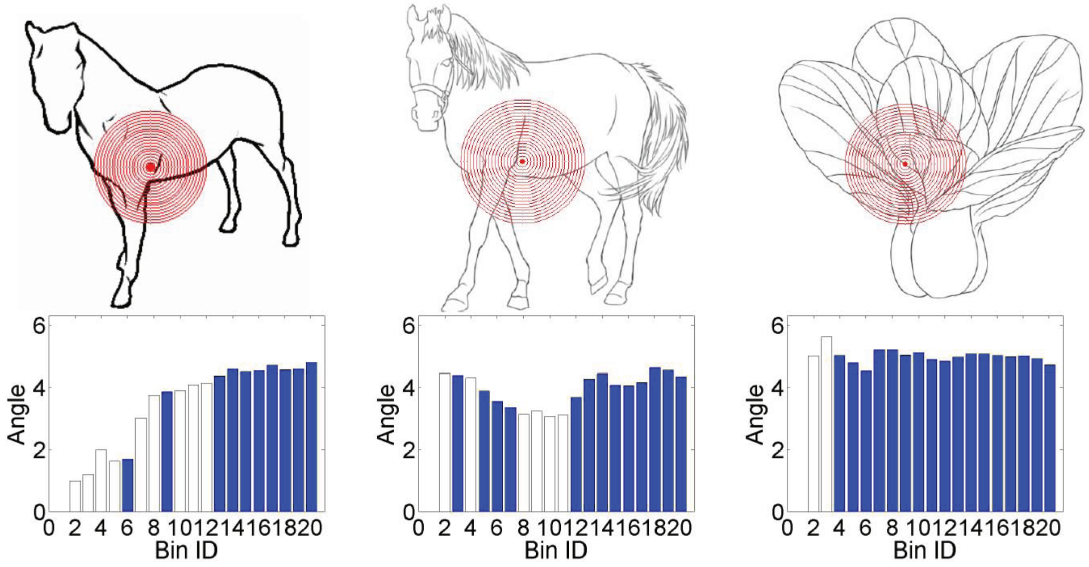

Figure 5: Three line drawings with sample points and corresponding feature histogram. The left horse line drawing is converted from the projection of a 3D object (Figure 3). The middle horse line drawing is provided by a paid artist (Figure 1). The right vegetable line drawing presents a different object representation. For the feature histograms, the blue solid bins are activated ones. Hollow bins are non-activated. Note that line drawings may have different line widths. This figure was previously published in a preliminary conference paper [51] and is republished with permission by Springer.

(Figure 4 right), in which each circular bin has the same difference of radii and the maximal circle has radius of one fifth of diagonal length of $B(P)$. The reasons that we use such a feature representation are:

- Random sampling provides a maximal entropy of point locations. For 3D object models, random sampling on object surfaces has been demonstrated to be an effective tool [4].

- Circular histogram makes the feature representation rotation invariant and less sensitive to the shape distortion: this is important since line drawings are not an accurate form and human used to matching them with elastic deformation (cf. two horses in Figure 5).

The feature histogram of each sample point has n_f bins. In the proposed computational model, there are many parameters including n_p and n_f, which can be adaptive and be large values (compare to 10^{11} neurons in human brain) to make a powerful discrimination capacity. In our experiment, however, $n_p = 300$ and $n_f = 20$ are sufficient to make the proposed computational model superior to other models in computer science area.

Denote by n_{tl} the total number of black pixels falling into the circular histogram. Different from the feature histogram used in [52], in this work we propose a saturation-firing strategy in the circular histogram as follows. For each bin, if the number of black pixels is larger than a threshold[5], then that bin is activated (akin to neuron firing). We index the bins $1, 2, \cdots, n_f$ from innermost to outmost and denote the first activated bin by b_f. Each black pixel in b_f corresponds to an angle in polar coordinate and the average of all angles of all black pixels in b_f is denoted by A_{aver}. We use A_{aver} to set up a local coordinate system for the feature histogram h. Each of all later activated bins (other than the first activated one) has an average angle and we use this angle as the bin values. Three examples are illustrated in Figure 5. Only activated bins are involved in the feature matching.

