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:

Miniaturized five fundamental issues about visual knowledge

Yun-he PAN

Institute of Artificial Intelligence, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

E-mail: panyh@zju.edu.cn

Published online Nov. 18, 2020
https://doi.org/10.1631/FITEE.2040000

1 Fundamental issue 1: expression of visual knowledge

Cognitive psychology has for a long time considered that an important part of human memory is visual knowledge, which is used for conducting concrete thinking. Therefore, visual-based artificial intelligence (AI) is a subject that AI cannot bypass and is of great significance. As a continuation of the previous article "On visual knowledge" (Pan, 2019), in this paper we discuss five fundamental issues about visual knowledge.

Cognitive psychology experiments have been carried out to distinguish the characteristics of visual knowledge from those of language knowledge: (1) Visual knowledge can express the dimensions, color, texture, spatial shape, and spatial relationships of an object; (2) Visual knowledge can convey the movement, speed, and temporal relationships of an object; (3) Visual knowledge can facilitate space-time transformation, manipulation, and reasoning of objects (for example, the transformation of shape, action, velocity, and scene, and various spatio-temporal analogies, associations, and predictions based on spatio-temporal reasoning results).

The expression of structural features realized by computer graphics (CG) is well suited to express the manipulation and transformation of characteristics (1) and (3), but it is challenging to apply CG to simulated imaginative transformation because CG expresses in a geometric form rather than a visual concept.

A visual concept should be composed of a prototype and a domain. For instance, apples have ever-changing forms, but they must have one or several fundamental shapes and colors comprising what is known as a prototype. Within the prototype, a variety of apples constitute a range of transformation which has a boundary. Forms within the boundary belong to the category "apple," beyond which they will become other fruits. This range of transformation is the scope of the concept of "apple." Visual concepts have a hierarchical structure, which is a spatial organization containing sub-concepts. Moreover, visual concepts have an action structure, which should contain the typical motion and action categories of each sub-concept in the structure.

The visual proposition is the expression of the spatial and temporal relationships of a visual concept. The spatial relationship can be regarded as the scene structure, which describes locational relationships (e.g., up and down, left and right, front and rear), distance relationships, internal and external relationships, and size relationships in geometric patterns. The temporal relationship can be regarded as the dynamic structure that conveys timing patterns such as growth, displacement, movement, transformation, competition, synergy, and evolution.

Prof. Yun-he PAN
Editor-in-Chief

Visual narration consists of a group of visual propositions. For example,

in a video, it could constitute the dynamics of different visual objects in various scenes. A silent movie is a case of visual narration.

Visual narration is concrete continuous expression, while language narration is discretely abstract expression. A sign language is a visual expression of language imitation.

Cognitive psychology studies show that human memory stores much more visual knowledge than language knowledge. Visual knowledge has been classified as common sense because it is difficult to express in language symbols. For instance, before the age of five, children who see different cups may grab them in different ways to drink water, which proves that they are proficient in the application of visual knowledge, although they may not yet be able to explain their actions in words (Fig. 1). What people learn in their early childhood is mostly visual knowledge.

Fig. 1 Children aged three or four can recognize and use different cups, but may not be able to describe these objects in language

A major weakness of previous AI research has been the deficiency of visual knowledge. As a result, the research and application of visual knowledge is a significant direction for development in AI 2.0 (Pan, 2016).

2 Fundamental issue 2: visual recognition

Since the initial development period of AI, pattern recognition has been one of the most important fields of research. Image and video recognition is the direction that has realized the most rapid development.

Image recognition technology based on digital image processing technology, is a method that integrates local features into an overall object. Recently, deep learning has provided another method in an end-to-end way: training a deep neural network (DNN) model using a large number of labeled images for image recognition. In this way, recognition accuracy can be significantly improved, and this approach has been widely applied.

The advantages of DNN are: (1) Model knowledge can be learned automatically from the labeled sample data; (2) It can be used for the recognition of non-symbolic data, e.g., image and speech recognition. However, DNN has its drawbacks: (1) Interpretation is difficult; (2) Inference is not feasible; (3) A large amount of labeled data is required to train a network to obtain knowledge.

