Perception from an AGI Perspective

Pei Wang(✉) and Patrick Hammer

Department of Computer and Information Sciences, Temple University, Philadelphia, USA
{pei.wang,tuh38867}@temple.edu

Abstract. This paper argues that according to the relevant discoveries of cognitive science, in AGI systems perception should be subjective, active, and unified with other processes. This treatment of perception is fundamentally different from the mainstream approaches in computer vision and machine learning, where perception is taken to be objective, passive, and modular. The conceptual design of perception in the AGI system NARS is introduced, where the three features are realized altogether. Some preliminary testing cases are used to show the features of this novel approach.

1 The Nature of Perception

In general, "perception" refers to the organization and interpretation of sensory information during the interaction between the system and its environment. The perceptual process is usually taken as a multi-level generalization or abstraction, by which the sensory information of various modularity is gradually transformed and integrated into a concept-level description of the environment, then used to carry out various types of task, like the recognition of objects and events [20].

A representative and influential work in this field is Marr's work on vision [15]. Marr described vision as a computation where the input is a two-dimensional signal array and the output is a three-dimensional description of the world. The system implements an algorithm that carries out this computation. In the early years, algorithms for perception (vision, speech, etc.) were designed directly by the researchers. These algorithms extract certain predetermined features from the input, and then decide the output according to them. Later, machine learning let the computer system itself choose the features for a given problem, using the training data as guidance [6]. Most of the recent achievements of deep learning are obtained by designing special learning algorithms to take the advantage of the abundance of training data and computational power [14].

Though the current techniques work well on many problems, they lack generality and flexibility, and the processes and results are hard to explain. These issues are of special significance in AGI systems, where perception often faces novel situations, and real-time response is required. These problems are all well known, though most researchers attempt to solve them within the framework of an existing technique, such as deep neural networks. What we want to propose in this paper is an alternative approach.

M. Iklé et al. (Eds.): AGI 2018, LNAI 10999, pp. 259–269, 2018.
https://doi.org/10.1007/978-3-319-97676-1_25

This new approach toward perception in AGI is mainly based on the research results on human perception [2–5,7–13,16–19,22,23]. Because of the length restriction of the paper, in the following we cannot survey these results, but summarize them into three key features:

Subjective: Perception is a constructive process carried out according to the current needs of the system, and the sensory information is organized using the available percepts (patterns, mental images) and concepts (notions, categories) of the system. Consequently, different systems may perceive the same situation differently, and even the same system may perceive the same situation differently in various time and context, though some similarity can be expected. According to this opinion, perception should not be treated as a *function* or *computation* that maps every sensory input into a unique "correct representation", and the aim of perception should not be considered as creating an "objective model" of the world. Here "subjective" does not mean "arbitrary" or "random", but "depending on the system's past experience and current status".

Active: Perception should not be taken as a process in which the system *passively* processes the sensory information imposed on it by the user or the environment, but a goal-guided process in which the system selectively acquires certain information via the execution of its own operations. According to this opinion, perception is not a pure input process, but should be studied together with the related actions of the system. Perception is not mainly about *signal processing* or *pattern recognition*, but *sensorimotor coordination* where the system predicts the sensory effects of its own actions.

Unified: Perception should not be considered as carried out by a separate module that is independent of the other cognitive processes, but as closely tangled with them. In particular, many basic perceptual operations can be treated as inference, and learning in perception is not that different from learning in cognition in general. Though perception can still be considered mainly as a multi-level generalization with a certain degree of modularity, it is not a purely bottom-up process, but heavily influenced by top-down forces. In a system with multiple types of sensor, the integration of the information happens at early stages of the process, rather than until each modality-specific modules completes its work.

This new opinion about perception challenges the basic assumptions of many existing AI techniques, and is not completely unknown to the AI community. Various types of "top-down" influence introduce subjective factors into perception [28], the "active vision" approach integrates action into perception [1], and to include reasoning in perception is a hot topic in the deep learning research [21]. Even so, we have not seen an approach with these three features altogether. Furthermore, in most projects perception is still treated as objective, passive, and isolated.

In this paper, we explore a new direction with the above natures from the very beginning. Such an attempt cannot be accomplished soon, but there are reasons to give it a try. In the following we introduce a preliminary design, as a first step in this direction. The following design is an addition to NARS (Non-Axiomatic Reasoning System), which is an AGI system that has been described in a large

number of publications, including [25,27]. Limited by paper length, here we only describe the components of NARS that are directly related to perception.

