What's Wrong with Computer Vision?

Marco Gori(✉)

Department of Information Engineering and Mathematics,
University of Siena, Siena, Italy
marcoxgori@gmail.com
http://sailab.diism.unisi.it/people/marco-gori/

Abstract. By and large, the remarkable progress in visual object recognition in the last few years is attributed to the availability of huge labelled data paired with strong and suitable computational resources. This has opened the doors to the massive use of deep learning which has led to remarkable improvements on common benchmarks. While subscribing this view, in this paper we claim that the time has come to begin working towards a deeper understanding of visual computational processes, that instead of being regarded as applications of general purpose machine learning algorithms, are likely to require appropriate learning schemes. The major claim is that while facing nowadays object recognition problems we have been working a problem that is significantly more difficult than the one offered by nature. This is due to learning algorithms that are working on images while neglecting the crucial role of frame temporal coherence. We address the limitations and discuss how the evolution of the tradition of image recognition towards visual recognition might give rise to remarkable advances in the field of computer vision.

Keywords: Computer vision · Object recognition · Machine learning
Motion invariance

1 Introduction

While the emphasis on a general theory of vision was already the main objective at the dawn of the discipline [13], it has evolved without a systematic exploration of foundations in machine learning. When the target is moved to unrestricted visual environments and the emphasis is shifted from huge labelled databases to a human-like protocol of interaction, we need to go beyond the current peaceful interlude that we are experimenting in vision and machine learning. A fundamental question a good theory is expected to answer is why children can learn to recognize objects and actions from a few supervised examples, whereas nowadays supervised learning approaches strive to achieve this task. In particular, why are they so thirsty for supervised examples? Interestingly, this fundamental difference seems to be deeply rooted in the different communication protocol at the basis of the acquisition of visual skills in children and machines.

L. Pancioni et al. (Eds.): ANNPR 2018, LNAI 11081, pp. 3–16, 2018.
https://doi.org/10.1007/978-3-319-99978-4_1

So far, the semantic labeling of pixels of a given video stream has been mostly carried out at frame level. This seems to be the natural outcome of well-established pattern recognition methods working on images, which have given rise to nowadays emphasis on collecting big labelled image databases (e.g. [4]) with the purpose of devising and testing challenging machine learning algorithms. While this framework is the one in which most of nowadays state of the art object recognition approaches have been developing, we argue that there are strong arguments to start exploring the more natural visual interaction that animals experiment in their own environment.

This suggests to process video instead of image collection, that naturally leads to a paradigm-shift in the associated processes of learning to see. The idea of shifting to video is very much related to the growing interest of *learning in the wild* that has been explored in the last few years[1]. The learning processes that take place in this kind of environments has a different nature with respect to those that are typically considered in machine learning. Learning convolutional nets on ImageNet typically consists of updating the weights from the processing of temporally unrelated images, whereas a video carries out information where we pass from one frame to the previous one by smooth changes. While ImageNet is a collection of unrelated images, a video supports information only when motion is involved. In presence of fixed images that last for awhile, the corresponding stream of equal frames basically supports only the information of a single image. As a consequence, it is clear that visual environments diffuse information only when motion is involved. There is no transition from one image to the next one—like in ImageNet—but, as time goes by, the information is only carried out by motion, which modifies one frame to the next one according to the optical flow. Once we deeply capture this fundamental features of visual environment, we early realize that we need a different theory of machine learning that naturally processes streams that cannot be regarded just as collection of independent images anymore.

A crucial problem that was recognized by Poggio and Anselmi [15] is the need to incorporate visual invariances into deep nets that go beyond simple translation invariance that is currently characterizing convolutional networks. They propose an elegant mathematical framework on visual invariance and enlightened some intriguing neurobiological connections. Overall, the ambition of extracting distinctive features from vision poses a challenging task. While we are typically concerned with feature extraction that is independent of classic geometric transformation, it looks like we are still missing the fantastic human skill of capturing distinctive features to recognize ironed and rumpled shirts! There is no apparent difficulty to recognize shirts by keeping the recognition coherence in case we roll up the sleeves, or we simply curl them up into a ball for the laundry basket. Of course, there are neither rigid transformations, like translations and rotation, nor scale maps that transforms an ironed shirt into the same shirt thrown into the laundry basket. Is there any natural invariance?