Note that, unlike the DNN method, during the human visual recognition process in their work memory, people not only analyze the data introduced into their short-term memory after retinal perception, but also activate the relevant mental imagery (i.e., visual knowledge) in their long-term memory needed for the processing of work memory. Therefore, human beings often need only a small amount of data to complete a visual recognition task, and it can be explained and inferred.

Therefore, in visual recognition, not only the use of data, but also the coordinated use of visual knowledge to form a data-driven and visual knowledge-guided computing paradigm has become an important research direction.

3 Fundamental issue 3: simulation of visual imagery thinking

Imagery thinking is a critical intelligent behavior of human beings in design, creativity, and problem solving. Simulating imagery thinking requires the following operations: (1) physical transformation of the visual imagery, such as geometric transformation, spatio-temporal transformation, scene transformation, comparison, prediction, dismantling, and assembly; (2) biological changes of the visual imagery, e.g., movement, generation, and interaction; (3) imaginary transformation of the visual imagery, such as various imaginative operations in the creation and design of new products, e.g., “Journey to the West,” “Avatar,” “The Lion King,” and “Dumbo.”

Simulation of visual imagery thinking has been widely applied in many fields including

computer-aided design (CAD), computer animation, games, children's education, and digital media creativity. According to the different digital media types, these applications can be divided into three categories: (1) Generation of visual imagery from text, for example, with a given text description, to automatically generate a graphic image background, or with a given paragraph of evaluation, to automatically modify the design of a product. (2) Transformation from one visual imagery to another. For instance, in "The Monkey Creates Havoc in Heaven," the Monkey King is automatically transformed into a temple from the imagery of a monkey. In another typical example, researchers at the Institute of Modern Industrial Design, Zhejiang University, changed the standard seal word in a square into a seal word in Chinese-seal style with seal design characteristics (Fig. 2). (3) Generation of text from visual imagery, for example, to assign a title or natural language description to a picture or video and classify it. The automatic generation of text description for short video content has already been applied for accurate online shopping recommendation (Zhang SY et al., 2020).

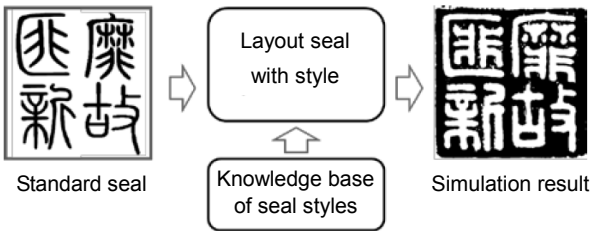

Fig. 2 Imagery simulation AI technology used for an engraving layout

CG has incorporated many fundamental technologies, but still needs to be connected to AI. Once realized, it is expected to form the basis of a new generation of design software.

4 Fundamental issue 4: learning of visual knowledge

In computer vision (CV), the importance of visual object shape reconstruction has been recognized. There have been many achievements including three-dimensional (3D) scanning for shape reconstruction, multi-camera reconstruction of shape, and video-based shape reconstruction. Shape reconstruction bridges the gap between CV and CG.

However, the aim of visual knowledge learning is to upgrade the goal from the task of reconstruction of visual shapes to the reconstruction of visual knowledge concepts and propositions. This requires further research on existing CV technologies, not only to reconstruct 3D shapes, but also to reconstruct the conceptual structure and hierarchical structure of 3D shapes.

This should lead to the development of methods for automatic learning of visual knowledge. Current research on scenario diagrams (Xu et al., 2017; Zellers et al., 2018) is an appropriate intermediate step towards this direction.

To this end, today's researchers in the three fields of AI, CG, and CV need to work together to study the automatic learning of visual knowledge.

