Preprint of the article. Published in Kuenstliche Intelligenz, Springer, 2015:
http://dx.doi.org/10.1007/s13218-015-0351-6

Noname manuscript No.
(will be inserted by the editor)

Towards an Embodied Developing Vision System

İlkay Atıl, Sinan Kalkan

Received: date / Accepted: date

Abstract Many cognitive scientists now agree that artificial cognition might be probably achieved developmentally, starting from a set of basic-level premature capabilities and incrementally self-extending itself with experience through discrete or continuous stages bred with experience. Although we are still far from seeing an artificial full-fledged self-extending cognitive system, the literature has provided promising examples and demonstrations. Nonetheless, not much thought is given to the modeling of how an artificial vision system, an important part of a developing cognitive system, can develop itself in a similar manner. In this article, we dwell upon the issue of a developing vision system, the relevant problems and possible solutions whenever possible.

Keywords Vision System · Developmental · Embodiment

1 Introduction

Vision is the process of understanding scenes from their 2D projections, which are in the form of a set of images. The intensity values in an image are formed by one or more of the following factors: (1) the geometry, and the illumination of the environment, (2) the spatiotemporal states of the objects and the viewer, (3) the reflectance of the surfaces, and (4) the type of the medium that light travels. By definition, this makes vision an *ill-posed*[1] *inverse* problem (Bertero et al, 1987).

I. Atil, S. Kalkan
Dept. of Computer Engineering
Middle East Technical University
E-mail: {ilkayatil,skalkan}@ceng.metu.edu.tr

[1] According to Hadamard (1923), a problem is well-posed if (1) a solution exists, (2) the solution is unique, and (3) it depends continuously on the data. A problem is ill-posed if it is not well-posed.

An important mechanism used by the human visual system to deal with the ambiguities and the missing information that is due to the ill-posed nature of the problem is to make use of the regularities in images. It is proposed that the human visual system is adapted to the statistics of the retinal projections of the environment, in order to make use of the regularities or the redundancy of information in the environment. With the availability of computational and technological means, it has been possible to prove such claims (Krüger, 1998; Wagemans et al, 2012), and the results of such investigations have proven to be useful in several computational vision problems (Elder et al, 2003; Pugeault et al, 2004; Zhu, 1999) - see Simoncelli (2003) for a review.

Another important mechanism to deal with the ill-posed nature of the problem is to make the problem well-posed. One such mechanism is to collect more data from the scene, which can easily be achieved if the camera is integrated on an embodied agent interacting with the environment. The motivation for this comes from active perception studies, which showed that, for example, with an actively moving camera, ill-posed problems such as structure from motion, shape from shading become well-posed (Aloimonos and Rosenfeld, 1994; Aloimonos, 1990).

These mechanisms suggest that a vision system should be embodied and it should utilize regularities of the environment. By doing so, we claim that one can talk about an artificial vision system that initially starts with only few vision capabilities, and develops itself gradually, by interacting with the environment and constantly self-extending itself by exploiting the regularities of the environment.

Piecemeal efforts have shown that, e.g., grouping can be demonstrated to be the manifestation of the exploitation of the regularities in the environment (Elder et al, 2003; Pugeault et al, 2004; Zhu, 1999). Moreover, with frameworks exploiting important principles like sparsity and slowness (such as sparse autoencoders and slow feature analysis), features can be discovered from a large collection of visual input - see, *e.g.*, (Bengio, 2009; Bengio et al, 2013; Schmidhuber, 2014). In this paper, we claim that we need a unified system that can bring together such principles and mechanisms to enable a vision system to develop itself with experience such that it can go beyond grouping of low-level features, recognition of events or objects towards a system that can learn new relations, cues, associations to solve high-level tasks such as depth perception and scene understanding. For this end, we first review and discuss the embodied approach to cognition and the importance of (statistical) regularities for vision and cognition in general. Then, we discuss some of the important aspects that a developing vision system should address, and provide guidelines whenever possible.

2 Developmental Approaches to Cognition and Intelligence

There is not an absolute recipe for creating a cognitive, intelligent system. Since ancient times, humanity has tried understanding the process of problem solving and logical thinking. However, millenniums have past and still, the underlying mechanisms of cognition and intelligence are largely unknown.

Inspired by the mathematical logic, the main approaches to artificial intelligence are (i) rule-based systems manipulating symbols for solving problems (Fikes and Nilsson, 1972; Newell and Simon, 1976; Fodor, 1981) in which the environment or the problem setup is expressed in terms of symbols and the rule-based system tries to find a valid derivation from the current state to the goal state; or (ii) learning-based where supervised and unsupervised mechanisms (see, *e.g.*, (Selfridge, 1958; Minsky, 1963; Samuel, 2000; Boden, 2006)) are used to learn a mapping from inputs to outputs. However, such systems can work only in very specific environments and they lack the ability to acquire new abilities and solve problems which their designers did not anticipate.

Developmental approaches to artificial intelligence is a promising alternative to classical artificial intelligence approaches (Asada et al, 2009; Lungarella et al, 2003; Cangelosi et al, 2010; Overton, 2003). It is promising because we have a working example: us, humans. Putting aside the problems of differences in embodiment, by endowing an artificial agent with similar developmental abilities, one can aim to achieve a cognitive system given sufficient time for development. In a sense, instead of fishing for the agents, we build them with fishing abilities. In the most general definition, there are two types of what can develop: (1) transformation and (2) variation (Overton, 2003). Transformational changes happen in the form, organization or structure of a system (*e.g.*, formation of new neuron connections in the brain, new tissue, organs etc.). Variational changes, on the other hand, represent degree of deviation from a standard or simply adaptation of a system (*e.g.*, adjusting weights of a neural network, better motor control accuracy of an infant, etc.). Transformational changes are important because they provide emergence of new abilities whereas variational changes provide improving those new abilities. Hence, a good developmental system should exhibit both transitional and variational changes.

