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

Review:

Towards human-like and transhuman perception in AI 2.0: a review*

Yong-hong TIAN†1, Xi-lin CHEN2, Hong-kai XIONG3, Hong-liang LI4, Li-rong DAI5, Jing CHEN1, Jun-liang XING6, Jing CHEN7, Xi-hong WU1, Wei-min HU6, Yu HU5, Tie-jun HUANG†‡1, Wen GAO1

(1School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)
(2Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)
(3Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)
(4School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611730, China)
(5Department of Electronic Engineering and Information Sciences, University of Science and Technology of China, Hefei 230027, China)
(6Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)
(7School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China)
†E-mail: yhtian@pku.edu.cn; tjhuang@pku.edu.cn

Received Dec. 12, 2016; Revision accepted Dec. 26, 2016; Crosschecked Dec. 26, 2016

Abstract: Perception is the interaction interface between an intelligent system and the real world. Without sophisticated and flexible perceptual capabilities, it is impossible to create advanced artificial intelligence (AI) systems. For the next-generation AI, called 'AI 2.0', one of the most significant features will be that AI is empowered with intelligent perceptual capabilities, which can simulate human brain's mechanisms and are likely to surpass human brain in terms of performance. In this paper, we briefly review the state-of-the-art advances across different areas of perception, including visual perception, auditory perception, speech perception, and perceptual information processing and learning engines. On this basis, we envision several R&D trends in intelligent perception for the forthcoming era of AI 2.0, including: (1) human-like and transhuman active vision; (2) auditory perception and computation in an actual auditory setting; (3) speech perception and computation in a natural interaction setting; (4) autonomous learning of perceptual information; (5) large-scale perceptual information processing and learning platforms; and (6) urban omnidirectional intelligent perception and reasoning engines. We believe these research directions should be highlighted in the future plans for AI 2.0.

Key words: Intelligent perception; Active vision; Auditory perception; Speech perception; Autonomous learning
http://dx.doi.org/10.1631/FITEE.1601804 **CLC number:** TP391

1 Introduction

The source of biological intelligence is the perception of external stimuli. For example, the human brain perceives the outside world in real time through more than three million nerve fibers (more than one million fibers per eye). Similarly, perception is the interaction interface between an intelligent system and the real world. Without sophisticated and flexible perceptual capabilities, it is impossible to create advanced artificial intelligence (AI) systems. Just like a person has visual, auditory, taste, and other different sensory systems (Bear *et al.*, 2001), perception in an AI system typically begins with (possibly distributed) sensor data in various modalities and forms. The

‡ Corresponding author

* Project supported by the Strategic Consulting Research Project of Chinese Academy of Engineering (No. 2016-ZD-04-03)

ORCID: Yong-hong TIAN, http://orcid.org/0000-0002-2978-5935

sensor data is processed and synthesized, often along with prior knowledge and models, to extract information such as geometric features, attributes, location, and velocity that is relevant to the task of the AI system. Therefore, integrated data from perception forms situational awareness that provides AI systems with comprehensive knowledge and models about the state of the world necessary to understand, plan, and execute tasks effectively and safely.

For an animal from any species to exhibit intelligent perception, it must be capable of being consciously aware of what it perceives and capable of learning from this experience (Kendrick, 1998). That is, intelligent perception is the ability to both be aware and learn from it. In the AI community, researchers have struggled for several decades with the challenge of designing and implementing intelligent perception systems that can effectively simulate the brain's mechanisms. Great success has been achieved with some specific problems and tasks such as face recognition in a constrained environment, especially given the recent advances in deep learning. However, these systems are still far from where they should be. For example, one of the main problems is that we often need to develop different computational algorithms or tools for different perceptual tasks, while ignoring the correlation or dependency between these tasks. According to Mountcastle (1978), the cortex does something universal that can be applied to any type of sensory or motor system. Essentially, the brain uses the same process to see as to hear, to touch, to motion, etc. More importantly, future intelligent perception systems should not only simulate the brain effectively in terms of mechanism (referred to as 'human-like perception'), but also surpass the human brain in terms of performance (referred to as 'trans-human perception'). This is one of the most significant features of the next-generation AI, called 'AI 2.0' by Pan (2016). Such a system is recognized as the next-generation, general-purpose AI, which will be beyond the theoretical capabilities and limitations of current AI.

The main purpose of this article is to envision several R&D trends in intelligent perception in the forthcoming era of AI 2.0. Towards this end, we briefly review the state-of-the-art advances in different areas of perception, including visual perception, auditory perception, speech perception, perceptual information processing, and learning engines. We believe that these research directions should be highlighted in the future plans for AI 2.0.

2 State of the art

This section will briefly review the state-of-the-art across different areas of perception. After more than 30 years of continued efforts, many ideas have materialized into numerous transformative perception technologies. Due to space limitations, the following subsections will highlight only the recent progress in some of these areas.