2 Representation and Semantics

NARS uses the formal language *Narsese* for both internal representation and external communication, and its grammar is given in [27]. Narsese is a term-based language, in which each *term* is the identifier of a concept within the system. Unlike traditional "symbolic AI" systems, the meaning of a term in NARS is determined not by an entity outside the system it refers to, but by its experienced relations with other terms, and sometimes also by its built-in relations with certain sensorimotor component. Beside *atomic* terms, there are also *compound* terms composed from other terms by logical connectors, whose compositional relations with its components also contribute to the meaning of such a term [24].

As far as perception is concerned, terms can be divided into the following types:

- A *sensory term* is an array that represents concurrent sensations produced by the same type of sensor. An array can be 1-dimentional (vector), 2-dimensional (matrix), or 3-dimentional (space). The familiar format A[i,j,k] will be used to indicate a component in array A. For example, after every visual observation the sensors for brightness produce a 1024-by-1024 matrix B, where each 'pixel' B[i,j] represents the brightness produced by a sensor at the location indicated by the indexes.
- A *perceptual term* is also an array, though it is not directly produced by sensors, but constructed from other sensory and perceptual terms. For example, a perceptual term P can be obtained by taking a part of a sensory term S. More descriptions on this type of term are in the following.
- An *operational term* represents an executable operator, and an operation is an operator applied on a list of terms (as argument), which can be either a physical operation on the external environment, or a mental operation on the internal environment, i.e., the memory of the system. Operations can be compounds, too, formed from other operations recursively and hierarchically [26].
- An *abstract term* does not have direct sensorimotor association as the above, so is just an identifier that gets its meaning from its experienced or compositional relations with other terms [24].

Conceptually, sensory and perceptual terms can be taken as multi-dimensional spaces with a coordinate defined on each dimension in the range of $[-1, 1]$, though each space is stored discretely in an array. In this way, many operations on these terms can be defined independently of the storage size of the arrays involved. For instance, an element of a matrix A can be identified either as $A[i, j]$ with index i and j, or as $A(x, y)$ with coordinates x and y. For each dimension, the coordinate x and the index i (from 1 to N) can be linearly mapped into each other according to the relation $(x + 1)/2 = (i - 1)/(N - 1)$,

that is, $x = (2i - N - 1)/(N - 1)$ and $i = ((N - 1)x + N + 1)/2$. Since an index must be an integer, the mapping result for i may either be rounded, or used at both integers around it with a confidence discount, depending on the nature of the operation.

Terms are related by a number of *copulas* (which can be *inheritance*, *similarity*, *implication*, or *equivalence*) to form a *statement*, and its truth-value measures the evidential support the statement gets according to the system's experience. A truth-value consists of a pair of values, where the *frequency* value represents the proposition of positive evidence among all evidence, so is in [0, 1], while the *confidence* value represents the proposition of currently available evidence among all evidence at a future time after a constant amount of new evidence arrives, so is in (0, 1). NARS stands for "Non-Axiomatic Reasoning System", since in the system no empirical belief has the status of *axiom* whose truth-value cannot be adjusted by future evidence [24].

For perception, each group of sensor can be invoked by an operator to receive certain signal (which can be physical, chemical, biological, electrical, etc.), and the result corresponds to a statement $S \rightarrow [T]$, where S is a sensory term, T the type of the sensation, and '$\rightarrow$' the *inheritance* copula. In this context, the statement just classifies the sensation as of a certain type. Since S is an array, each element in it stores a Narsese truth-value, where *frequency* is intuitively the "strength" of the sensation, and *confidence* is intuitively the "reliability" of the sensation. The truth-values at different locations of the same array can be different, where the *frequency* distribution corresponds to the spatial pattern of the sensation, and the *confidence* distribution may summarize various factors like noise, resolution, attention, etc. In particular, a perceptive field of any shape can fit into a multi-dimensional array by assigning the irrelevant elements a 0 as confidence, so they will make no impact to the following perception process.

3 The Construction of Perceptual Terms

Terms in NARS can be obtained directly from the system's experience, or constructed by the system from the existing terms using composing/decomposing rules [27]. For the current discussion, sensory terms are produced by the sensors, while perceptual terms are constructed by the system from the existing sensory or perceptual terms.