The aim of the developmental approaches to cognition and intelligence is to provide developmental tools which start with simple and low-level abilities and transform into complex and high-level abilities. Garrett (1946) states that intelligence changes in its organization from a general ability into a more organized group of abilities as time passes. The developmental nature of infant cognition and intelligence has been shown and claimed by many (Fry and Hale, 1996; Kirkham et al, 2002; König and Krüger, 2006). Such developmental approaches are also used in machine learning studies (Elman, 1993). However, achieving a theory of developmental cognition and intelligence is not a straightforward work.

There are many theories inspiring for developmental cognition and perception. As reviewed by (Cohen and Cashon, 2003), Piagetian theory, Gibsonian theory, dynamical systems and connectionist modeling are influential theories of cognition and perception. Additionally, nativism, cognitive neuroscience and information processing are approaches to understand and provide computational models of cognition and perception.

Among these theories, the Gibsonian theory is notable since it has been very influential in cognitive robotics since 2000s. The Gibsonian theory relies on the concept of affordances, which connects agents to the environment through what the environment provides to the agents and emphasize the importance of an agent's embodiment (Gibson, 1977). In the Gibsonian theory, a developing agent (or an infant) performs two things; (1) it discovers new *affordances*, *i.e.*, the opportunities provided by the environment, and (2) it learns to differentiate relevant information from non-relevant information based on the affordances (Gibson and Gibson,

1955; Gibson, 1969). Experiments show that infants interact with the environment and learn new affordances and how to differentiate information to improve their behaviors (Gibson and Walk, 1960; Adolph et al, 1993).

Artificial cognitive systems have made significant progress using developmental approaches. Many studies pointed out that developmental approaches are important and beneficiary for robotics (Thelen and Smith, 1994; Ferrell and Kemp, 1996; Brooks, 1997; Sandini, 1997; Asada et al, 2009; Lungarella et al, 2003; Cangelosi et al, 2010). Robotics systems with developmental properties are able to acquire new abilities as they interact with their environment. These systems use their interaction experiences to further develop their perception and cognitive abilities to solve new problems. See (Asada et al, 2009; Lungarella et al, 2003; Cangelosi et al, 2010; Kraft et al, 2010) for detailed reviews.

3 Statistical Regularities of the Environment

Our environment, even nature without man-made structures, has structure, or regularities: objects, their shape, their texture, their locations, their behaviors etc. have certain patterns, or regularities. A good proof of this is the fact that we can recognize objects, events or scenes when we see them; we can recognize things because they have distinctive characteristics, structure, regularities that allow us to distinguish them.

We, humans, perceive the regularity in the environment with our limited senses. Our neuronal processing mostly relies on linking a neuron to the frequently firing neurons in its receptive field, in a sense, basing its activity on the statistics of the firing activities in its receptive field. From this, it is not far-fetched to claim that our neuronal machinery is designed for capturing regularities via statistical mechanisms, and we call them statistical regularities in this paper.

Statistical regularities are important for a cognitive entity facing real-world problems since they are used for learning, adaptation, inference, etc. (Vapnik, 2000). Statistical regularities enable a learning system to make the transition from low-level rapidly changing information to high-level, more stable symbols or concepts (König and Krüger, 2006) which are considered as the grounding of cognition, logical thinking and intelligence (Harnad, 1990; Rocha, 1997; Glenberg and Robertson, 2000). Hence, detecting statistical regularities is a necessary ability for any cognitive/intelligent system.

In fact, it has been shown and claimed by many that our brain is adapted to exploit statistical regularities in the environment (Saffran et al, 1996; Barlow, 2001; Altamura et al, 2014). It has also been shown that our brains constantly seek statistical regularities in the environment even when we do not pay attention to such regularities (Turk-Browne et al, 2009, 2010; Barakat et al, 2013), for vision (Fiser and Aslin, 2002; Kirkham et al, 2002; Kellman and Arterberry, 1998; Brunswik and Kamiya, 1953), speech and language (Tomasello, 2009; Saffran, 2003; Smith and Yu, 2008), touch, theory formation, etc. (Gopnik and Schulz, 2004; Conway and Christiansen, 2005).

3.1 Statistical Regularities of the Environment and Vision System

The amount of images that can be observed in nature is a very small subset of the possible images that can be constructed using arbitrary combinations of intensity values (Field, 1994). This suggests that the natural images bear intrinsic regularities which are believed to be exploited by our visual system for perceiving the environment (see, *e.g.*, Krüger and Wörgötter (2004)), especially for resolving ambiguities inherent in local processing of various visual modalities.

For example, it is widely acknowledged that Gestalt principles for perceptual organization are manifestations of our visual system's adaptation to the statistical regularities in natural scenes. This hypothesis was first pointed out by Brunswik and Kamiya (1953), but could not be tested or justified until 90s due to insufficient computational means. Field et al (1993) used computer-generated randomly-oriented data to develop a theory of contour grouping in the human visual system, called the *association field*. In 1998, Krüger (1998) used natural images instead of computer generated data to prove the relation between grouping mechanisms and the natural image statistics. Such investigations were extended in (Elder and Goldberg, 2002; Geisler et al, 2001; Krüger and Wörgötter, 2002), and the results were utilized in several computer vision tasks, including contour grouping, object recognition and stereo (see, *e.g.*, Elder et al (2003); Pugeault et al (2004); Zhu (1999)).

Statistical regularities of natural images also helped researches to understand the principles of sensory coding in the early stages of visual processing. It was shown that Independent Component Analysis and Principle Component Analysis of image patches from natural images produce Gabor-wavelet like responses which are believed to be what the V1 cells in the human visual system are doing (see, *e.g.*, Simoncelli and Ohlshausen (2001); Simoncelli (2003) for a review).

