Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Perspective:

On visual knowledge

Yun-he PAN
Zhejiang University, Hangzhou 310027, China
E-mail: panyh@cae.cn

https://doi.org/10.1631/FITEE.1910001

This paper presents the concept of "visual knowledge." Visual knowledge is a new form of knowledge representation that is different from all other visual representations or knowledge representations that have emerged in artificial intelligence (AI) development. A visual concept is composed of prototypes, category structures, hierarchical structures, action structures, etc. It can further constitute a visual proposition, incorporating scene structures and their dynamics, and the visual proposition can then be used to narrate a visual scene. This paper suggests that careful utilization of developments from computer graphics technology will contribute to realizing visual knowledge representation, and to its reasoning and analysis, and that careful utilization of progression from computer vision will promote the learning of visual knowledge. Representation, reasoning, learning, and utilization of visual knowledge will form a key step toward remarkable breakthroughs in the era of AI 2.0.

1 Impact of visual knowledge on the development of AI

1.1 Upsurge in AI due to rapid upgrading of image recognition

In recent years, the rapid upgrading of image recognition has promoted an upsurge in AI. For instance, in 2012, the convolutional neural network AlexNet famously won the 2012 ImageNet LSVRC-2012 competition by a considerable margin (error rate of 15.3% vs. 26.2% (second place)). This established deep learning as a focal interest in academia and industry. In May 2016, the White House published an article entitled "Preparing for the Future of Artificial Intelligence." It claimed that in view of the unprecedented influence of AI on social life in the form of medical science, image and speech understanding, and so forth, a subcommittee for AI and machine learning was chartered by the National Science and Technology Council (NSTC) to coordinate and guide the development of AI technologies in industries, research communities, and Federal Government.

The breakthroughs in image recognition technology not only improve their precision in recognizing human faces, image captions, fingerprints, biological characteristics, and medical images, but have also promoted the development of intelligent cars, security monitoring, intelligent traffic systems, robots, unmanned aerial vehicles, intelligent manufacturing, and more. Fig. 1 shows the number of AI-related companies in China and the USA according to the Chinese Academy of Science and Technology for Development. More than half of the companies are relevant to image recognition.

1.2 Deep learning based on multi-layer neural networks as a new means of knowledge representations

Traditional image recognition is based on image processing techniques that date back to the transmission of digitized newspaper images via submarine cable lines between London and New York in 1920. In 1977, Rafael C. Gonzalez systematically described contemporary developments

Prof. Yun-he PAN
Editor-in-Chief

in all mainstream areas of digital image processing, such as image editing, transformation, enhancement, segmentation, boundary detection, and object localization, in his textbook *Digital Image Processing*. Based on this work, the classical image recognition and computer vision technology was developed.

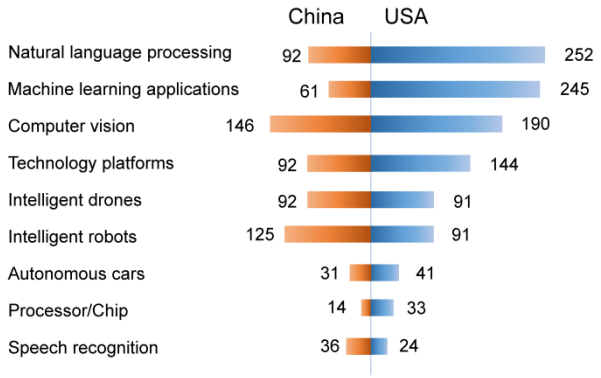

Fig. 1 Number of AI companies in China and the USA

The use of convolutional neural networks (CNNs) has made changes to the traditional methods of computer vision. For feature learning in terms of data-driven learning, CNNs can effectively improve the precision of image classification and recognition processes. In AI, the success of CNNs can be attributed to successful application of the new methods of knowledge representation. Before CNNs, general knowledge representation in AI was symbolic, which included rules, frameworks, and semantic networks. Deep learning, however, achieves a weighted-network knowledge representation, e.g., using network structures and connection weights in CNNs.

There are two advantages to CNNs: (1) They can learn automatically from a labeled amount of data; (2) They can be applied to the recognition of non-symbolic information, e.g., the recognition of images and speech.