2.1 Visual signal acquisition

In the field of signal acquisition, the Nyquist sampling theorem is a fundamental bridge between continuous-time signals (often called 'analog signals') and discrete-time signals (often called 'digital signals'). Candès *et al.* (2006) proposed the theory of compressed sensing as a signal processing technique for efficiently acquiring and reconstructing a signal. This theory moves beyond certain limitations in the Nyquist sampling theorem, by assuming that the amount of signal to be sampled does not depend on the signal bandwidth but on its internal structure. If the signal is sparse in the original or a transformation domain, it can be projected onto the low-dimensional space using a measurement matrix that is dependent on the transform and that satisfies the restricted isometry property. Technologically speaking, compressed sensing provides a new way to consider the relationship between the information and the signal.

Basically, compressed sensing can be used in photography to reduce hardware complexity, increase the imaging frame rate, and improve image reconstruction. This technology, called 'compressive imaging', provides a new approach incorporating more intelligent image acquisition, by transitioning from traditional imaging to information imaging. As a pioneer work, a single pixel camera was developed at Rice University (Duarte *et al.*, 2008), which could obtain an image or video with a single detection element (the 'single pixel') while measuring the scene fewer times than the number of pixels/voxels. Following that, a high-resolution, short-waved infrared compressed sensing camera was developed in

McMackin *et al.* (2012), while a multi-view lenseless camera was developed in Jiang *et al.* (2014), in which aberration and focusing problems associated with lenses could be completely avoided. In Kadambi *et al.* (2013) and Tadano *et al.* (2015), the authors aimed to develop a unified theory and practical designs for adaptive coded imaging and display, which could adapt to scene geometry, motion, and illumination to maximize the information throughput. Obviously, new computational imaging theory and technologies such as ultra-high speed imaging and light-field imaging are highly desirable due to some recent large technological demands such as automated driving and virtual reality.

2.2 Active vision—from looking to looking around

In the past decades, computer vision has played an important role in AI. We have witnessed the dramatic change in various recognition tasks. After about 30-year work on general object recognition and 3D recovery, researchers in the 1990s began focusing their attention on recognition of special objects such as faces, pedestrians, and vehicles (Turk and Pentland, 1991). Now, computer vision systems are superior to human beings in large-scale face recognition tasks under controlled environments. Meanwhile, object categorization has been extended from several classes to thousands of classes (Deng *et al.*, 2009). As one of the deep learning models, the convolutional neural network (CNN) and its derivatives have achieved great success in ImageNet and in other tasks (Krizhevsk *et al.*, 2012). This achievement seemed to break the barrier to automatic recognition. However, this is misleading. Human beings, and even other animals, never look at something from a passive acquisition of information point of view. They look around under various conditions, sometimes even touching and using other sensory organs.

Therefore, active sensing and recognition will be the terminator in computer vision. In contrast to an animal's active vision which can move only on six degrees of freedom (DoFs), an active computer vision system can have more DoFs. Some efforts in this regard have made significant progress in recent years. One example is Microsoft's Kinect. With an active ejected infrared pattern, 3D reconstruction has become easier than ever before. This kind of device takes image/video based human-computer interface (HCI) and other indoor applications a step forward (Han *et al.*, 2013). Another example is city-scale reconstruction from multiple uncalibrated cameras (Musialski *et al.*, 2013). Although these images are captured passively, the whole set can still be viewed as an active set. Self-driving cars, drones, and other mobile robots bring computer vision systems to a full capability of moving around in their environment. This mobility enables the computer vision system to observe an environment actively and continuously. Note that in the DARPA Robotics Challenge competition, continuous vision systems have shown their power preliminarily (Pratt and Manzo, 2013). Therefore, seeking methods to implement more comprehensive active sensing and recognition systems should be one of the most important tasks in the AI 2.0 era.

2.3 Auditory perception and computation

Auditory perception is a central pathway for information interaction in human beings. It usually occurs within a complicated auditory setting, which includes multiple sound sources and reverberation. However, machine auditory perception shows a significantly lower performance in actual environments. To reduce the detrimental effects caused by competing sound sources, signal processing algorithms for speech enhancement based on the input of a single microphone were studied in the 1960s, and they have achieved a good performance for near-field recordings but were useless for far-field recordings (Robinson and Treitel, 1967). In the 1970s, algorithms based on microphone arrays were studied to detect and enhance target sounds in far-field situations, but they usually worked well only for high signal-to-noise ratio (SNR) scenarios and, critically, they required identical microphones for recording (Roy and Kailath, 1989).

With the recent development in CNNs, some algorithms have helped enhance target sound sources in a reverberant environment, and the limitation of identical microphones was partially relaxed (Niwa *et al.*, 2016). However, they were still not efficient for settings with multiple sound sources. The auditory mechanism of binaural processing has revealed that the physical structure of the human ear and body is important for sound localization, and this information is conveyed by the human head related transfer function (HRTF). A recent work using CNN to demodulate this function showed a promising result for multiple sound source settings and low SNR (Song *et al.*, 2016). Nevertheless, the question of how to develop effective auditory perception and computation algorithms in complicated auditory settings remains open, and it should be addressed in future research in the era of AI 2.0.