Availability of relatively cheap range scanners made it possible to analyze the statistical properties of the 3D world together with its 2D image projections. Such

analyses are important (1) for quantifying and understanding the assumptions that the vision researchers have been making and (2) for understanding the intrinsic properties of the 3D world. In (Yang and Purves, 2003; Huang et al, 2000; Potetz and Lee, 2003), the correlation between the properties of the 3D surfaces (like roughness, 3D orientation, distance, size, curvature etc.) and the intensity of the images are analyzed. Such studies mainly justify assumptions made by shape from shading studies and confirm that natural scene geometry is quite regular and less complex than luminance images. In (Kalkan et al, 2006), a higher-order representation of the 2D local image patches and the 3D local patches were considered, and the probability of observing a certain kind of the 3D structure given its 2D projection is provided. Similarly, Kalkan et al (2007) studied the predictability of depth at constant-intensity image regions from the depth available at the edges, and showed that, using a co-planarity constraint, depth at constant-intensity image regions can be predicted, modulated by the distance to the edge segments. Moreover, range image statistics allow explanation of several visual illusions (Howe and Purves, 2002, 2004).

Krüger and Wörgötter (2004) provide a summary of the evidences from developmental psychology which suggest that depth extraction based on statistical regularities used in perceptual organization develops at a later stage than depth extraction based on stereopsis and motion. In particular, it is discussed that perceptual organization based on edge structures are in place after approximately 6 months of visual experience but not before (Kellman and Arterberry, 1998; Spelke, 1990). As we have mentioned before (Kalkan et al, 2007):

> "*This indicates that experience may play an important role in the development of these cues, i.e., that we have to understand depth perception as a statistical learning problem (Knill and Richards, 1996; Purves and Lotto, 2002; Rao et al, 2002). A step towards such an understanding is the investigation and use of the statistical relations between the local image structures and the underlying 3D structure for each of these depth cues (Knill and Richards, 1996; Purves and Lotto, 2002; Rao et al, 2002).*"

4 Statistical Regularities in Computer Vision

Statistical regularities of the environment that we discussed in the previous section are exploited in Computer Vision literature to a certain extent. Gestalt laws and grouping are good examples for this. For example,

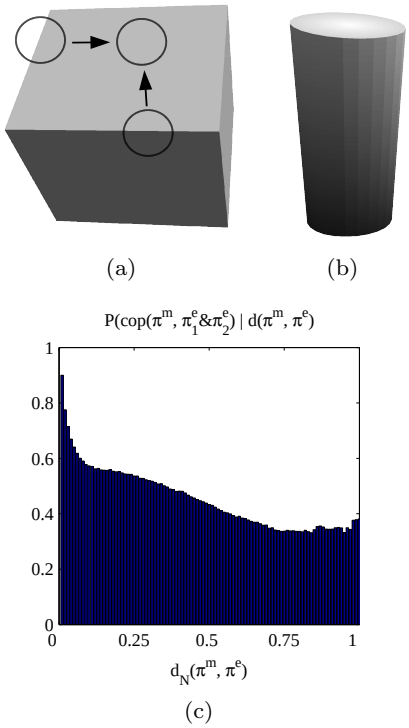

Fig. 1 Depth predictability using co-planarity of edge segments. (a) On planar surfaces, co-planarity of edge segments could be used for predicting the depth at an homogeneous image region. (b) Such a cue might not work on round surfaces directly since the boundaries are not coplanar with the surface. (c) If we look at how much co-planarity account for the predictability of depth from the boundaries (plotted against distance from the boundaries), we see that it is very high closer to the boundaries (Source: Kalkan et al (2007)).

Elder and Goldberg (2002) used the statistics of images to learn Gestalt laws and studied the contour extraction problem using the Gestalt grouping rules learned from statistics of edge segments in images (see, *e.g.*, (Sarkar and Boyer, 1993) for a review on alternative approaches).

Another example is the utilization of the analysis of the depth structure in relation to the corresponding 2D image. In this line, Kalkan et al. studied the predictability of depth at constant-intensity image regions from the depth available at the edges, showed that, using a coplanarity constraint, depth at constant-intensity image regions can be predicted, modulated by the distance to the edge segments (Kalkan et al, 2007) (see also Figure 1) and then used these results to inspire a computational model that predicted depth at constant-intensity image regions using the coplanarity constraint (Kalkan et al, 2008) - see also Figure 2.

A more prominent example for the utilization of the statistics of the environment started with the discovery that the independent components of natural im-

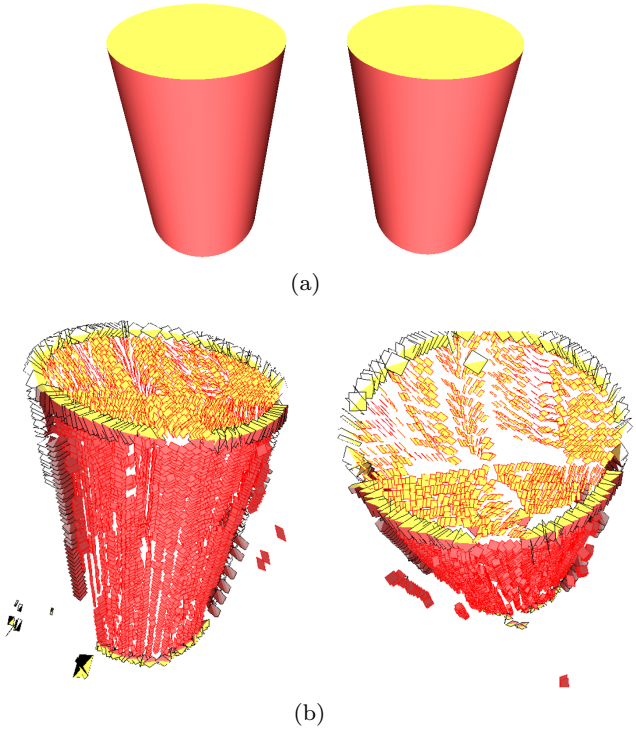

Fig. 2 Using the results in (Kalkan et al, 2007) (see Figure 1) as inspiration, one can predict depth at homogeneous image regions using coplanarity of edge segments even on round surfaces. **(a)** Input stereo pair. **(b)** The predictions of our model as a disparity map. (Source: (Kalkan, 2008))