For feature histogram matching, due to inaccurate forms of line drawings, it is desired that a small deformed histogram has a small distance to the original one. For example, denote by h_a the histogram

[5] The threshold gives a measures of signal intensity and in our experiment we set it as 5% of n_{tl}.

which has nonzero bin values $\pi/2$ at bins 4 and 16, and suppose h_b has nonzero bin values $\pi/2$ at bins 5 and 17, and h_c has nonzero bin values $\pi/2$ at bins 4 and 5. Intuitively, h_a and h_b are similar and should have small distance value, while the distance between h_a and h_c should be large. However, the L_2 Euclidean distance $L_2(h_a, h_b) = \pi > L_2(h_a, h_c) = \sqrt{2}\pi/2$. In the proposed computational model, we apply a generalized distance [53]:

$$D(h_1, h_2) = h_1^T M_{n_f \times n_f} h_2 \tag{1}$$

where $M_{n_f \times n_f}$ is a SPD matrix whose elements are

$$m_{ij} = \frac{1}{2\pi\sigma^2} e^{-(i-j)^2/2\sigma^2}, \quad \sigma = 0.5$$

Using metric (1), $D(h_a, h_b) = 0.4252 < D(h_a, h_c) = 1.7834$, giving us desired results. For the three histograms shown in Figure 5, with metric (1), the distance between two horse line drawings is 209.3710, and the distances of vegetable to the two horses are 267.0254 and 297.2649, respectively.

Given a set M of 3D objects that a human has seen so far (or a database stored in a computer), let F be all feature histograms of all line drawings projected by these 3D objects. We apply the affinity propagation (AP) method [54] to classify F into m clusters, where the number m is automatically and optimally determined by the AP method. For each cluster, its center is selected as a codeword c_i and m clusters constitute a codebook $C = \{c_1, c_2, \cdots, c_m\}$. What stores in the computer memory for object recognition is the codebook. The codebook is not static but evolves in a learning process when the human see more 3D objects. We discuss this in the following section.

3.3 Learning

By learning human beings convert some short-term memory into long-term memory. Accordingly, the codebook $C = \{c_1, c_2, \cdots, c_m\}$ is not static, but evolves during a continuous learning process. The dynamically changes in the codebook C are subject to two factors:

- Human may see more 3D objects during a longer time. We thus model the number m of codes in the codebook C by a Poisson process $m(t)$.
- The codes in C are themselves not static and may convert into each other (some may be lost and others may be enhanced by learning). We thus model the transience (short-term) and stationary distribution (long-term) in the codes by a state space in a continuous-time Markov chain.

We explain these two factors in detail below. Poisson distribution is based on the assumption that for small time interval, the probability of an arrival (a new code is generated in our case) is proportional to the length of waiting time. The set of Poisson points on the time t-axis satisfies the following requirements.

- The number $n(t_1, t_2)$ of the Poisson points in a time interval (t_1, t_2) of length $t = t_2 - t_1$ is a Poisson random variable with parameter λt;
- If the intervals (t_1, t_2) and (t_3, t_4) do not overlap, then $n(t_1, t_2)$ and $n(t_3, t_4)$ are independent.

By modeling the number $m(t)$ of codes as increased by one at each location of Poisson points along the time axis, $m(t) = n(0, t)$ is a stochastic Poisson process, where the parameter λ gives a measure of the learning speed.