2.4 Speech perception and computing

Speech perception and computing are together one of the core technologies to achieve man-machine interaction in the AI 2.0 era. Typically, speech recognition and speech synthesis are the two main tasks of speech perception and computing. The aim of speech recognition is to convert spoken language into text using automatic algorithms. Current speech recognition systems usually adopt an acoustic model and a language model to represent the statistical properties of speech. Since 2009, deep learning techniques have been applied to the acoustic modeling of speech recognition and they have achieved great success (Hinton *et al.*, 2012). The word error rates (WERs) in speech recognition systems are significantly reduced compared to conventional hidden Markov model (HMM) based acoustic modeling. The WERs of several representative systems using deep learning techniques are listed in Table 1, where Switchboard is a standard large-vocabulary conversational speech recognition task. The latest progress reported by the Microsoft speech team devoted to this task is that it has reached human parity (Xiong *et al.*, 2016).

Table 1 Performance of the speech recognition systems using deep learning techniques on the Switchboard part of the Hub5-2000 evaluation test set*

System	Amount of training data	WER (%)
Seide *et al.* (2011)	309 h	16.1
Veselý *et al.* (2013)	309 h	12.6
Soltau *et al.* (2014)	309 h	10.4
IBM's system (Saon *et al.*, 2015)	309 h	9.6
	2000 h	8.0
Microsoft's system (Xiong *et al.*, 2016)	2000 h	5.9

* A comparable HMM-based system achieved a WER of 23.6% (Seide *et al.*, 2011). WER: word error rate

The aim of speech synthesis is to generate intelligible and natural-sounding artificial speech for input text. A typical speech synthesis system is composed of two main modules, text analysis and speech waveform generation. Statistical parametric speech synthesis and unit selection are two mainstream approaches to speech waveform generation nowadays (Tokuda *et al.*, 2013). Statistical models, such as HMMs or CNNs, play an essential role in both approaches (Ling *et al.*, 2015). Current speech synthesis systems are able to produce reading-style utterances with high intelligibility and naturalness when enough training data is available and appropriate algorithms are applied (King, 2014).

At present, there are still many challenges in the technological progress and industrial development of intelligent speech perception and computing. In the field of speech recognition, current techniques still have limitations. First, most of the existing methods are language or dialect dependent, and the self-learning capability of intelligent speech perception is very limited (Makhoul, 2016). Second, the problem of distant and noise-robust speech recognition has still not been solved well (Amodei *et al.*, 2015). In the field of speech synthesis, there are still gaps between the naturalness of synthetic speech and the human voice. Moreover, the performance of synthesized speech uttered with high expressiveness like a human being remains unsatisfactory.

2.5 Machine learning for perceptual information

Although deep learning has achieved great success in many tasks, we cannot explain exactly why such networks are effective theoretically. Recently, studies have emerged concerned with the mathematical theory of deep models. Mahendran and Vedaldi (2015) tried to understand deep representations by inverting them, reconstructing both the shallow and deep features to investigate the connections to the original signal. Bruna and Mallat (2013) proposed a scattering network based on cascades of wavelet filters and average operations. More details and mathematical demonstrations have been discussed to interpret scattering networks. These studies proved that the combination of signal processing tools and machine learning methods helps build the theoretical basis of deep learning.

Traditional deep learning models often rely on labeled data (Krizhevsk *et al.*, 2012), which is very difficult and expensive to obtain, and thus the ability to use unlabeled data holds a significant promise. The ability to learn with unlabeled data could be treated as autonomous learning, since machines are not told what to learn. On the other hand, unsupervised learning had a catalytic effect in reviving interest in deep learning, but that has been overshadowed by the successes of purely supervised learning. Raina *et al.* (2007) described an autonomous learning approach to self-teach the models via sparse coding to construct higher-level features using the unlabeled data. Since unlabeled data is significantly easier to obtain than the typical supervised learning data, autonomous learning is widely applicable to many practical learning problems (Fig. 1).

Moreover, traditional neural networks lack the memory for dynamic information input streams (such as video streams and voice streams) with strong correlations. To address this issue, Hochreiter and Schmidhuber (1997) proposed a long short-term memory (LSTM) neural network, which can process sequence data. LSTM units can extract and store some previous correlative information.

Note that deep networks have been successfully applied to learn models from a single modality, while in the real world, the information comes from different sources. For example, audio and visual data coming from the same video has correlations at a ‘mid-level’. It is better for the networks to learn features over multiple modalities. Ngiam *et al.* (2011) presented a multimodal learning framework and demonstrated that cross-modality feature learning is able to obtain better features compared with single-modality feature learning. Deep Boltzmann machines (Salakhutdinov and Hinton, 2009) are usually used to find a unified representation of different modes based on the probability density of a multimodal input.

In summary, the development of more effective autonomous learning models and algorithms for various types of perceptual information and data should be one of the central tasks for AI 2.0.