However, CNNs have certain limitations as well. They are (1) non-interpretable, (2) incapable of reasoning, and (3) the large-scale data-driven approach used to train network parameters, and the large quantities of data involved inevitably introduce data biases (Hutchinson and Mitchell, 2019). From another perspective, the limitations in knowledge representation via CNNs lead us to investigate another form of knowledge representation, i.e., visual knowledge. Our motivations are as follows: (1) Visual research is an important field for promoting the development of AI; (2) Better methods of knowledge representation are critical for promoting visual research; (3) Overcoming the defects in deep neural networks is essential to visual knowledge.

2 Irreplaceable characteristics of visual knowledge

2.1 Extensive research in cognitive science on the importance and uniqueness of visual cognition

In the 1970s, cognitive psychologists performed a series of experiments on visual memory, and subsequently demonstrated that visual memory can be rotated, folded, scanned, and drawn as an analogy, which is different from verbal memory. The experiments are listed below (Anderson, 1989).

1. Mental rotation tests of R. N. Shepard in 1971

This experiment specifically tested mental rotation of three-dimensional (3D) objects. Each participant was presented with multiple pairs of 3D objects and asked to decide whether the pairs represented the same object or two different objects. The research demonstrated that the reaction time required for participants to determine whether or not the pair of items matched was linearly proportional to the angle of rotation from the original position. That is, the greater the rotation of an object from the original position, the longer it takes an individual to determine whether the two images are the same object.

2. Psychological experiments of mental paper folding of Shepard et al. in 1972

Each participant viewed one of the patterns of six connected squares that result when the faces of a cube are unfolded onto a flat surface. The participants decided as quickly as possible whether or not two arrows, each marked on the edge of a (different) square, would meet if the squares were folded back into a cube. The experiment showed that the participants could fold the visual memory in their mind. The average time required to make such decisions is a function of the total number of squares involved in each fold, if those folds were actually performed physically.

3. Mental image scanning experiments of Kosslyn et al. in 1978

In this experiment, participants were asked to shift their attention from one object in a picture to

another object. It turned out that participants took significantly longer to shift attention between more distant objects (for example, the hut and the rock) than between more proximal objects (for example, the hut and the well). The mental scanning experiment revealed a strong linear correlation between the time required to scan two objects in a mental image and the distance between those two objects in the picture.

4. Experiments of animal size comparison in memory by Moyer in 1973

Humans can compare objects by retrieving information from memory. In this experiment, subjects were asked to compare the relative sizes of animals when they heard pairs of animal names. It was found that the reaction time decreased as the difference between the rated sizes of the animals increased.

The above-mentioned visual memory is termed "mental imagery" in cognitive psychology. Mental imagery is a knowledge form of imagery thinking (Pan, 1991). In AI, visual mental imagery is referred to as visual knowledge.

2.2 Characteristics of visual knowledge

The psychology experiments described above demonstrated that visual knowledge in human memory has a set of characteristics: (1) It can describe the relation between spatial shapes, sizes and correlation, as well as colors and textures. (2) It can describe the relations between action, speed, and time in terms of objects. (3) It can conduct temporal–spatial transformations and reasoning, including shape transformations, action transformations, speed transformations, scene transformations, various temporal–spatial analogies and associations, and predictions based on temporal–spatial reasoning results.

Cognitive psychology also indicates that the quantity of visual knowledge present in human memory is larger than that of verbal knowledge, and that understanding of verbal knowledge further requires assistance from this visual knowledge (Horoufchin et al., 2018; Kosiorek et al., 2019). Due to the difficulty in its expression via linguistic symbols, visual knowledge had been classified as common sense. In the past years, it has been observed that one of the weaknesses of AI research is the lack of studies on visual knowledge. Therefore, the study and application of visual knowledge will be an important direction in which to develop AI 2.0 (Pan, 2016).