As a special continuous-time Markov chain, the Poisson process ia a natural probability model for any steam of independent discrete events in continuous time. For a fixed m, continuous-time Markov chain is also used to model the codes' transience and stationary distribution. The codebook $C = \{c_1, c_2, \cdots, c_m\}$ is treated as a state space and each code $c_i \in C$ is a state. Let $\lambda = \{\lambda_i : c_i \in C\}$ be a possibility measure on C, where $\lambda_i \geqslant 0$ is the possibility of code c_i being representative, $\sum_{c_i \in C} \lambda_i = 1$ and thus λ defines a distribution of a random state C.

To model the dynamically changed codebook, a continuous time t is added to the distribution $(\lambda_t)_{t \geqslant 0} = (X_t : 0 \leqslant t < \infty)$, which describes a continuous-time random process. To handle the discrete events occurring in continuous time, we restrict our attention to right-continuous process $(x_t)_{t\geqslant 0}$, i.e., for all i with which $c_i \in C$ and $t \geqslant 0$, $\exists \varepsilon > 0$ such that $C_s(i) = X_t(i)$ for $t \leqslant s \leqslant t+\varepsilon$.

Let H be all feature histograms in line drawings $L = \bigcup_i L(m_i), i = 1, 2, \cdots, n$. H is mapped to the clusters each of which corresponds to a code, and gives rise to a coded representation $C(H) = \{n_1(H)c_1, n_2(H)c_2, \cdots, n_m(H)c_m\}$, where there are $n_i(H)$ feature histograms $H_i = \{h_1^i, h_2^i, \cdots, h_{n_i(H)}^i\}$ falling into the cluster corresponding to the code c_i. The initial distribution of C_0 is given by $\lambda_0 = \{n_1(H)/n(H), n_2(H)/n(H), \cdots, n_m(H)/n(H)\}$, $n(H) = n_1(H) + n_2(H) + \cdots + n_m(H)$. With λ_0, the generator matrix Q determines how the continuous-time Markov chain $(X_t)_{t\geqslant 0}$ evolves from its initial state. The Q-matrix $\{q_{ij}\}_{1\leqslant i,j\leqslant m}$ is determined by its associated jump matrix $\Pi = \{\pi_{ij}\}_{1\leqslant i,j\leqslant m}$ given by

$$\pi_{ii} = \begin{cases} 0 & \text{if} \quad q_i \neq 0 \\ 1 & \text{if} \quad q_i = 0 \end{cases}$$

$$\pi_{ij} = \begin{cases} q_{ij}/q_i & \text{if} \quad j \neq i \quad \text{and} \quad q_i \neq 0 \\ 0 & \text{if} \quad j \neq i \quad \text{and} \quad q_i = 0 \end{cases}$$

Here, an important note is that the codes in C are not static and may have chance to convert into each other. The transition probability π_{ij} between any two codes $c_i, c_j \in C$ is defined by their normalized similarity:

$$\pi_{ij} = \frac{c_{ij}}{\sum_{j=1}^{m} c_{ij}}$$

where

$$\begin{aligned} c_{ij} &= M_{ij} - \tfrac{1}{n_i(H)} \tfrac{1}{n_j(H)} \textstyle\sum_{f_u \in F_i} \sum_{f_v \in F_j} D(f_u, f_v), \\ M_{ij} &= \max\{D(f_u, f_v), \forall f_u \in F_i, \forall f_v \in F_j\} \end{aligned}$$

and $D(,)$ is defined in Eq. (1). The jump matrix Π is stochastic, with which after an exponential time of parameter q_i, the code c_i jumps to a new state of code c_j with probability π_{ij}. When the continuous-time Markov chain converges to an equilibrium, the stationary distribution λ_∞ offers a measure of importance of codes in C.