2.6 Large-scale processing and learning platform

To achieve high performance, deep learning algorithms often require incredible amounts of data and computational power to train a recognition or classification model, and in this regard, hardware acceleration is highly desirable. Clusters of graphics processing units (GPUs) are the most popular solution and have been widely used in many open-source deep learning platforms, such as Google TensorFlow, Microsoft DMTK, Samsung Veles, Baidu Paddle, and DMLC MXNet. As reported in the NVIDIA DGX-1 Deep Learning System, the training speed of a GPU

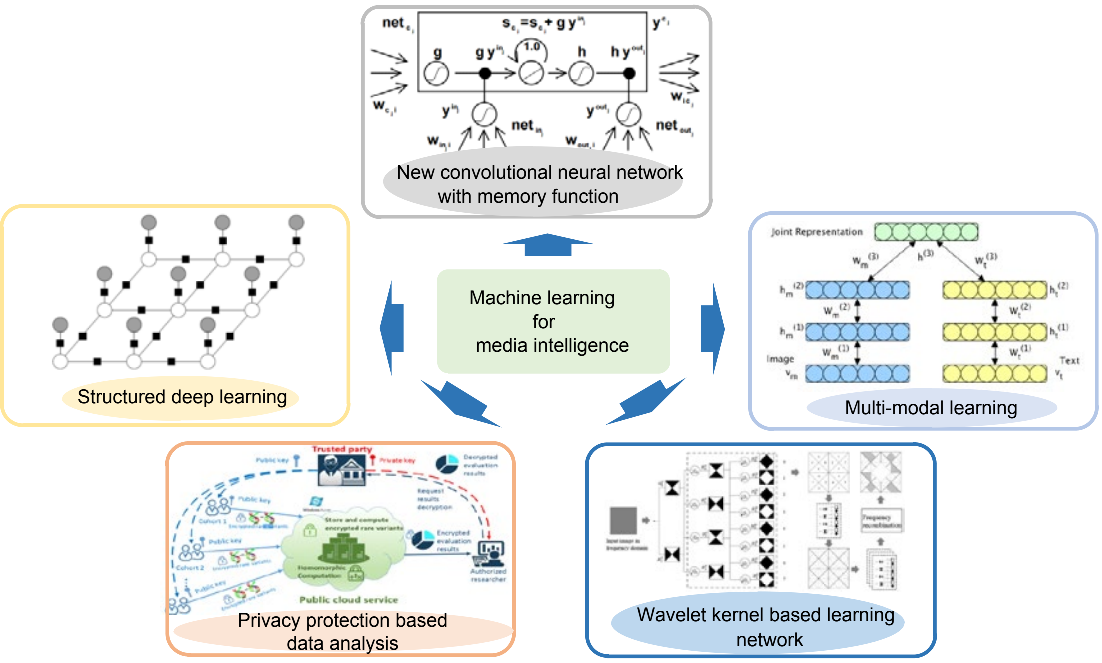
New convolutional neural network with memory function
Structured deep learning
Machine learning for media intelligence
Multi-modal learning
Joint Representation
Image
Text
Trusted party
Public cloud service
Privacy protection based data analysis
Wavelet kernel based learning network

Fig. 1 The potential autonomous learning directions for intelligent perception

could be 75 times faster than that of a CPU. In addition, studies on deep-learning-friendly field programmable gate arrays (FPGAs) (Lacey *et al.*, 2016) are also active, focusing on data parallelism, model parallelism, and pipeline parallelism. Except for computing units, some neural-network-like sensors are also designed to speed up the data processing of deep learning.

Another key for large-scale applications in machine learning is the use of commercial machine learning clouds. On the cloud, data scientists no longer need to manage the infrastructure or implement their codes; instead, the cloud system automatically performs this work for them and generates a new model in real time, which is faster and will provide more accurate results. To date, Microsoft, Google, HP, and IBM have released their own machine learning clouds. Chinese corporations and researchers need to build their own commercial machine learning clouds to boost the application of large-scale machine learning. Towards this end, the ability of these machine learning cloud systems to cope with the huge amounts of perception data needs to be extensively verified.

2.7 A typical application paradigm: urban intelligent surveillance system

Public security is a growing problem for cities worldwide. A smart city should first be a safe city. To this end, urban surveillance is becoming increasingly important in a modernized safe city. With the fast development and deployment of all kinds of digitalized devices in every aspect of people's daily life, intelligent surveillance systems and the Internet of Things are drawing extensive attention from both research and industrial communities. Governments from the US, Canada, the EU, Japan, and China have all launched a series of related projects to improve the social and public security. By 2014, there were more than 300 cities around the world that aimed to build an intelligent city (Hou and Jiao, 2014). Global transnational corporations like IBM, Cisco, Siemens, Huawei, and Hikvision have invested enormously in the study and development of related solutions and products for intelligent surveillance systems.

From a technological perspective, an urban surveillance system usually involves a set of independent or weakly related sub-systems such as traffic monitoring systems (Zhang *et al.*, 2011), crowd analysis systems (Li *et al.*, 2015), criminal tracking systems (Zheng *et al.*, 2016), and property protection systems (Kale and Sharma, 2014). These sub-systems focus mainly on multi-source heterogeneous information processing, e.g., exploiting potential information behind surveillance video data to organize it into a structured video surveillance repository. However, they always encounter different sorts of bottlenecks in handling the scenarios of spatio-temporal large-span sensing, multi-layer multi-view analysis, and multi-source heterogeneous information fusion from a panoramic view. Therefore, there is an urgent need to promote an accelerated development of urban surveillance systems in order to reach a new fully intelligent surveillance sensing and reasoning engine.