3 Representation and operation of visual knowledge—reconstruction based on computer graphics technologies

After extensive research, computer graphics design has accumulated numerous 3D-shape expressions and operation algorithms, e.g., half-edge data structure, geometric transformations, Euler operations, and projection transformations (Pan et al., 2011), as well as techniques of animation and morphing. They provide a technical basis for visual knowledge representation and operation. However, carrying out the representation and reasoning of visual knowledge requires that we rebuild and reconstruct the research in computer graphics.

3.1 Graphical expressions reconstructed as visual concepts

3.1.1 Visual concepts

Graphical expressions should be reconstructed as visual concepts. A visual concept usually consists of prototype and category. For instance, apples have various shapes, but they have only a handful of core shapes and colors that are named prototypes. Based on the core shapes and colors, any apple can be generated. There is a margin of variation among the core shapes of apples. Shapes within the margin belong to the category of apples, but the shapes outside it fall into the categories of shapes of other fruits. As a result, visual concept={prototype, category}.

Categories can be expressed not only as a margin of variation of different parameters but also as a synthesis field consisting of a typical and several atypical shapes and colors (Pan, 1996).

3.1.2 Hierarchical structures of visual concepts

The visual concept should contain sub-concepts, i.e., the spatial structure of concepts. For instance, an apple is a structure that is composed of several sub-concepts, such as kernels, fruit flesh, rind, and pedicel. The visual concept needs to express the relations of spatial structures of these sub-objects; e.g., spatial structural expressions of apples are useful for solving problems in botany, agriculture, and food science.

3.1.3 Action structures of visual concepts

The visual concept for some objects, such as animals, should contain actions, consisting of typical

motion and action categories of each sub-concept within a structure, e.g., expression of actions for the animal's head, limbs, body, or claws (fingers), as well as their relations.

3.2 Visual operation and reasoning

Visual operations and reasoning include the decomposition, replacement, combination, deformation, motion, comparison, damage, repair, and prediction of shapes. They need to be reconstructed based on the techniques of graphical transformation, deformation, animation, etc.

4 Learning of visual knowledge: reconstruction based on computer vision

Visual knowledge learning is a prominent issue to be addressed in the construction and application of visual knowledge systems. Despite some methods being proposed to enable the bottom-up learning of visual knowledge (Greff et al., 2019; Zhao et al., 2019), the field currently lacks a method to systematically learn visual knowledge as the building of knowledge often requires a top-down design. During this process, the achievements of computer vision, e.g., 3D rebuilding, are beneficial to the systematic learning of visual knowledge.

4.1 Visual concept learning

After long research, computer vision has accumulated many techniques, including the technique to restore the shape of a 3D object from multiple images, to scan a real object with a 3D scanner and to acquire the scanning point cloud to create a 3D shape grid. It can also be used to capture a real object (as well as its actions) synchronously using video cameras from different angles and calculating the corresponding points of the same objects to rebuild and animate the shape and actions of a 3D object (Ma and Zhang, 1998).

Visual knowledge learning requires not only reconstruction of visual shapes but also reconstruction of visual knowledge. This requires further study on the present technique of computer vision to accomplish the following goals: (1) to reconstruct not only 3D shapes but also hierarchical structures of 3D shapes and (2) to not only reconstruct 3D shape structures but also know where they fit in the concept category, e.g., typical or non-typical concepts.

4.2 Learning the representation of visual propositions

Besides visual concepts, the representation and learning of a visual narrative need to be studied. A visual narrative consists of a series of visual propositions that describe the spatial and temporal relations of the visual concepts.

The spatial relations are represented as scene structures that describe relations of location, distance, inside/outside, up/down, left/right, and front/back, among objects.

The temporal relations are represented as dynamic structures that describe the evolution, displacement, action, variation, competition, and cooperation of different objects.

An image corresponds to a visual proposition that describes a given scene structure. A video is a typical visual narrative that corresponds to a sequence of visual propositions including both scene structures and potential dynamic structures. Silent films have demonstrated a strong ability for expressing a visual narrative. The task for visual knowledge learning also includes the automatic learning of knowledge of visual propositions from visual concept knowledge, and the automatic learning of visual narrative knowledge from visual propositions.

5 Usage analyses of visual knowledge

Visual knowledge has many potential applications in current AI development. Below we give three examples to illustrate various means of visual knowledge applications: recognition by generation, reconstruction by knowledge, and design by knowledge.