[1] See. e.g. https://sites.google.com/site/wildml2017icml/.

In this paper, we claim that motion invariance is in fact the only one that we need. Translation and scale invariance, that have been the subject of many studies, are in fact examples of invariances that can be fully gained whenever we develop the ability to detect features that are invariant under motion. If my inch moves closer and closer to my eyes then any of its representing features that is motion invariant will also be scale invariant. The finger will become bigger and bigger as it approaches my face, but it is still my inch! Clearly, translation, rotation, and complex deformation invariances derive from motion invariance. Humans life always experiments motion, so as the gained visual invariances naturally arise from motion invariance. Animals with foveal eyes also move quickly the focus of attention when looking at fixed objects, which means that they continually experiment motion. Hence, also in case of fixed images, conjugate, vergence, saccadic, smooth pursuit, and vestibulo-ocular movements lead to acquire visual information from relative motion. We claim that the production of such a continuous visual stream naturally drives feature extraction, since the corresponding convolutional filters are expected not to change during motion. The enforcement of this consistency condition creates a mine of visual data during animal life. Interestingly, the same can happen for machines. Of course, we need to compute the optical flow at pixel level so as to enforce the consistency of all the extracted features. Early studies on this problem [8], along with recent related improvements (see e.g. [2]) suggests to determine the velocity field by enforcing brightness invariance. As the optical flow is gained, it is used to enforce motion consistency on the visual features. Interestingly, the theory we propose is quite related to the variational approach that is used to determine the optical flow in [8]. It is worth mentioning that an effective visual system must also develop features that do not follow motion invariance. These kind of features can be conveniently combined with those that are discussed in this paper with the purpose of carrying out high level visual tasks. Early studies driven by these ideas are reported in [6], where the authors propose the extraction of visual features as a constraint satisfaction problem, mostly based on information-based principles and early ideas on motion invariance.

In this paper we mostly deal with an in-depth discussion of the principles that one should follow to construct a sound theory of vision that, later on, can likely be also applied to computer vision. In addition, we discuss some of the reasons of the limitations of current approaches, where perceptual and linguistic tasks interwound with vision are not properly covered. This issue is enlighten by proposing a hierarchy of cognitive tasks connected to vision that contributes to shed light on the intriguing connection between gaining perceptual and linguistic skills. The discussion suggests that most problems of computer vision are likely to be posed according to the historical evolution of the applications more than on a formal analysis of the underlying computational processes. While this choice has been proven to be successful in many real-world cases, stressing this research guideline might lead, on the long run, to wrong directions.

2 Top Ten Questions a Theory on Vision Should Address

The extraction of informative and robust cues from visual scenes has been attracting more and more interest in computer vision. For many years, scientists and engineers have contributed to the construction of solutions to extract visual features, that are mostly based on clever heuristics (see e.g. [12]). However, the remarkable achievements of the last few years have been mostly based on the accumulation of huge visual collections enriched by crowdsourcing. It has created labels to carry out massive supervised learning in deep convolutional networks, that has given rise to very effective internal representations of visual features. The have been successfully used in an impressive number of application (see e.g. [10,17]).