3 R&D trends

Despite many research efforts devoted to intelligent perception in the past several decades, further progress is still needed in the development of more advanced theories, algorithms, and technologies that can effectively determine what an AI system can perceive and predict about the future states of the world. Essentially, we can envision that future intelligent perception systems should not only simulate the brain's mechanisms effectively, but also surpass the human brain in terms of performance.

Towards this end, we suggest the following two-step R&D strategy:

Short-term goal: To achieve intelligent perception methods and technologies that can successfully generate a uniform semantic representation of objects, scenes, behavior, and events in the real world, realize audio analysis and speech recognition in a natural auditory setting, and develop new machine learning algorithms and methods for large-scale perception data.

Long-term goal: To establish human-like, and even transhuman, intelligent perception theories, methods, and technologies. These will include active perception and learning models, human-like auditory perception and understanding technologies in actual auditory settings, and autonomous, self-evolving, and collaborative learning theories and models on intelligent perception.

Moreover, these intelligent perception methods and technologies should be applied to some important applications, such as urban surveillance systems, to significantly improve their intelligent services.

Fig. 2 provides a graphical illustration of the technological framework of intelligent perception in AI 2.0. The core of the framework is to derive a uniform semantic representation of the real world through next-generation intelligent perception technologies, including active vision, auditory perception, speech perception and computation, and autonomous learning. These technologies take the roles like the human eyes, ears, and mouth, as well as their corresponding neural information processing systems. Meanwhile, they may work on large-scale perceptual information processing and learning platforms (i.e., iMedia), and then can be applied to the urban omnidirectional intelligent perception and reasoning engine (i.e., iEye). Here, iMedia acts as the computational engine or infrastructure of intelligent perception in AI 2.0, while iEye is a comprehensive system that can apply these new-generation intelligent perception methods and technologies in a smart city. In the following subsections we describe our vision from these aspects.

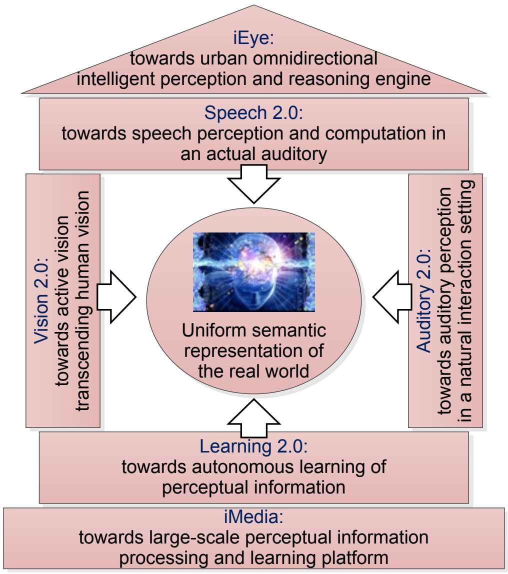
iEye:
towards urban omnidirectional
intelligent perception and reasoning engine
Speech 2.0:
towards speech perception and computation in
an actual auditory
Vision 2.0:
towards active vision
transcending human vision
Uniform semantic
representation of
the real world
Auditory 2.0:
towards auditory perception
in a natural interaction setting
Learning 2.0:
towards autonomous learning of
perceptual information
iMedia:
towards large-scale perceptual information
processing and learning platform

Fig. 2 A vision about the technological framework of intelligent perception in AI 2.0

3.1 Vision 2.0: towards human-like and trans-human active vision

In the coming decade, most robots, including self-driving cars and drones, will be equipped with various computer vision systems. The requirements for these computer vision systems are significantly different from those cases in the past decades. The system should perform like an ordinary person, able to deal with most daily visual tasks easily, rather than as a specialized expert, capable of dealing with only specific tasks. For this purpose, we need to have a new generation of active vision systems. These systems need to have the capability of understanding the environment and recognizing thousands of objects in almost real time. In some cases, these systems are expected to be superior to the human vision system. To achieve this goal, efforts should focus on the following areas:

1. On-site and active learning for vision tasks

In contrast to those offline learned tasks, the active vision system needs to have the capability of recognizing zero-shot objects on-site. For this purpose, it needs to capture and model objects interactively with other components. An active loop of capturing and modeling ego motion will be a key issue for on-site learning.

2. Beyond human sensing

Human beings capture visual information with two eyes, and across a limited spectrum. This leads to a very complex procedure to recover 3D in later stages. With the progress in computing power, memory size, and sensing, it is possible to have next-generation cameras which can record full information from the environment. This will cause vision systems to be superior to human sensing in many aspects.