To directly construct a perceptual term B from a sensory term A, four parameters are needed. Taken 2-dimensional terms as example: a pair of coordinate (x, y) is taken to set a focus point at $A(x, y)$ to be used as the center of B. The other two parameters are used to decide the scope of perception: a *center* value will be the radius of the circular area around $A(x, y)$, in which the truth-values of A will be copied into B as they are; a *boundary* value will be the width of the peripheral zone around the central area, in which the truth-values of A will be copied into B after a discount factor is multiplied to the confidence value, and this factor decreases linearly from 1 to 0 when the point moves away from the center. This operator will get a circular copy of a part of A, with the

boundary blurred gradually. The elements of B outside the boundary will all have confidence 0. For default, we set $x = y = 0, center = boundary = 0.5$.

Perceptual terms can also be constructed from other perceptual terms by a mental operator that adjusts the parameters, where '⇑' is the prefix of operators:

- $\Uparrow focus(x, y)$ will set the focus point to the given coordinates.
- $\Uparrow shift(dx, dy)$ is effectively $focus(x + dx, y + dy)$. This operator allows the focus point to be adjusted relatively to the current position.
- $\Uparrow zoom(z)$ changes *center* and *boundary* by multiplying z to them. When $z > 1$, it is "zoom out"; When $z < 1$, it is "zoom in".
- $\Uparrow rotate(a)$ turns the perception around the focus point clockwise to the angle a.

Another group of constructors corresponds to the term connectors that are already defined in NARS among statements: *disjunction*, *conjunction*, and *negation* [27]. For the latter, the NARS negation rule is applied to every element of an array to get the negated perceptual term; for the formers, elements of two given arrays are processed pair by pair by the disjunction or conjunction rule to get the new array. If the given arrays have different sizes in terms of storage space, coordinates are used to map one to the other before they are combined.

Using these constructors, a sensation of arbitrary complexity can be perceived as a compound term consists of existing terms combined using the term connectors and mental operations recursively and hierarchically. Perceptual knowledge will be integrated with the other types of knowledge in NARS, including declarative (eternal), episodic (temporal), and procedural (operational). A typical statement in NARS will not be part of a description of the world as it is, but is more like "When the condition c is satisfied, if I execute operation o, I will perceive its effect e", which is an extension of the previous form of procedural knowledge described in [27].

4 Perception via Inference

All terms in NARS are treated by the inference rules basically in the same way, no matter whether the term is associated directly with a sensorimotor component (like the sensory, perceptual, and operational terms). Consequently, inference can be carried out among mental images and operations, just like among abstract concepts.

There are special variants of rules that are dedicated to sensorimotor mechanism. For example, temporal induction/comparison do not require shared term in the premises, but their closeness in time. Similarly, spatial induction/comparison can be carried out among array elements that are close spatially to each other, so as to achieve functions like auto-filling, associative memory, and "perceptual set", which is a perceptual bias or predisposition or readiness to perceive particular features of a stimulus.

Inheritance/similarity statements between arrays can be built between sensations and perceptions of the same type. From $S_1 \rightarrow [T]$ and $S_2 \rightarrow [T]$, by

abduction $S_1 \rightarrow S_2$ and $S_2 \rightarrow S_1$ can be derived. While in ordinary abduction each premise only has one truth-value, here both S_1 and S_2 are arrays, so abduction between the corresponding element pairs are carried out first [27], then the results are merged by the revision rule to get an overall truth-value for the relation between the two arrays.

As perception is closely related to the system's operations, 3-D perception may start at the three degrees of freedom of body movements, combined with the feedbacks in the related sensorimotor channels (visual, auditory, kinesthetic, tactile, etc.). Consequently, an object is usually represented according to the system's interaction with it, or its "affordance" [7], rather than "as it is".

As movements are sequence of events, object movements are similarly perceived with compensation of movements of sensor and perceptive field. Like other knowledge, such compensation is learned by the system in its interaction with the environment.

NARS supports multiple input/output channels. Besides the primary channels that directly recognize Narsese tasks, there can also be multiple sensory channels, each dedicated to a special type of sensor or several types of related sensor. Within the system, there is also an "overall experience" channel that is not directly connected to any sensor, but integrates significant events from all other channels.

Perception is the process where relations are derived among the sensory terms, as well as between them and the other (non-sensory) terms. Beside the semantic relations provided by the copulas and the syntactic relations by the term connectors, there are also temporal-spatial relations directly coming from the input channels.

As a result of processing sensory experience, spontaneous forward inference happens as far as the significance of the signal is above the threshold of the sensory channel, which can be adjusted by factors including the system's anticipation, extent of busyness, emotional status, etc. This spontaneous inference can be triggered by the results of the system's observation operations.