Let $J_0, J_1, \cdots$ be the jump times of $(X_t)_{t\geqslant 0}$ and define the jump process of $(X_t)_{t\geqslant 0}$ by a discrete-time process $(Y_n)_{n\geqslant 0}$, where $Y_n = X_{J_n}$. $(Y_n)_{n\geqslant 0}$ is a discrete-time Markov chain (λ_0, Π). Since each code is obtained by clustering feature histograms, each code has at least one occurrence and all π_{ij} are strictly positive. Accordingly, Π is irreducible. On the other hand, since $\pi_{ij} > 0$, all states in C are reachable and there must exist some state that is positive recurrent. Given these two properties, $(Y_n)_{n\geqslant 0}$ is ergodic and the stationary distribution λ_∞ can be directly computed by solving the left-eigenvector problem $\lambda_\infty \Pi = \lambda_\infty$.

3.4 Judgment

Recall that for each line drawing projected by a 3D object, n_p sample points are sampled in a Halton's quasi-random sequence and a feature histogram is generated for each point. Given the codebook $C = \{c_1, c_2, \cdots, c_m\}$, a line drawing l can be presented by $l = (n_1c_1, n_2c_2, \cdots, n_mc_m)$, $n_1 + n_2 + \cdots + n_m = n_p$, by finding the clusters (corresponding to codes) in which the feature histograms of l fall; e.g., n_i means that there are n_i feature histograms of l falling into the cluster represented by the code c_i.

Let m_j be a 3D model in the database $M = \{m_1, m_2, \cdots, m_n\}$ and $L(m_j)$ be all line drawings of m_j. The frequency of occurrence of the code c_i in the model m_j, denoted by $f_{i,j}$, is the number of times that c_i appears in $L(m_j)$. Not all codes in C have equally usefulness for describing a 3D model. Obviously, the larger $f_{i,j}$ is, the more important the code c_i is to the model m_j. On the other hand, if a code c_j appears in most or all models in M, then it is less discriminative than a code that just appears in a few models in M. We thus use the TF-IDF weight [55]

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{n}{r_i} & \text{if} \quad f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $w_{i,j}$ is the weight assigned to the pair (c_i, m_j), r_i is the number of 3D models in M in which a code c_i occurs, and n is the total number of models in M.

Given the codebook $C = \{c_1, c_2, \cdots, c_m\}$ and the TF-IDF weights, a 3D model m_j in M is encoded as an m-vector

$$C(L(m_j)) = (w_{1,j}, w_{2,j}, \cdots, w_{m,j})$$

Given a query model q, after the same encoding process

$$C(L(q)) = (w_{1,q}, w_{2,q}, \cdots, w_{m,q}),$$

the models in M are sorted by the similarity value

$$S(m_j, q) = \frac{C(L(m_j)) \cdot C(L(q))}{\|C(L(m_j))\| \|C(L(q))\|} \tag{2}$$

and the semantic tags in the model that has the largest similarity value are used for recognizing the query model. Note that in the similarity metric (2), the encoding vectors are normalized since different models may have different complexity and thus have different vector magnitudes.

4 Experiment

One of the main goals of the proposed computational cognitive model is to explore the high-recognition-performance mechanism of the human brain, and use the key components in this mechanism to develop efficient computer programs that should have some advantages over the state-of-the-art recognition algorithms in computer science area. In this experiment, we compare the computational cognitive model to four classic methods (EGI, SPIN, D2 and G2) and one state-of-the-art method (VSKL):

- EGI [1]. The extended Gaussian image (EGI) maps each vertex's normal of a surface patch to the Gaussian sphere, weighted by the triangle area for each normal. This map has a clear relation to the surface Gaussian curvature and can be used for object recognition. Especially, if the object is convex, the EGI can uniquely determine the object shape.

- SPIN [2]. This method uses a dense collection of surface points and associated normals to represent 3D shape. For each point, a local orientation is established using the normal information and other points that are visible to this point (the angle of two normals of two points is less than a threshold) are projected into the tangent plane and form a spin image of that point. PCA method is applied to compress all spin images of the 3D shape that are used for shape recognition.

- D2 [4]. By randomly sampling a set of points on the object surface, the pair-wise Euclidean distances between each two in the sampling set are used in this method as a shape signature for object recognition.