Example 1 (Recognition by generation, or image recognition via visual knowledge, e.g., recognition of a cat) The process can be summarized as follows: (1) According to the prototypes and categories of the visual object cat, we use synthesis reasoning to generate positive and negative samples of cat images (Pan, 1996). (2) We train a deep neural network based on the generated positive and negative samples. (3) The trained deep neural network is used to recognize images of cats.

Example 2 (Reconstruction by knowledge) Visual knowledge is applied to 3D reconstruction, e.g., reconstruction of dressed body model. (1) We use computer vision technology to obtain body features of one person. (2) We retrieve the parameters of a similar 3D body model from body concept and use these parameters to generate a 3D body model. (3) We iteratively modify the generated 3D body model until the parameters of the generated 3D body are the same as the parameters obtained from the person. Clearly, the aforementioned process is "knowledge transfer learning," which significantly reduces the potential information overload of visual knowledge learning.

Example 3 (Design by knowledge) Visual knowledge is applied to designs, e.g., the design of characters, which is needed for games, movies, and advertising. (1) Determine the requirements of the characters, e.g., a group of young men jumping with various statures, weights, muscularities, and physical agilities, from various states of vigorous walking and running. (2) According to the prototype of young men with various characteristics in the visual concept, images of jumping men with various statures, jumping times, and jumping locations are generated using the vigorous type direction of the category. (3) Users provide advice on the jumping men generated. Based on the advice, the category is adjusted to again generate the images of jumping men using step (2) until the users feel satisfied.

6 Conclusions

From the above analyses, the unique advantages of visual knowledge are the ability to generate by synthesis, align the temporal–spatial correlation, and visualize non-symbolic information. Symbolic knowledge lacks these advantages. Visual knowledge can be used in creation, prediction, and human–computer integration. Therefore, the study of visual knowledge will promote the development of new visual intelligence, which is one of the critical technologies required to promote important breakthroughs in AI 2.0 (Pan, 2016).

Building a visual knowledge dictionary is extremely important. This is a huge practical knowledge platform and data platform, and should be built jointly by the communities of AI, computer graphics, and computer vision researchers. A visual knowledge dictionary can be effectively built using crowdsourcing.

Acknowledgements

I thank Profs. Yue-ting ZHUANG, Fei WU, and Si-liang TANG for their helpful advice.

References

Anderson JR, 1989. Cognitive Psychology (Yang Q, Trans.). Jilin Education Press, China.

Greff K, Kaufmann RL, Kabra R, et al., 2019. Multi-object representation learning with iterative variational inference. https://arxiv.org/abs/1903.00450

Horoufchin H, Bzdok D, Buccino G, et al., 2018. Action and object words are differentially anchored in the sensory motor system—a perspective on cognitive embodiment. *Sci Reports*, 8:6583. https://doi.org/10.1038/s41598-018-24475-z

Hutchinson B, Mitchell M, 2019. 50 years of test (un)fairness: lessons for machine learning. Proc Conf on Fairness, Accountability, and Transparency, p.49-58. https://doi.org/10.1145/3287560.3287600

Kosiorek AR, Sabour S, Teh YW, et al., 2019. Stacked capsule autoencoders. https://arxiv.org/abs/1906.06818

Ma SD, Zhang ZY, 1998. Computer Vision: Fundamentals of the Computational Theory and Algorithms. China Science Publishing & Media Ltd., Beijing, China (in Chinese).

Pan YH, 1991. A study on the imagery information model of imagery thinking. *Patt Recogn Artif Intell*, 4(4):7-12 (in Chinese).

Pan YH, 1996. A study on integrated reasoning. *Patt Recogn Artif Intell*, 9(3):201-208 (in Chinese).

Pan YH, 2016. Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409-413. https://doi.org/10.1016/J.ENG.2016.04.018

Pan YH, Tong RF, Tang M, 2011. Computer Graphics: Principles, Methods and Applications (3rd Ed.). Higher Education Press, Beijing, China (in Chinese).

Zhao Y, Birdal T, Deng H, et al., 2019. 3D point capsule networks. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1009-1018.