Perception will summarize the sensory experience into descriptions at multiple levels of generalization and abstraction in parallel, where the array-based "sensory" representation and the concept-based "symbolic" representation will co-exist. The system represents the situation both as a mental image and as a judgment like "A cat is on a mat", where the latter is formed by matching the parts of the image with concepts in the system and recognizing their relations. These two types of representation will interweave at all levels and are irreducible into each other. An image corresponds to an existing concept will be remembered better and accessed easier than an incomprehensible image. This feature should allow the model to explain phenomena like Gestalt shapes, visual illusions, Bongard figures, and so on.

During perception, the bottom-up signal-compression and the top-down anticipation will form a mutual confirmation process. The sensory input first suggests some patterns with associated concepts, and anticipation and inference then increased the confidence of the suggestions, which in turn lead the fill-in of

details. As the system changes its internal states, it is normal for the same situation to be perceived differently, with different objects and events recognized. The result of perception is under constant revision with the coming of new experience, as well as with the continuous thinking process of the system. Therefore, the perception mechanism is not a function that maps the input signals into a unique "correct" representation. Instead, it will be more similar to the human perception process.

Beside the automatic self-organizing process in perception, the most common deliberative tasks are "recognition" and "imagination". Roughly speaking, the former is to find a concept for an image, while the latter is to find an image for a concept, where the relation from the image to the concept is the inheritance copula. In NARS, both processes are carried out by inference, with all types of uncertainty involved, and the final answer is chosen among the available candidates by balancing truthfulness, simplicity, and usefulness [25].

5 A Simple Example

The conceptual design described above is being experimented in NARS, and currently the sensory terms have been implemented, while the perceptual terms have not. While our prototype is at an early stage, we can nevertheless demonstrate some results on gray scale images, as well as using such a concrete example to explain the proposed approach to perception.

The first example is to choose a label for a given image. To keep the example simple, 5×5 images are used. Initially, a diamond, M_1, and a cross, M_2, are entered as Narsese sentences and categorized. In the input, the pixels not mentioned are black by default:

```
//Input: Bright pixels in M1:
<{M1[-1.0,0.0]}  --> [bright]>.
<{M1[1.0,0.0]}   --> [bright]>.
<{M1[0.0,1.0]}   --> [bright]>.
<{M1[0.0,-1.0]}  --> [bright]>.
<{M1[0.5,0.5]}   --> [bright]>.
<{M1[-0.5,0.5]}  --> [bright]>.
<{M1[0.5,-0.5]}  --> [bright]>.
<{M1[-0.5,-0.5]} --> [bright]>.
//It is a diamond:
<{M1} --> diamind>.
```

```
//Input: Bright pixels in M2:
<{M2[0.0,1.0]}  --> [bright]>.
<{M2[0.0,0.5]}  --> [bright]>.
<{M2[-1.0,0.0]} --> [bright]>.
<{M2[-0.5,0.0]} --> [bright]>.
<{M2[0.0,0.0]}  --> [bright]>.
<{M2[0.5,0.0]}  --> [bright]>.
<{M2[1.0,0.0]}  --> [bright]>.
<{M2[0.0,-1.0]} --> [bright]>.
<{M2[0.0,-0.5]} --> [bright]>.
//It is a cross:
<{M2} --> cross>.
```

Then a noisy pattern M_3 is entered, and followed by a question asking what it is:

```
//Input: Pixels at these locations in M3 are bright or half-bright:
<{M3[-1.0,1.0]}  --> [bright]>. %0.5%
<{M3[0.0,1.0]}   --> [bright]>.
<{M3[-0.5,0.5]}  --> [bright]>.
<{M3[0.5,0.5]}   --> [bright]>. %0.5%
<{M3[-1.0,0.0]}  --> [bright]>. %0.5%
<{M3[1.0,0.0]}   --> [bright]>.
<{M3[-0.5,-0.5]} --> [bright]>.
<{M3[0.5,-0.5]}  --> [bright]>. %0.5%
<{M3[1.0,-0.5]}  --> [bright]>. %0.5%
//How to categorize M3?
<{M3} --> ?what>?
```

From these inputs, by merging pixel-wise comparisons of the matrices, two similarity judgments are derived, then by analogy, the new pattern is recognized as most likely to be a diamond (among the existing categories):

```
//M3 is quite similar to M1
<M1 <-> M3>. %0.61;0.88%
//M3 is not similar to M2
<M2 <-> M3>. %0.19;0.91%

<{M3} --> diamond>. %0.61;0.48%  //M3 is likely a diamond
<{M3} --> cross>. %0.19;0.16%    //M3 is unlikely a cross

Answer <{M3} --> diamond>. %0.61;0.48% //System answer, M3 is taken as a diamond
```