- G2 [3]. The Euclidean distance in D2 is affected by the object's embedding space. Elbaz et al. [3] proposed the use of the intrinsic geodesic distance between any two sample points as a signature for object recognition.

- VSKL [5]. With the aid of the intrinsic geodesic distance field, this method samples a set of salient points on the object surface and builds an intrinsic Voronoi diagram of these sample points. The inherent geodesic Reeb graph is then extracted and forms a 1D tree-like skeleton. The 1D skeleton has been shown [5] to be effective for object recognition.

We use McGill 3D Shape Benchmark [8] to test the above methods and our computational cognitive model. The McGill Benchmark contains articulated and non-articulated objects. Ten classes are included in articulated objects: they are ants (30), crabs (30), hands (20), humans (30), octopus (25), pliers (20), snakes (25), spectacles (25), spiders (31), and teddy (20). The number in the bracket is the number of

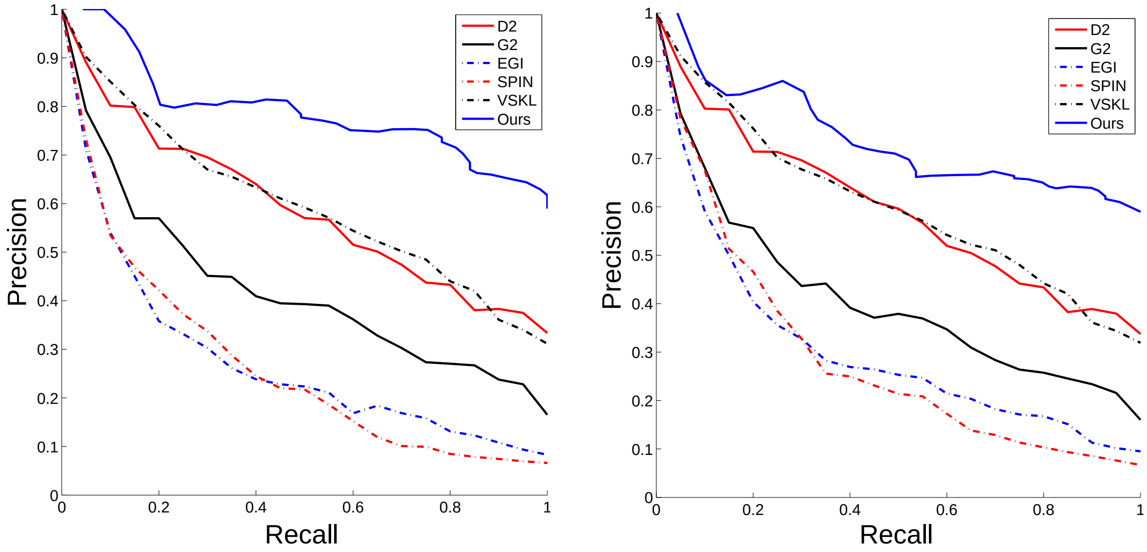

Figure 6: The PR curves of five methods (EGI, SPIN, D2, G2, VSKL) and the proposed computational cognitive model. Left: tested on the McGill 3D Shape Benchmark. Right: tested on the noised McGill 3D Shape Benchmark.

models in that class. Nine classes are included in non-articulated objects: they are airplanes (26), birds (21), chairs (23), cups (25), dinosaurs (19), dolphins (12), fishes (23), four-limbs (31), and tables (22). To test the learning mechanism in the proposed computational cognitive model, we generate a codebook $C = \{c_1, c_2, \cdots, c_m\}$ using all 3D objects in the McGill Benchmark. Then we compute the associated jump matrix $\Pi = \{\pi_{ij}\}_{1 \leqslant i,j \leqslant m}$ for the codes in C. The values in the stationary distribution λ_∞ (i.e., the left eigenvector of Π) is used as an index to rank the codes in C.