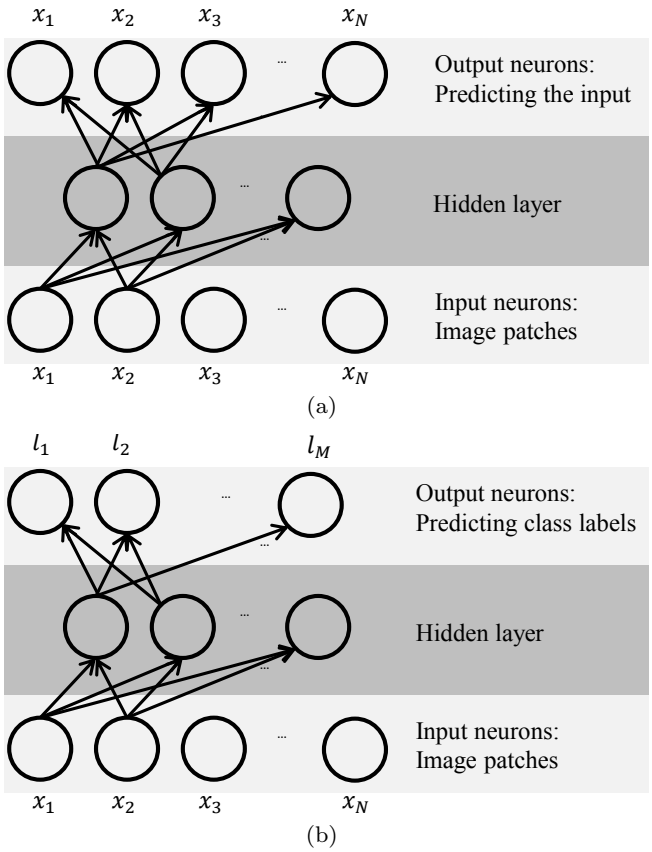

Fig. 3 (a) Generally, in feature learning, first a sparse representation of the input is learned. (b) After learning the "features" relevant for the problem, the output layer is replaced with neurons that predict class labels.

ages resemble V1-like cells in visual cortex (Olshausen and Field, 1997; Hyvärinen and Hoyer, 2001; Hyvärinen et al, 2009). This discovery ignited a line of research, called feature or representation learning, that led to many successful applications in Computer Vision. In this line of work, a neural network is (generally) trained to predict its input first, through which, in hidden layers, it learns a representation of what is common in the input - see Figure 3. In this approach, the neural network learns how to encode the input itself in a lower-dimensional space using sparsity as a principle - hence, the method is called *sparse auto-encoding*. The sparseness principle, the slowness principle (Berkes and Wiskott, 2005) and similar "transformation to a lower dimensional space for representation learning" approaches are abundant in the literature and a very trendy research topic in Computer Vision these days - see, *e.g.*, Scalzo and Piater (2005); Fidler and Leonardis (2007); Franzius et al (2008); Ciresan et al (2012); LeCun et al (1998); Hinton et al (2006); Salakhutdinov and Hinton (2009); Rifai et al (2012). For reviews on this emerging approach, please refer to (Bengio, 2009; Bengio et al, 2013; Schmidhuber, 2014). As an example, see Figure 4 and (Firat et al, 2014), where, for recognizing landing lanes in airports, a set of features are learned and used.

These successful applications in Computer Vision inspired from vision system's adaptation to statistical regularities of the environment are very promising, but limited to recognition and classification tasks. We propose that inspiration from vision system's adaptation to statistical regularities of the environment should be tailored towards building a vision system that develop itself to not only recognize objects but also solve the high-level vision tasks that we, humans, can solve.

5 Aspects for a developing vision system

It has been suggested that our vision system starts only with few premature vision capabilities (*e.g.*, motion detection, depth from motion, stereopsis - (Kellman and Arterberry, 1998)) and through extensive interactions with the environment and experience, develops new capabilities, not only for recognizing and categorizing objects, scenes or events but also for spatio-temporal understanding of scenes, events, and eventually a 3D in-

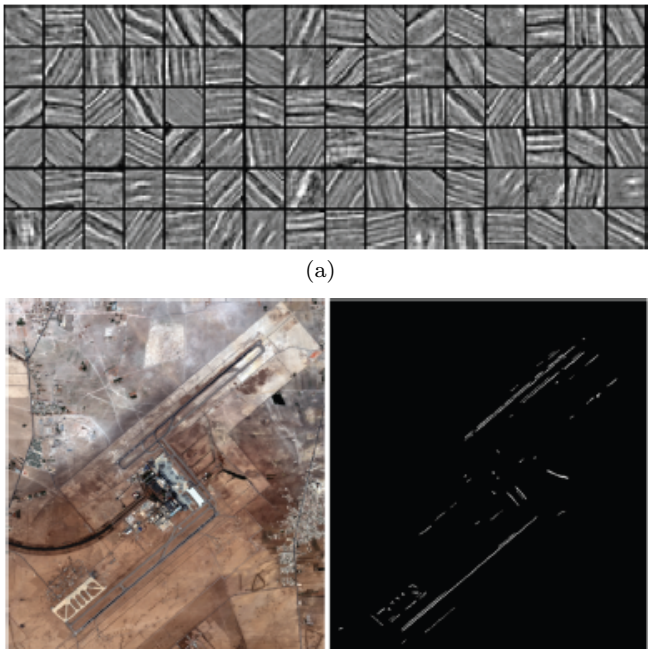

Fig. 4 A feature learning example in a sparse auto-encoding network. (a) A set of features, *i.e.*, basis functions, learned from a set of airport images from remote sensing data. (b) The thresholded responses of the features in (a) on a sample image. [Taken from (Firat et al, 2014) with authors' permission]

terpretation of the scene. Development of vision capabilities is achieved in stages of learning new features and new relations between features that explain the regularities of the environment as well as the *predictions* of the agent (Cohen and Cashon, 2003). In the following, we cover the aspects that should be addressed by such a developing vision system.