3.2 Auditory 2.0: towards auditory perception and computation in an actual auditory setting

To date, the performance of machines is still far below that of human beings in natural auditory settings, especially in understanding audio in reverberant environments and noisy backgrounds (Lippmann, 1997). To resolve this problem, it is necessary to study the mechanism of auditory binaural processing, as the binaural advantage is prominent in natural settings with multiple sound sources. Related studies include psychoacoustic models for sound localization and for interpreting the precedence effect (Litovsky *et al.*, 1999), adaptive learning models based on HRTF, computational models for multi-scale harmonic analysis, demodulation methods for HRTF, and computational models for sound localization in reverberant environments. Additionally, as speech signals inherently contain rich information and perceptual cues, it is necessary to study the mechanisms underlying the perception of speech with interfering sounds (Mattys *et al.*, 2012). Eventually, novel algorithms and auditory computational models will be produced to improve the performance of machines in sound localization and audio understanding in complicated auditory environments.

3.3 Speech 2.0: towards speech perception and computation in a natural interaction setting

The ultimate purpose for Speech 2.0 is to develop perception and computation for a natural interaction setting. Towards this end, future research directions for speech recognition may include exploring brain-like models and learning algorithms by integrating the mechanism of speech perception and selective attention, and developing novel end-to-end speech recognition frameworks with auditory context perception and adaptation.

For speech synthesis, one future direction may be to develop advanced deep generative models for speech generation, such as modeling speech waveforms directly (Oord *et al.*, 2016) and constructing a unified model for end-to-end speech synthesis (Wang *et al.*, 2016). Another direction may be to explore approaches to extract rich information from texts to boost expressive speech synthesis, such as emotion classification, semantic understanding, and paraphrase-level text analysis.

3.4 Learning 2.0: towards autonomous learning of perceptual information

Current deep learning models are composed of multiple processing layers to learn the representations of data with multiple levels of abstraction, while ignoring a crucial point: the structure of the data. In fact, structured prediction methods have been widely studied in the traditional signal processing field. Such methods use graph models like conditional random fields to construct a structured model to represent and predict the latent knowledge and correlations of multiple output data. Following this idea, we need to establish such a set of intelligent perceptual information processing and learning frameworks with sufficient theoretical support and autonomous learning capabilities so that the deep network is no longer trapped in a hyper-parameter selection framework.

In fact, autonomous learning has already been applied to deep learning, but mostly to show the advantage of unlabeled examples and it is far from achieving satisfactory performance. A promising direction for autonomous learning is to analyze the properties of a signal itself, and to try to reconstruct signals in terms of structured sparsity and topology properties. Furthermore, autonomous learning is able to determine not only what to learn, but also where to learn it (LeCun *et al.*, 2015). Since human vision is an active process that sequentially samples the optic array in an intelligent, task-specific way, we hope future studies on autonomous learning focus on deciding where to learn. Moreover, the memory management mechanism of the human brain provides a quite appealing property in its ability to rank the priorities of our prior knowledge, which is important for refining and reusing human-rated data to efficiently train a learning model. Overall, autonomous learning systems are in their infancy, but they are appealing because they represent true AI.

3.5 iMedia: towards large-scale perceptual information processing and learning platforms

To date, the research in large-scale perceptual information processing and learning focuses mainly on two areas: (1) cognition and perception inspired learning frameworks that are more suitable for exploring the relevance of massive data; (2) highly efficient and low-power hardware that supports deep learning on mobile and portable devices. With respect to the first area, the collaborative computing model over multiple datasets presents great potentials for improving the training efficiency of a learning algorithm by reducing the redundancy of big data. For the second area, the processors that support highly parallel floating point arithmetic would facilitate the development of learning-based applications. For example, chips that are specifically designed for convolution and probability computations are highly desired. The construction of distributed parallel computing systems (DPCSs) is also very important for integrating existing computing resources and facilitating big data processing. More powerful DPCSs would lower the barriers to entry for big data related businesses and services. Meanwhile, low-power system-on-chip (SoC) technology will promote the popularity of deep learning in many consumer electronics, such as smart phones and tablet computers.

3.6 iEye: towards urban omnidirectional intelligent perception and reasoning engines

To deal with the challenges in current urban perception systems, such as information fragmentation and islanding problems (Suzuki, 2015; Priano *et al.*, 2016), the trend is towards building a multi-dimensional intelligent perception and reasoning engine, which is called 'iEye' in this article. Specifically, based on the collected massive image and video data in an urban scale, and through associative analyzing and synthesis reasoning, the iEye system is expected to have features that include big capacity, large view-angles, big data, and excellent service.

The core technologies behind the iEye system include intelligent perception within the scope of a whole city, associative analysis among multiple targets, cross spatial-temporal behavioral understanding, synthesis of heterogeneous information from multiple sources, and urban panorama modeling. With algorithmic innovations in these core technologies, the iEye system will open up new service models for smart cities.

4 Conclusions

AI has great potential to help address some of the biggest challenges that society faces. Towards this end, AI systems would greatly benefit from advancements in theory, algorithms, and hardware to enable more robust, reliable, and intelligent perception. In this article, we envision several R&D trends in intelligent perception in the forthcoming era of AI 2.0. Actually, it is a summary of the specialists' opinions from a subcommittee on intelligent perception technologies, supported by the research project on the National Artificial Intelligence 2.0 Research and Development Strategy from the Chinese Academy of Engineering.