After the perceptual terms are fully implemented, this example will be enriched further, using the mental operators introduced previously. We can imagine an input matrix M_4 which looks like a diamond above a small cross (which will surely need a large matrix than 5×5). At the beginning the system will attempt to classify the new sensation using the existing categories. Since in NARS every conclusion is true to a degree, such an attempt often can produce some answer, even though the quality of the solution will not be very high. For this example, M_4 will probably have a relatively higher similarity to M_1 (by ignoring the small cross) than to the other candidate. If the system is not satisfied enough by this conclusion, it will continue to look for better answers by decomposing M_4 into simpler shapes plus some structures combining them.

Starting at default parameters at the sensation M_4, an operation "$\Uparrow shift(0, 0.5)$" will turn its top part into a perceptual term M_{41}, which matches reasonably well with M_1, so "$\{M_{41}\} \rightarrow diamond$" can be derived, which will have less negative evidence than "$\{M_4\} \rightarrow diamond$".

After that, operation "$\Uparrow shift(0, -0.8)$" followed by operation "$\Uparrow zoom(0.4)$" on the current sensation will generate M_{42} that matches M_2, a cross of the default size. Now the question "$\{M_4\} \rightarrow ?what$" will be answered by judgment

$$\{M_4\} \rightarrow (\Uparrow shift(0, 0.5), M_{41}, \Uparrow shift(0, -0.8), \Uparrow zoom(0.4), M_{42})$$

which will have less negative evidence than the other candidate answers, though being more complicated in syntax.

Of course, the above result assumes a proper sequence of mental operations. In the initial experiment, it can be either predetermined or obtained from exhaustive search, while in the future it will be learned together with the components themselves. That means the system's knowledge about an image also includes information on how it is usually perceived as a sequence of events and operations.

With a properly trained natural language interface, M_4 can be described as "A diamond above a small cross". The given knowledge used in the example, such as "$\{M1\} \rightarrow diamond$", can also be learned from the repeated co-occurrence of an image and a word in the system's experience, as they will both be associated with a concept in the system which is named by *diamond*. However, it is important to remember that in NARS, neither the image of a diamond nor the word "diamond" will be used to "define" the term *diamond* (or whatever the term is labeled), as the meaning of the an abstract term like *diamond* is not determined only by its (visual) exemplifiers or (verbal) labels, but also by its relations with the other terms, including the abstract ones.

Though only partially implemented, this example still shows the desired features of this new approach to perception when compared with the conventional computer vision techniques:

- **Subjective**: The answer to a question like "$\{M_4\} \rightarrow ?what$" not only depends on M_4, but also on the existing knowledge of the system and its resource allocation situation when the question is processed.
- **Active**: The answer "$(\Uparrow shift(0, 0.5), M_{41}, \Uparrow shift(0, -0.8), \Uparrow zoom(0.4), M_{42})$" contains operational components, so perception is based on action.
- **Unified**: The question answering process is carried out by the inference rules, and mingled with all the other co-existing processes in the system.

6 Discussion

This paper proposes a new conceptual design for perception in AGI systems. Though this approach has not been fully implemented in NARS, and no enough empirical results have not been obtained to support a definite conclusion about its feasibility, the design nevertheless has the desired features observed in human perception.

Psychologists have reached the consensus long ago that perception is multi-level abstraction, and deep learning just realizes this in special-purpose systems [14]. The approach we proposed also has the potential to carry out multi-level abstraction, though with the following characteristics that distinguishes it from deep learning and the other traditional approaches:

- Using meaningful term connectors to carry out abstraction from level to level. It is assumed that the existing term connectors of NARS [27] are sufficient for all necessary patterns — convolution and neuron models are basically weighted average functions followed by a non-linear step, which should be achievable using the set-theoretic operators of NARS.
- Carrying out multiple tasks, so the intermediate results are not bounded to a single task, but have independent meaning. Therefore, learning results are naturally transferable. As there is no distinction between "hidden layer" and "input/output layer", results at any layer are understandable (to various degrees), and are adaptive with experience-grounded meaning.