5.1 Embodiment, interaction, movement: Active vision

As discussed before, vision is an ill-posed problem by definition. However, ill-posed vision problems become well-posed once interaction with the environment or the object provides more observations (Aloimonos and Rosenfeld, 1994; Aloimonos, 1990). For example, shape from shading, structure from motion or shape from texture become well-posed once the observer is allowed to interact with the environment. The approaches that use this aspect is called *active vision* in the literature and has led to many successful applications - see Aloimonos (2013) for a review.

However, there is more to what acting in an environment can provide, in addition to reducing ambiguities. Through action, our vision system is provided with opportunities to discover new regularities, leading to the development of new features and relations that allow us to predict the consequences of our actions as well as the affordances of the environment (Cohen and Cashon, 2003; Gibson, 2000). In fact, as mentioned by Gibson (2000), one can relate the learning or development of perception with the discovery of the affordances of the environment.

In this direction, we have previously used a robot's interactions with the environment to eliminate irrelevant features in an unsupervised manner and discover object concepts (Atil, 2010) - see Figure 5. In that study, the robot applied its behaviors on the objects in the environment and in an unsupervised fashion, clustered the kind of effects it can generate with these behaviors. Then, by looking at which perceptual property of objects best explains the different effects on the different objects, it could learn the relevant features and from them, build first-level object concepts.

5.2 Relation learning and feature learning

As discussed in Section 4, the Computer Vision community working on feature or representation learning has been making use of the statistical regularities of the environment successfully. The literature has seen performances better than hand-crafted features on state-of-the-art datasets - see (Bengio, 2009; Bengio et al, 2013; Schmidhuber, 2014) for reviews. Unfortunately, such exploitation of the statistical regularities has been limited to only recognition or classification tasks - classification and recognition of objects, scenes, faces, speech etc.

However, there is more to the vision problem than recognition and classification that requires more than a feature-learning system. One important example is the monocular depth perception using cues, such as familiar size, texture gradient, atmospheric effect, occlusion, shading, shadow, etc. Utilization of these cues requires learning relations between features at different levels of representation at different scales: For example, for texture gradient, relations (gradient) between low-level filter responses might be sufficient but, for others, *e.g.*, familiar size and occlusion, relations between object-level representations should be used.

In this regard, Cohen and Cashon (2003) provided an information processing perspective to the development of perception in infants and mentioned that development of features and relations are the key aspects that can explain many findings in Psychology. They propose that, at stages, an infant develops new features, from which new relations are discovered, and the new relations lead to newer features - as a sketch of the basic idea, see Figure 6.

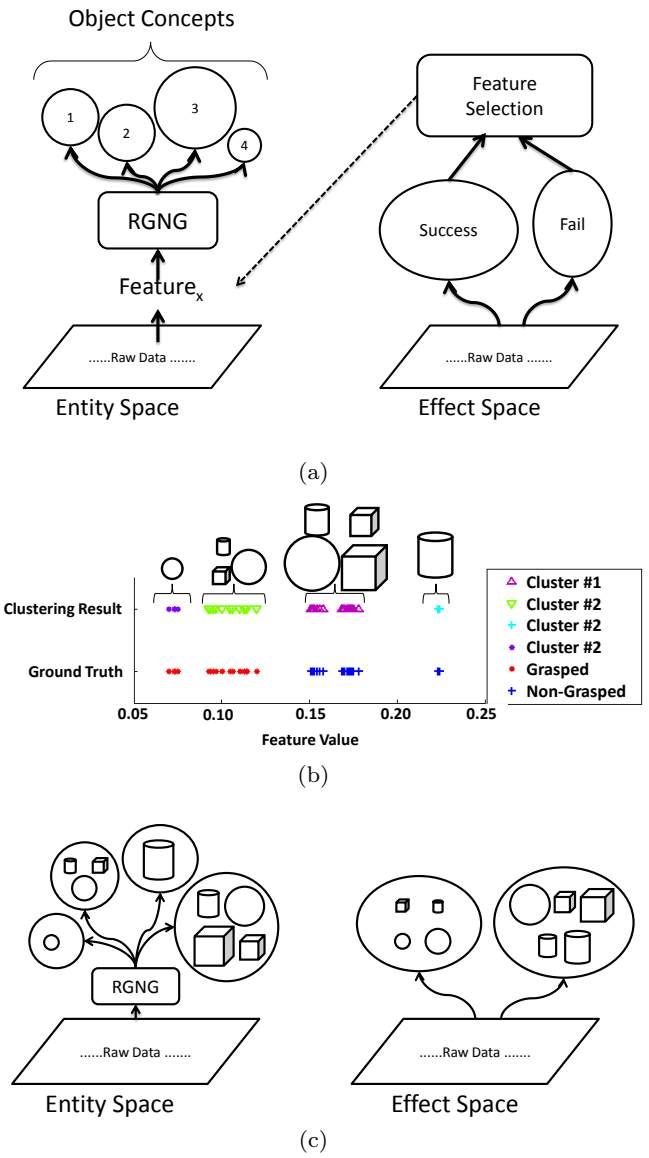

Fig. 5 An unsupervised learning system to discover object concepts (*e.g.*, big-small, round-cornered, heavy-light) through a robot's interaction with the environment. (a) The system is able to select the relevant feature (*i.e.*, the statistical regularity) explaining why different objects cause different effects. For example, the system detects object height feature as the relevant information (regularity) for the grasping behavior. (b) The selected object height feature is successfully able to separate graspable and non-graspable objects. Hence, the object height feature can be used to form concepts of objects (*e.g.*, graspable objects and non-graspable objects for the example of grasping) in an unsupervised manner (using Robust Growing Neural Gas as the clustering algorithm - (Qin and Suganthan, 2004)). (c) By using the selected feature, object appearances are clustered into object concepts of small, medium and big which are all relative (embodied) to the robot. These concepts create a grounding to predict the effects of different behaviors.