Note that our opinions are also very close to those in a recent strategic plan (https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf) released by the subcommittee on Networking and Information Technology Research and Development (NITRD), under the National Science and Technology Council (NSTC), USA. In this strategic plan, enhancing the perceptual capabilities of AI systems is highlighted as one of the important areas for long-term investments. Therefore, we believe that the research directions listed in this article should be highlighted in AI 2.0.

References

Amodei, D., Anubhai, R., Battenberg, E., *et al.*, 2015. Deep Speech 2: end-to-end speech recognition in English and Mandarin. arXiv:1512.02595.

Bear, M.F., Connors, B.W., Paradiso, M.A., 2001. Neuroscience. Lippincott Williams and Wilkins, Maryland, p.208.

Bruna, J., Mallat, S., 2013. Invariant scattering convolution networks. *IEEE Trans. Patt. Anal. Mach. Intell.*, **35**(8): 1872-1886. http://dx.doi.org/10.1109/TPAMI.2012.230

Candès, E., Romberg, J., Tao, T., 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, **52**(2):489-509. http://dx.doi.org/10.1109/TIT.2005.862083

Deng, J., Dong, W., Socher, R., *et al.*, 2009. ImageNet: a large-scale hierarchical image database. IEEE Conf. on Computer Vision and Pattern Recognition, p.248-255. http://dx.doi.org/10.1109/CVPR.2009.5206848

Duarte, M., Davenport, M., Takhar, D., *et al.*, 2008. Single-pixel imaging via compressive sampling. *IEEE Signal Proc. Mag.*, **25**(2):83-91. http://dx.doi.org/10.1109/MSP.2007.914730

Han, J., Shao, L., Xu, D., *et al.*, 2013. Enhanced computer vision with Microsoft Kinect sensor: a review. *IEEE Trans. Cybern.*, **43**(5):1318-1334. http://dx.doi.org/10.1109/TCYB.2013.2265378

Hinton, G., Deng, L., Yu, D., *et al.*, 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc. Mag.*, **29**(6):82-97. http://dx.doi.org/10.1109/MSP.2012.2205597

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neur. Comput.*, **9**(8):1735-1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

Hou, Y.Z., Jiao, L.F., 2014. Survey of smart city construction study from home and abroad. *Ind. Sci. Trib.*, **13**(24):94-97 (in Chinese).

Jiang, H., Huang, G., Wilford, P., 2014. Multi-view in lensless compressive imaging. *Apsipa Trans. Signal Inform. Proc.*, **3**(15):1-10. http://dx.doi.org/10.1109/PCS.2013.6737678

Kadambi, A., Whyte, R., Bhandari, A., *et al.*, 2013. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Trans. Graph.*, **32**(6):1-10. http://dx.doi.org/10.1145/2508363.2508428

Kale, P.V., Sharma, S.D., 2014. A review of securing home using video surveillance. *Int. J. Sci. Res.*, **3**(5):1150-1154.

Kendrick, K.M., 1998. Intelligent perception. *Appl. Animal Behav. Sci.*, **57**(3-4):213-231. http://dx.doi.org/10.1016/S0168-1591(98)00098-7

King, S., 2014. Measuring a decade of progress in text-to-speech. *Loquens*, **1**(1):e006. http://dx.doi.org/10.3989/loquens.2014.006

Krizhevsk, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, p.1097-1105.

Lacey, G., Taylor, G.W., Areibi, S., 2016. Deep learning on FPGAs: past, present, and future. arXiv:1602.04283.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature*, **521**(7553):436-444. http://dx.doi.org/10.1038/nature14539

Li, T., Chang, H., Wang, M., *et al.*, 2015. Crowded scene analysis: a survey. *IEEE Trans. Circ. Syst. Video Technol.*, **25**(3):367-386. http://dx.doi.org/10.1109/TCSVT.2014.2358029

Ling, Z.H., Kang, S.Y., Zen, H., *et al.*, 2015. Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends. *IEEE Signal Proc. Mag.*, **32**(3):35-52. http://dx.doi.org/10.1109/MSP.2014.2359987

Lippmann, R.P., 1997. Speech recognition by machines and humans. *Speech Commun.*, **22**(1):1-15. http://dx.doi.org/10.1016/S0167-6393(97)00021-6

Litovsky, R.Y., Colburn, H.S., Yost, W.A., *et al.*, 1999. The precedence effect. *J. Acoust. Soc. Am.*, **106**:1633-1654. http://dx.doi.org/10.1121/1.427914

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. IEEE Int. Conf. on Computer Vision Pattern Recognition, p.5188-5196. http://dx.doi.org/10.1109/CVPR.2015.7299155

Makhoul, J., 2016. A 50-year retrospective on speech and language processing. Int. Conf. on Interspeech, p.1.