5 Fundamental issue 5: multiple expressions of knowledge (Pan, 2020)

Knowledge in the human brain is often described by multiple expressions. Therefore, in AI 2.0, knowledge should be expressed in multiple ways. There are three kinds of knowledge expression and processing methods: (1) Language expression of knowledge. This is characterized by the use of symbolic data. Therefore, its structure is clear, with perceivable semantics and inferable knowledge. Typical examples include semantic networks and knowledge maps (Zhang NY et al., 2020). At present, the acquisition of such knowledge is developing from artificial construction to automatic extraction (Tang et al., 2018). (2) Deep neural network expression of knowledge. This is suitable for the classification and recognition of unstructured data, such as image and audio, but it is difficult to interpret semantics. DNN is a typical example. Currently, the acquisition of such knowledge is developing from supervised to unsupervised learning (Brown et al., 2020). (3) Imagery expression of knowledge. This is suitable for graphics and animation, including data for describing shape, space, and motion. This kind of knowledge has a clear structure, interpretable semantics, and deductive knowledge. A typical example is visual knowledge. The acquisition and use of knowledge of this type

is a direction that requires urgent research and development.

These three kinds of knowledge expression of AI are in line with the three different but interlinked components of human memory, which can be described as follows: (1) knowledge graph—the memory component of semantics, which is applicable to the retrieval and reasoning of character information; (2) visual knowledge—the memory component of visual scenes, which is available for spatio-temporal inference and visualization of visual imagery information; (3) deep neural network—the perception of memory component, which is suitable for learning and classifying patterns in the original data through layer-by-layer abstraction.

Components (1) and (2) correspond to the two major elements of human long-term memory, namely, the encoding mode of language and mental imagery. Component (3) corresponds to the perceptual element of human short-term memory.

AI 2.0 aims to implement intercommunion among multiple knowledge types, i.e., multiple knowledge expression. This will form the technical basis of cross-media intelligence (Zhuang et al., 2017) and big data intelligence.

Our analysis of the five major issues of visual intelligence shows that there are good foundations for addressing issues 1, 2, and 4, while issues 3 and 5 still need to be tackled by scholars from various fields. Therefore, visual knowledge and cross-media knowledge expression are the key areas.

6 Conclusions

Our analysis shows that the distinct advantages of visual knowledge are its capacity to generate comprehensive imagery, its spatio-temporal evolution capacity and imagery display capacity. These are the features currently lacking in character knowledge and DNN. Integration of AI and CAD/CG/CV technologies will provide a vital foundation for the development of AI in terms of creation, prediction, and man–machine integration.

The study of visual knowledge and multiple expressions of knowledge is the key to the development of visual intelligence and the main theory and technology to enable AI 2.0 to make major breakthroughs.

It is a desolate, clammy, and fertile "Great Northern Wilderness," but also a "depopulated land" full of hope and worthy of multi-disciplinary cooperation.

Acknowledgements

Thanks to Profs. Ling-yun SUN, Jun XIAO, Ke-jun ZHANG, and Mr. Huang-huang DENG for providing valuable information for this paper.

References

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. https://arxiv.org/abs/2005.14165

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, 2019. On visual knowledge. *Front Inform Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan YH, 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Tang SL, Zhang Q, Zheng TP, et al., 2018. Two step joint model for drug drug interaction extraction. https://arxiv.org/abs/2008.12704

Xu DF, Zhu YK, Choy CB, et al., 2017. Scene graph generation by iterative message passing. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5410-5419.

Zellers R, Yatskar M, Thomson S, et al., 2018. Neural motifs: scene graph parsing with global context. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5831-5840.

Zhang NY, Deng SM, Zhang W, et al., 2020. Relation adversarial network for low resource knowledge graph completion. Proc Web Conf, p.1-12. https://doi.org/10.1145/3366423.3380089

Zhang SY, Tan ZQ, Zhou Z, et al., 2020. Comprehensive information integration modeling framework for video titling. Proc SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.2744-2754. https://doi.org/10.1145/3394486.3403325

Zhuang YT, Jain R, Gao W, et al., 2017. Panel: cross-media intelligence. Proc 25th ACM Int Conf on Multimedia, p.1173. https://doi.org/10.1145/3123266.3133336