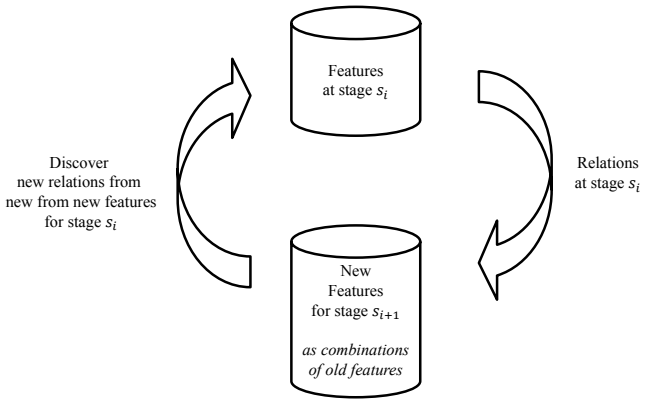

Fig. 6 A developing vision system can be explained by a mechanism which constantly updates features and relations between features.

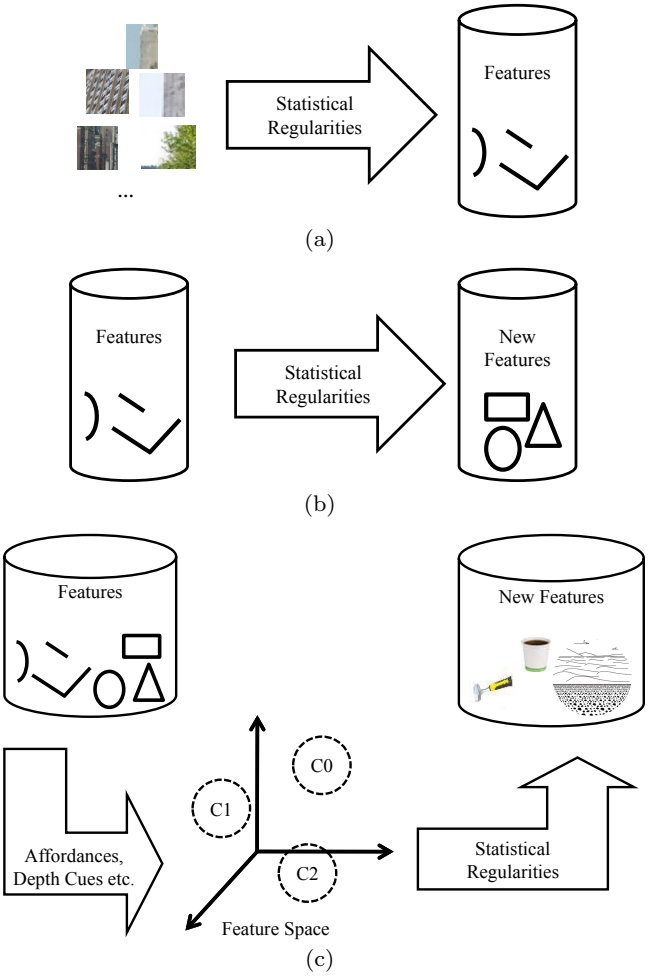

Fig. 7 Using sparsity or slowness principle allows learning simple features from images (a) or complex ones from simple features. Learning features can benefit from interactions with the environment, which provide masking mechanisms to reduce the search space for looking consistencies between features (c).

5.3 Cues that provide regularities

Relations can be discovered as abstractions of consistent spatial and temporal configurations of features or

entities, by looking out for second-order, third-order or even higher-order co-occurrences in many visual aspects, such as orientation, distance, color, etc. Discovered relations might be updated over time with how best they explain future observations or interactions, and as such, spurious relations might be filtered out easily. We can illustrate this in Figure 7(a) and (b), which correspond to what the current feature learning literature has achieved: By constraints or principles of sparsity or slowness, one can discover feature detectors (Figure 7(a)); by repeatedly using such principles, one can learn more complex detectors, even those that can detect objects (Figure 7(b)).

However, checking configurations of many-orders between features amounts to checking consistencies in millions of combinations of features even in a small image. Considering also that discovered relations might be spurious and they need to be updated over time, this combinatorial *explosion* becomes an important issue in relation learning. In the case of an embodied agent, interactions with the environment provide "masking" mechanisms to reduce the size of the combinatorial search. For example, if you touch an object, you would discover that only a subset of low-level features has moved consistently and the rest of the features stayed as they are. As another example, one can consider monocular depth cues: You can walk on a planar surface (floor) and look for the features that give you the walking affordance. Such mechanisms can be used as a masking mechanism to limit the combinatorial search problem - Figure 7(c).

5.4 Principles

The literature has shown that sparsity and slowness have accounted well for development of representations for classification and recognition tasks. An important question for a developing vision system is whether or not such principles are sufficient for a full-fledged developing vision system, especially for its relation-learning aspect. Could there be a unified principle governing the development of a vision system? How can we relate invariance, affordances, sparsity and slowness together in a developing vision system?

6 Conclusion

In this article, we argued that a full-fledged vision system should be embodied and developmental, and it should constantly exploit the regularities of the environment. We propose that such a system can develop itself starting from basic-level vision capabilities and self-extending its capabilities by discovering new features and relations from its interactions with the environment. More importantly, we discussed the main aspects that need to be addressed for such a developing vision system.

The article could be considered as research agenda, providing a direction for the biologically-motivated vision community. As such, it has skipped many technical details and minor aspects that such a developing vision system will need to tackle.

Our motivation comes from developmental robotics and cognitive science studies which have shown that an agent can learn many complex abilities developmentally starting from simpler ones, by exploiting its sensorimotor interactions with the environment.

Acknowledgements We would like to thank Orhan Firat and Fatos Yarman Vural for contributing the figure on the feature learning example.