- Using multi-level abstraction to solve "over-fitting" and "inductive bias", and to keep multiple hypotheses for a given problem. For the same observation, more abstract results are less confident, though they are simpler and can be supported by other observations later, so can become preferred than the more specific results.
- Using dynamic resource allocation to carry out local and incremental adjustments to provide real-time responses. Compared to the global iterations demanded by neural network models, this approach can meet various time requirements associated with the tasks. The control mechanism of NARS is not introduced in this paper, but can be found in other publications on NARS, such as [25].
- Having stronger top-down influences, in the form of anticipation, familiarity, emotion, etc. The existing conceptual hierarchy plays a significant role in deciding what is perceived, while being adjusted in the process, as Piaget's assimilation-accommodation process, with stable perceptions as their equilibrium [18].
- Integrating perception with action, in the sense that (1) perception is carried out by operation, (2) perception and operation have unified representation, and (3) perceptive patterns are identified as invariants during related operations.

Like the other processes, perception in NARS will not attempt to simulate human perception in all details, but its general principles and major features. Consequently, it will still be closer to human than the existing AI techniques.

This research is still at its early stage, so the purpose of this paper is to raise this possibility for the AGI community to consider and discuss. Though there are many issues to be resolved, there are reasons to believe that this is a suitable approach for AGI systems to carry out perception.

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. Int. J. Comput. Vis. **1**(4), 333–356 (1988)
2. Barsalou, L.W.: Perceptual symbol systems. Behav. Brain Sci. **22**, 577–609 (1999)
3. Brette, R.: Subjective physics. arXiv:1311.3129v1 [q-bio.NC] (2013)
4. Chalmers, D.J., French, R.M., Hofstadter, D.R.: High-level perception, representation, and analogy: a critique of artificial intelligence methodology. J. Exp. Theor. Artif. Intell. **4**, 185–211 (1992)
5. Di Paolo, E.A., Barandiaran, X.E., Beaton, M., Buhrmann, T.: Learning to perceive in the sensorimotor approach: Piaget's theory of equilibration interpreted dynamically. Front. Hum. Neurosci. **8**, 551 (2014)
6. Flach, P.: Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press, New York (2012)
7. Gibson, J.J.: The theory of affordances. In: The Ecological Approach To Visual Perception, New edn. Chap. 8, pp. 127–143. Psychology Press (1986)
8. Goldstone, R.L., Barsalou, L.W.: Reuniting perception and conception. Cognition **65**, 231–262 (1998)

9. Hatfield, G.: Perception as unconscious inference. In: Heyer, D., Mausfeld, R. (eds.) Perception and the Physical World: Psychological and Philosophical Issues in Perception, pp. 113–143. Wiley, New York (2002)
10. Hockema, S.A.: Perception as prediction. In: Proceedings of the Cognitive Science conference (2004)
11. Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W.: The theory of event coding (TEC): a framework for perception and action planning. Behav. Brain Sci. **24**(5), 849–78 (2001)
12. Jarvilehto, T.: Efferent influences on receptors in knowledge formation. Psycoloquy **9**(41), Article 1 (1998)
13. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York (1998)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
15. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman & Co., San Francisco (1982)
16. Noë, A.: Action in Perception. MIT Press, Cambridge (2004)
17. O'Regan, J., Noë, A.: A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. **24**(05), 939–973 (2001)
18. Piaget, J.: The Construction of Reality in the Child. Basic Books, New York (1954)
19. Rock, I.: The Logic of Perception. MIT Press, Cambridge (1983)
20. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice Hall, Upper Saddle River (2010)
21. Santoro, A., et al.: A simple neural network module for relational reasoning. CoRR abs/1706.01427 (2017), http://arxiv.org/abs/1706.01427
22. Shams, L., Shimojo, S.: Sensory modalities are not separate modalities: plasticity and interactions. Curr. Opin. Neurobiol. **1**, 505–509 (2001)
23. Shanahan, M.: Perception as abduction: turning sensor data into meaningful representation. Cogn. Sci. **29**(1), 103–134 (2005)
24. Wang, P.: Experience-grounded semantics: a theory for intelligent systems. Cogn. Syst. Res. **6**(4), 282–302 (2005)
25. Wang, P.: Rigid Flexibility: The Logic of Intelligence. Springer, Dordrecht (2006). https://doi.org/10.1007/1-4020-5045-3
26. Wang, P.: Solving a problem with or without a program. J. Artif. Gen. Intell. **3**(3), 43–73 (2012)
27. Wang, P.: Non-Axiomatic Logic: A Model of Intelligent Reasoning. World Scientific, Singapore (2013)
28. Wu, T.: Integration and goal-guided scheduling of bottom-up and top-down computing processes in hierarchical models. Ph.D. thesis, University of California, Los Angeles (2011)