We use the precision and recall (PR) metric to compare the different recognition methods. Let I be a 3D object in a class C_i of the McGill Benchmark. We use I as input to get a set of recognized objects R. Ideally, if R is much closer to C_i, we can obtain better recognition performance. The precision value is defined as $p = \frac{|R \cap C_i|}{|R|}$ and the recall value is defined as $r = \frac{|R \cap C_i|}{|C_i|}$, where $|S|$ is the cardinality of set S. We rank the recognized objects and define the set R to be top matched objects with increased set cardinality. We use each of the 3D objects in the benchmark in turn as input and the final PR curve is the average of all individual PR curves. The corresponding PR curves of different methods are summarized in Figure 6 (left). The data presented in Figure 6 (left) clearly shows that the proposed computational cognitive model has the best recognition performance.

A good object recognition method should also have a high robustness against noises. Accordingly, we generate a noised version of the McGill Benchmark by adding a Gaussian noise (with the maximum magnitude of 10% the diagonal length of the object's bounding box) to each vertex in the 3D object's mesh. We then apply different recognition method on this noised Benchmark and the corresponding PR curves are summarized in Figure 6 (right). The data presented in Figure 6 (right) demonstrates that the added noise has little influence on our methods as well as D2 and VSKL. Meanwhile, G2, SPIN and EGI still have a low accuracy when dealing with the noised database. This observation meets our expectation since both the geodesic distances [56] along the object surface and the vertex normals [1] on the object surface seriously suffer from noise.

5 Conclusions

In this paper, we present a computational cognitive model based on a comprehensive review on cognitive functionalities in the human brain for 3D object recognition. Compared to previous object recognition methods in computer science area, the proposed cognitive model has several distinct features: (1) A distributed pictorial representation in terms of line drawings is used to represent 3D objects; (2) the 3D object information is locally encoded by a set of points randomly sampled from line drawings and each sample point is associated with a circular histogram that is similarity-invariant; (3) from the large set of feature histograms, a distributed abstract form called codebook is extracted by optimal feature clustering; (4) for the codebook representation, learning mechanism is exploited by applying a Poisson process and a Markov chain model. Object recognition performance is tested using the McGill 3D Shape Benchmark and the comparison between our method and five representative methods in computer science area shows that the proposed computational cognitive model has a better performance than these representative recognition methods. Finally we remark that the line drawing representation proposed in this paper may be intrinsic in the human brain for fast visual media understanding, and thus in addition to being useful for a natural and efficient 2D/3D geometric modeling [57, 58], it also sheds some light on a promising multimedia (including image and video) authoring and summarization scheme [59, 60].

Acknowledgements

This work was supported in part by the National Basic Research Program of China (2011CB302202), the Natural Science Foundation of China (Grant Nos. 61272228, 31270024) and the 863 program of China (2012AA011801). The work of Y.J. Liu was supported in part by the Program for NCET-11-0273 and TNList Cross-discipline Foundation.

References

1 Horn B. Extended Guassian images. Proc. IEEE, 1984, 72: 1671–1686

2 Johnson A, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Analysis and Machine Intelligence, 1999, 21: 433–449

3 Elbaz A, Kimmel R. On bending invariant signatures for surfaces. IEEE Trans. Pattern Analysis and Machine Intelligence, 2003, 25: 1285–1295

4 Funkhouser T, Min P, Kazhdan M, Chen J, Halderman A, Dobkin D, Jacobs D. A search engine for 3D models. ACM Trans. Graphics, 2003, 22: 83–105

5 Liu Y, Chen Z, Tang K. Construction of iso-contours, bisectors, and Voronoi diagrams on triangulated surfaces. IEEE Trans. Pattern Analysis and Machine Intelligence, 2011, 33: 1502–1517

6 Goodale M, Milner A. Separate visual pathways for perception and action. Trends in Neurosciences, 1992, 15: 20–25

7 Fu X, Cai L, Liu Y, et al. A computational cognition model of perception, memory, and judgment. Sci China Ser F-Inf Sci, 2013, 56(5): , doi: 10.1007/s11432-009-0095-8.