References

Adolph KE, Eppler MA, Gibson EJ (1993) Crawling versus walking infants' perception of affordances for locomotion over sloping surfaces. Child development 64(4):1158–1174

Aloimonos J (1990) Purposive and qualitative active vision. In: 10th International Conference on Pattern Recognition, IEEE, vol 1, pp 346–360

Aloimonos Y (2013) Active perception. Lawrence Erlbaum Associates, New Jersey

Aloimonos Y, Rosenfeld A (1994) Principles of computer vision. In: Young TY (ed) Handbook of Pattern Recognition and Image Processing: Computer Vision, vol 2, Academic Press, San Diego, USA, chap 1, pp 1–15

Altamura M, Carver FW, Elvevåg B, Weinberger DR, Coppola R (2014) Dynamic cortical involvement in implicit anticipation during statistical learning. Neuroscience letters 558:73–77

Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: A survey. IEEE Transactions on Autonomous Mental Development 1(1):12–34

Atil I (2010) Function and appearance based emergence of object concepts through affordances. Master's thesis, Dept. of Computer Engineering, Middle East Technical University

Barakat BK, Seitz AR, Shams L (2013) The effect of statistical learning on internal stimulus representations: Predictable items are enhanced even when not predicted. Cognition 129(2):205–211

Barlow H (2001) The exploitation of regularities in the environment by the brain. Behavioral and Brain Sciences 24(04):602–607

Bengio Y (2009) Learning deep architectures for ai. Foundations and Trends in Machine Learning 2(1):1–127

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8):1798–1828

Berkes P, Wiskott L (2005) Slow feature analysis yields a rich repertoire of complex cell properties. Journal of Vision 5(6):579–602

Bertero M, Poggio T, Torre V (1987) Ill-posed problems in early vision. Tech. rep., Cambridge, MA, USA

Boden MA (2006) Mind As Machine: A history of Cognitive Science (Vols 1–2). Oxford University Press, Oxford

Brooks RA (1997) From earwigs to humans. Robotics and autonomous systems 20(2):291–304

Brunswik E, Kamiya J (1953) Ecological cue–validity of 'proximity' and of other Gestalt factors. American Journal of Psychologie LXVI:20–32

Cangelosi A, Metta G, Sagerer G, Nolfi S, Nehaniv C, Fischer K, Tani J, Belpaeme T, Sandini G, Nori F, et al (2010) Integration of action and language knowledge: A roadmap for developmental robotics. IEEE Transactions on Autonomous Mental Development 2(3):167–195

Ciresan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 3642–3649

Cohen LB, Cashon CH (2003) Infant perception and cognition. In: Lerner RM, Easterbrooks MA, Mistry J (eds) Handbook of Psychology: Developmental Psychology, vol 6, John Wiley and Sons, Inc., USA, chap 3, pp 65–89

Conway CM, Christiansen MH (2005) Modality-constrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition 31(1):24

Elder J, Goldberg R (2002) Ecological statistics of Gestalt laws for the perceptual organization of contours. Journal of Vision 2(4):324–353

Elder JH, Krupnik A, Johnston LA (2003) Contour grouping with prior models. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(25):1–14

Elman JL (1993) Learning and development in neural networks: the importance of starting small. Cognition 48(1):71 – 99

Ferrell C, Kemp C (1996) An ontogenetic perspective to scaling sensorimotor intelligence. In: Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium, vol 5

Fidler S, Leonardis A (2007) Towards scalable representations of object categories: Learning a hierarchy of parts. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Field D (1994) What is the goal of sensory coding? Neural Computation 6(4):561–601

Field DJ, Hayes A, Hess RF (1993) Contour integration by the human visual system: evidence for a local "association field". Vision Research 33(2):173–193

Fikes RE, Nilsson NJ (1972) Strips: A new approach to the application of theorem proving to problem solving. Artificial intelligence 2(3):189–208

Firat O, Can G, Vural FY (2014) Representation learning for contextual object and region detection in remote sensing. Int Conference on Pattern Recognition (ICPR)

Fiser J, Aslin RN (2002) Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences 99(24):15,822–15,826

Fodor JA (1981) Representations: Philosophical Essays on the Foundations of Cognitive Science. Mit Press

Franzius M, Wilbert N, Wiskott L (2008) Invariant object recognition with slow feature analysis. In: Proceedings of the 18th international conference on Artificial Neural Networks, Part I, Springer-Verlag, pp 961–970

Fry AF, Hale S (1996) Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. Psychological science 7(4):237–241

Garrett HE (1946) A developmental theory of intelligence. American Psychologist 1(9):372–378

Geisler W, Perry J, Super B, Gallogly D (2001) Edge co–occurrence in natural images predicts contour grouping performance. Vision Research 41(6):711–724

Gibson EJ (1969) Principles of perceptual learning and development. Appleton-Century-Crofts

Gibson EJ (2000) Perceptual learning in development: Some basic concepts. Ecological Psychology 12(4):295–302

Gibson EJ, Walk RD (1960) The "visual cliff"

Gibson JJ (1977) The theory of affordances. In: Shaw R, Bransford J (eds) Perceiving, acting, and knowing: Toward an ecological Psychology, John Wiley & Sons Inc, New Jersey, pp 67–82

Gibson JJ, Gibson EJ (1955) Perceptual learning: Differentiation or enrichment? Psychological review 62(1):32–41

Glenberg AM, Robertson DA (2000) Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of Memory and Language 43(3):379 – 401

Gopnik A, Schulz L (2004) Mechanisms of theory formation in young children. Trends in cognitive sciences 8(8):371–377

Hadamard J (1923) Lectures on the Cauchy Problem in Linear Partial Differential Equations. Yale, New Haven

Harnad S (1990) The symbol grounding problem. Physica D: Nonlinear Phenomena 42(1):335–346

Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural computation 18(7):1527–1554

Howe CQ, Purves D (2002) Range image statistics can explain the anomalous perception of length. PNAS 99(20):13,184–13,188

Howe CQ, Purves D (2004) Size contrast and assimilation explained by the statistics of natural scene geometry. Journal of Cognitive Neuroscience 16(1):90–102

Huang J, Lee AB, Mumford D (2000) Statistics of range images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1(1):1324–1331