Mattys, S.L., Davis, M.H., Bradlow, A.R., *et al.*, 2012. Speech recognition in adverse conditions: a review. *Lang. Cogn. Proc.*, **27**:953-978. http://dx.doi.org/10.1080/01690965.2012.705006

McMackin, L., Herman, M.A., Chatterjee, B., *et al.*, 2012. A high-resolution SWIR camera via compressed sensing. *SPIE*, **8353**:835303. http://dx.doi.org/10.1117/12.920050

Mountcastle, V., 1978. An organizing principle for cerebral function: the unit model and the distributed system. *In*: Edelman, G.M., Mountcastle, V.B. (Eds.), The Mindful Brain. MIT Press, Cambridge.

Musialski, P., Wonka, P., Aliaga, D.G., *et al.*, 2013. A survey of urban reconstruction. *Comput. Graph. Forum*, **32**(6): 146-177. http://dx.doi.org/10.1111/cgf.12077

Ngiam, J., Khosla, A., Kim, M., *et al.*, 2011. Multimodal deep learning. 28th In. Conf. on Machine Learning, p.689-696.

Niwa, K., Koizumi, Y., Kawase, T., *et al.*, 2016. Pinpoint extraction of distant sound source based on DNN mapping from multiple beamforming outputs to prior SNR. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.435-439. http://dx.doi.org/0.1109/ICASSP.2016.7471712

Oord, A., Dieleman, S., Zen, H., *et al.*, 2016. WaveNet: a generative model for raw audio. arXiv:1609.03499.

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

Pratt, G., Manzo, J., 2013. The DARPA robotics challenge. *IEEE Robot. Autom. Mag.*, **20**(2):10-12. http://dx.doi.org/10.1109/MRA.2013.2255424

Priano, F.H., Armas, R.L., Guerra, C.F., 2016. A model for the smart development of island territories. Int. Conf. on Digital Government Research, p.465-474. http://dx.doi.org/10.1145/2912160.2912187

Raina, R., Battle, A., Lee, H., *et al.*, 2007. Self-taught learning: transfer learning from unlabeled data. 24th Int. Conf. on Machine Learning, p.759-766. http://dx.doi.org/10.1145/1273496.1273592

Robinson, E.A., Treitel, S., 1967. Principles of digital Wiener filtering. *Geophys. Prospect.*, **15**(3):311-332. http://dx.doi.org/10.1111/j.1365-2478.1967.tb01793.x

Roy, R., Kailath, T., 1989. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process.*, **37**(7):984-995. http://dx.doi.org/10.1109/29.32276

Salakhutdinov, R., Hinton, G., 2009. Deep Boltzmann machines. *J. Mach. Learn. Res.*, **5**:448-455.

Saon, G., Kuo, H.K.J., Rennie, S., *et al.*, 2015. The IBM 2015 English conversational telephone speech recognition system. arXiv:1505.05899.

Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks. Int. Conf. on Interspeech, p.437-440.

Soltau, H., Saon, G., Sainath, T.N., 2014. Joint training of convolutional and nonconvolutional neural networks. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.5572-5576. http://dx.doi.org/10.1109/ICASSP.2014.6854669

Song, T., Chen, J., Zhang, D.B., *et al.*, 2016. A sound source localization algorithm using microphone array with rigid body. Int. Congress on Acoustics, p.1-8.

Suzuki, L.R., 2015. Data as Infrastructure for Smart Cities. PhD Thesis, University College London, London, UK.

Tadano, R., Pediredla, A., Veeraraghavan, A., 2015. Depth selective camera: a direct, on-chip, programmable technique for depth selectivity in photography. Int. Conf. on Computer Vision, p.3595-3603. http://dx.doi.org/10.1109/ICCV.2015.410

Tokuda, K., Nankaku, Y., Toda, T., *et al.*, 2013. Speech synthesis based on hidden Markov models. *Proc. IEEE*, **101**(5):1234-1252. http://dx.doi.org/10.1109/JPROC.2013.2251852

Turk, M., Pentland, A., 1991. Eigenfaces for recognition. *J. Cogn. Neurosci.*, **3**(1):71-86. http://dx.doi.org/10.1162/jocn.1991.3.1.71

Veselý, K., Ghoshal, A., Burget, L., *et al.*, 2013. Sequence-discriminative training of deep neural networks. Int. Conf. on Interspeech, p.2345-2349.

Wang, W., Xu, S., Xu, B., 2016. First step towards end-to-end parametric TTS synthesis: generating spectral parameters with neural attention. Int. Conf. on Interspeech, p.2243-2247. http://dx.doi.org/10.21437/Interspeech.2016-134

Xiong, W., Droppo, J., Huang, X., *et al.*, 2016. Achieving human parity in conversational speech recognition. arXiv:1610.05256.

Zhang, J.P., Wang, F.Y., Wang, K.F., *et al.*, 2011. Data-driven intelligent transportation systems: a survey. *IEEE Trans. Intell. Transp. Syst.*, **12**(4):1624-1639. http://dx.doi.org/10.1109/TITS.2011.2158001

Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: past, present and future. arXiv:1610. 02984.