8 McGill 3D Shape Benchmark, http://www.cim.mcgill.ca/∼shape/benchMark/

9 Kanwisher N, McDermott J, Chun M. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. Journal of Neuroscience, 1997, 17: 4302–4311

10 Puce A, Allison T, Gore J, McCarthy G. Face-sensitive regions in human extrastriate cortex studied by functional MRI. Journal of Neurophysiology, 1995, 74: 1192–1199

11 Epstein R, Harris A, Stanley D, Kanwisher N. The parahippocampal place area: recognition, navigation, or encoding? Neuron, 1999, 23: 115–125

12 Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature, 1998, 392: 598–601

13 O'Craven K, Kanwisher N. Mental imagery of faces and places activates corresponding stiimulus-specific brain regions. Journal of Cognitive Neuroscience, 2000, 12: 1013–1023

14 Malach R, Reppas J, Benson R, Kwong K, Jiang H, Kennedy W, Ledden P, Brady T, Rosen B, Tootell R. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences, USA, 1995, 92: 8135–8139

15 Haxby J, Gobbini M, Furey M, Ishai A, Schouten J, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 2001, 293: 2425–2430

16 Ishai A, Ungerleider L, Martin A, Schouten J, Haxby J. Distributed representation of objects in the human ventral visual pathway. Proceedings of the National Academy of Sciences, USA, 1999, 96: 9379–9384

17 Biederman I. Recognition-by-components: a theory of human image understanding. Psychological Review, 1987, 94: 115–147

18 Tarr M, Williams P, Hayward W, Gauthier I. Three-dimensional object recognition is viewpoint dependent. Nature Neuroscience, 1998, 1: 275–277

19 Cahill L, McGaugh J. Mechanisms of emotional arousal and lasting declarative memory. Proceedings of the National

Academy of Sciences, USA, 1992, 89: 60–64

20 Jolicoeur P. Orientation congruency effects on the indentification of disoriented shapes. Journal of Experimental Psychology: Human Perception and Performance, 1990, 16: 351–364

21 Tarr M, Pinker S.Mental rotation and orientation-dependence in shape recognition. Cognitive Psychology, 1989, 21: 233–282

22 Haxby J, Ishai A, Chao L, Ungerleider L, Martin A. Object-form topology in the ventral temporal lobe. Trends in Cognitive Sciences, 2000, 4: 3–4

23 Walther D, Chai B, Caddigan E, Beck D, Fei-Fei L. Simple line drawings suffice for functional MRI decoding of natural scene categories. Proceedings of the National Academy of Sciences, USA, 2011, 108: 9661–9666

24 Liu Y, Luo X, Xuan Y, Chen W, Fu X. Image retargeting quality assessment. Computer Graphics Forum (regular issue of Eurographics 2011), 2011, 30: 583–592

25 Biederman I, Ju G. Surface versus edge-based determinants of visual recognition. Cognitive Psychology, 1988, 20: 38–64

26 Mehta R, Zhu R. Blue or Red? Exploring the effect of color on cogntive task performances. Science, 2009, 323: 1226–1229

27 Liu Y, Zheng Y, Lv L, Xuan Y, Fu X. 3D Model Retrieval based on Color+Geometry Signatures. The Visual Computer, 2012, 28: 75–86

28 Fu Q, Liu Y, Chen W, Fu X. The time course of natural scene categorization in human brain: simple line-drawings vs. color photographs. Journal of Vision, 2013, to appear.