Hyvärinen A, Hoyer PO (2001) A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. Vision research 41(18):2413–2423

Hyvärinen A, Hurri J, Hoyer PO (2009) Natural Image Statistics: A Probabilistic Approach to Early Computational Vision., vol 39. Springer

Kalkan S (2008) Multi-modal statistics of local image structures and its applications for depth prediction. PhD thesis, Dept. of Informatics, University of Goettingen, Germany

Kalkan S, Wörgötter F, Krüger N (2006) Statistical analysis of local 3d structure in 2d images. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) pp 1114–1121

Kalkan S, Krüger N, Wörgötter F (2007) First-order and second-order statistical analysis of 3d and 2d structure. Network: Computation in Neural Systems 18(2):129–160

Kalkan S, Wörgötter F, Krüger N (2008) Depth prediction at homogeneous image structures. int. conf. on computer vision theory and applications. International Conference on Computer Vision Theory and Applications (VISAPP)

Kellman P, Arterberry M (eds) (1998) The Cradle of Knowledge. MIT-Press

Kirkham NZ, Slemmer JA, Johnson SP (2002) Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition 83(2):B35 – B42

Knill DC, Richards W (eds) (1996) Perception as bayesian inference. Cambridge: Cambridge University Press

König P, Krüger N (2006) Symbols as self-emergent entities in an optimization process of feature extraction and predictions. Biological Cybernetics 94(4):325–334

Kraft D, Detry R, Pugeault N, Baseski E, Guerin F, Piater JH, Kruger N (2010) Development of object and grasping knowledge by robot exploration. Autonomous Mental Development, IEEE Transactions on 2(4):368–383

Krüger N (1998) Collinearity and parallelism are statistically significant second order relations of complex cell responses. Neural Processing Letters 8(2):117–129

Krüger N, Wörgötter F (2002) Multi modal estimation of collinearity and parallelism in natural image sequences. Network: Computation in Neural Systems 13(4):553–576

Krüger N, Wörgötter F (2004) Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. Advances in Imaging and Electron Physics 131:82–147

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connection Science 15(4):151–190

Minsky ML (1963) Steps towards artificial intelligence. In: Feigenbaum E, Feldman J (eds) Computers and Thought, McGraw-Hill, New York, pp 406–450

Newell A, Simon HA (1976) Computer science as empirical inquiry: Symbols and search. Communications of the ACM 19(3):113–126

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision research 37(23):3311–3325

Overton WF (2003) Development across the life span. Handbook of psychology: One pp 11–42

Potetz B, Lee TS (2003) Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. Journal of the Optical Society of America 20(7):1292–1303

Pugeault N, Krüger N, Wörgötter F (2004) A non-local stereo similarity based on collinear groups. Proceedings of the Fourth International ICSC Symposium on Engineering of Intelligent Systems

Purves D, Lotto B (eds) (2002) Why we see what we do: an empirical theory of vision. Sunderland, MA: Sinauer Associates

Qin AK, Suganthan PN (2004) Robust growing neural gas algorithm with application in cluster analysis. Neural Networks 17(8):1135–1148

Rao RPN, Olshausen BA, Lewicki MS (eds) (2002) Probabilistic models of the brain. MA: MIT Press

Rifai S, Bengio Y, Courville A, Vincent P, Mirza M (2012) Disentangling factors of variation for facial expression recognition. In: European Conference on Computer Vision, Springer, pp 808–822

Rocha A (1997) The brain as a symbol-processing machine. Progress in Neurobiology 53(2):121 – 198

Saffran JR (2003) Statistical language learning mechanisms and constraints. Current directions in psychological science 12(4):110–114

Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. Science 274(5294):1926–1928

Salakhutdinov R, Hinton GE (2009) Deep boltzmann machines. In: International Conference on Artificial Intelligence and Statistics, pp 448–455

Samuel AL (2000) Some studies in machine learning using the game of checkers. IBM Journal of research and development 44(1.2):206–226

Sandini G (1997) Artificial systems and neuroscience. In: Proc. of the Otto and Martha Fischbeck Seminar on Active Vision

Sarkar S, Boyer KL (1993) Perceptual organization in computer vision: A review and a proposal for a classificatory structure. IEEE Transactions on Systems, Man and Cybernetics 23(2):382–399

Scalzo F, Piater JH (2005) Statistical learning of visual feature hierarchies. In: IEEE Workshop on Learning in Computer Vision and Pattern Recognition, vol 3, pp 44–44

Schmidhuber J (2014) Deep learning in neural networks: An overview. CoRR abs/1404.7828

Selfridge OG (1958) Pandemonium: a paradigm for learning in mechanisation of thought processes. In Proceedings of a Symposium Held at the National Physical Laboratory pp 513–526

Simoncelli E, Ohlshausen B (2001) Natural image statistics and neural representations. Anual Reviews of Neuroscience 24:1193–1216

Simoncelli EP (2003) Vision and the statistics of the visual environment. Current Opinion in Neurobiology 13(2):144–149

Smith L, Yu C (2008) Infants rapidly learn word-referent mappings via cross-situational statistics. Cognition 106(3):1558–1568

Spelke E (1990) Principles of object perception. Cognitive Science 14(1):29–56

Thelen E, Smith L (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge, MA

Tomasello M (2009) Constructing a language: A usage-based theory of language acquisition. Harvard University Press

Turk-Browne NB, Scholl BJ, Chun MM, Johnson MK (2009) Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. Journal of Cognitive Neuroscience 21(10):1934–1945

Turk-Browne NB, Scholl BJ, Johnson MK, Chun MM (2010) Implicit perceptual anticipation triggered by statistical learning. The Journal of Neuroscience 30(33):11,177–11,187

Vapnik V (2000) The nature of statistical learning theory. springer

Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R (2012) A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. Psychological bulletin 138(6):1172

Yang Z, Purves D (2003) Image/source statistics of surfaces in natural scenes. Network: Computation in Neural Systems 14(3):371–390

Zhu SC (1999) Embedding gestalt laws in markov random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(11):1170–1187