29 Davenport J, Potter M. Scene consistency in object and background perception. Psychological Science, 2004, 15: 559–564

30 Peelen M, Fei-Fei L, Kastner S. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature, 2009, 460: 94–97

31 Walther D, Caddigan E, Fei-Fei L, Beck D. Natural scene categories revealed in distributed patterns of activity in the human brain. Journal of Neuroscience, 2009, 29: 10581–10573

32 Bar M. Visual objects in context. Nature Reviews Neuroscience, 2004, 5: 617–629

33 McClelland J, Rumelhart D. Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 1985, 114: 159–188

34 Medin D, Schaffer M. Context theory of classification learning. Psychological Review, 1978, 85: 207–238

35 Possner M, Keele S. Retention of Abstract Ideas. Journal of Experimental Psychology, 1970, 83: 304–308

36 Chklovskii D, Mel B, Svoboda K. Cortical rewiring and information storage. Nature, 2004, 431: 782–788

37 McGaugh J. Memory – a century of consolidation. Science, 2000, 287: 248-251

38 Bulthoff H, Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Trends in Neurosciences, 1998, 21: 294–299

39 Trachtenberg J, Chen B, Knott G, Feng G, Sanes J, Welker E, Svoboda K. Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. Nature, 2002, 420: 788–794

40 Wallis G, Bulthoff H. Learning to recognize objects. Trends in Cognition Sciences, 1999, 3: 22–31

41 Schyns P. Categories and percepts: a bi-directionnal framework for categorization. Trends in Cognitive Sciences, 1997, 1: 183–189

42 Miyashita Y. Neural correlate of visual associative long-term memory in the primate temporal. Nature, 1988, 335: 817–820

43 Miyashita Y. Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience, 1993, 16: 245–263

44 Stryker M. Temporal associations. Nature, 1991, 354: 108–109

45 Tanaka K. Inferotemporal Cortex and Object Vision. Annual Review of Neuroscience, 1996, 19: 109–139

46 Leopold D, O'Toole A, Vetter T, Blanz V. Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience, 2001, 4: 89–94

47 Pellicano E, Rhodes G. Holistic Processing of Faces in Preschool Children and Adults. Psychological Science, 2003, 14: 618–622

48 Anderson J. The Architecture of Cognition. Harvard University Press, Cambridge, MA, 1983.

49 Massaro D. Some criticisms of connectionist models of human performance. Journal of Memory and Language, 1988, 27: 213-234

50 Kang H, Lee S, Chui C. Coherent Line Drawing. ACM Symposium on Non-Photorealistic Animation and Rendering (NPAR), 2007, 43–50

51 Liu Y, Fu Q, Liu Y, Fu X. 2D-Line-Drawing-Based 3D Object Recognition. CVM 2012, LNCS 7633, Springer, 2012, 146-153

52 Liu Y, Luo X, Joneja A, Ma C, Fu X, Song D. User-adaptive sketch-based 3D CAD model retrieval. IEEE Transactions on Automation Science and Engineering, 2013, 10: , DOI: 10.1109/TASE.2012.2228481.

53 Wang L, Zhang Y, Feng J. On the Euclidean distance of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27: 1334–1339

54 Frey B, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976

55 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval, ACM Press, 1999

56 Liu Y. Exact geodesic metric in 2-manifold triangle meshes using edge-based data structures. Computer-Aided Design, 2013, 45: 695–704

57 Ma C, Liu Y, Yang H, Teng D, Wang H, Dai G. KnitSketch: a sketch pad for conceptual design of 2D garment patterns. IEEE Transactions on Automation Science and Engineering, 2011, 8: 431–437

58 Liu Y, Ma C, Zhang D. EasyToy: Plush toy design using editable sketching curves. IEEE Computer Graphics and Applications, 2011, 31: 49–57

59 Ma C, Liu Y, Wang H, Teng D, Dai G. Sketch-based annotation and visualization in video authoring. IEEE Transactions on Multimedia, 2012, 14: 1153–1165

60 Ma C, Liu Y, Fu Q, Liu Y, Fu X, Dai G, Wang H. Video sketch summarization, interaction and cognition analysis. Sci China Ser F-Inf Sci (in Chinese), 2013, DOI: 10.1360/112013-1.