In this paper, we argue that while stressing this issue we have been facing artificial problems that, from a pure computational point of view, are likely to be significantly more complex than natural visual tasks that are daily faced by animals. In humans, the emergence of cognition from visual environments is interwound with language. This often leads to attack the interplay between visual and linguistic skills by simple models that, like for supervised learning, strongly rely on linguistic attachment. However, when observing the spectacular skills of the eagle that catches the pray, one promptly realizes that for an in-depth understanding of vision, that likely yields also an impact in computer implementation, one should begin with a neat separation with language! This paper is mostly motivated by the curiosity of addressing a number of questions that arise when looking at natural visual processes [3]. While they come from natural observation, they are mostly regarded as general issues strongly rooted in information-based principles, that we conjecture are of primary importance also in computer vision.

*Q*1 *How can animals conquer visual skills without requiring "intensive supervision"?*
Recent remarkable achievements in computer vision are mostly based on tons of supervised examples—of the order of millions! This does not explain how can animals conquer visual skills with scarse "supervision" from the environment. Hence, there is plenty of evidence and motivations for invoking a theory of truly unsupervised learning capable of explaining the process of extraction of features from visual data collections. While the need for theories of unsupervised learning in computer vision has been advocated in a number of papers (see e.g. [7,11,16,19]), so far, the powerful representations that arise from supervised learning, because of many recent successful applications, seem to attract much more interest. While information-based principles could themselves suffice to construct visual features, the absence of any feedback from the environment make those methods quite limited with respect to supervised learning. Interestingly, the claim of this paper is that motion invariance offers a huge amount of free supervisions from the visual environment, thus explaining the reason why humans do not need

the massive supervision process that is dominating feature extraction in convolutional neural networks.

*Q*2 *How can animals gradually conquer visual skills in a visual environments?* Animals, including primates, not only receive a scarse supervision, but they also conquer visual skills by living in their own visual environment. This is gradually achieved without needing to separate learning from test environments. At any stage of their evolution, it looks like they acquire the skills that are required to face the current tasks. On the opposite, most approaches to computer vision do not really grasp the notion of time. The typical ideas behind on-line learning do not necessarily capture the natural temporal structure of the visual tasks. Time plays a crucial role in any cognitive process. One might believe that this is restricted to human life, but more careful analyses lead us to conclude that the temporal dimension plays a crucial role in the well-positioning of most challenging cognitive tasks, regardless of whether they are faced by humans or machines. Interestingly, while many people struggle for the acquisition of huge labeled databases, the truly incorporation of time leads to a paradigm shift in the interpretation of the learning and test environment. In a sense, such a distinction ceases to apply, and we can regard unrestricted visual collections as the information accumulated during all the agent life, that can likely surpass any attempt to collect image collection. The theory proposed in this paper is framed in the context of agent life characterized by the ordinary notion of time, which emerges in all its facets. We are not concerned with huge visual data repositories, but merely with the agent life in its own visual environments.

*Q*3 *Can animals see in a world of shuffled frames?*
One might figure out what human life could have been in a world of visual information with shuffled frames. Could children really acquire visual skills in such an artificial world, which is the one we are presenting to machines? Notice that in a world of shuffled frames, a video requires order of magnitude more information for its storing than the corresponding temporally coherent visual stream. This is a serious warning that is typically neglected; any recognition process is remarkably more difficult when shuffling frames, which clearly indicates the importance of keeping the spatiotemporal structure that is offered by nature. This calls for the formulation of a new theory of learning capable of capturing spatiotemporal structures. Basically, we need to abandon the safe model of restricting computer vision to the processing of images. The reason for formulating a theory of learning on video instead of on images is not only rooted in the curiosity of grasping the computational mechanisms that take place in nature. It looks like that, while ignoring the crucial role of temporal coherence, the formulation of most of nowadays current computer vision tasks leads to tackle a problem that is remarkably more difficult than the one nature has prepared for humans! We conjecture that animals could not see in a world of shuffled frames, which indicates that such an artificial formulation might led to a very hard problem. In a sense, the very good results that we already can experiment

nowadays are quite surprising, but they are mostly due to the stress of the computational power. The theory proposed in this paper relies of the choice of capturing temporal structures in natural visual environments, which is claimed to simplify dramatically the problem at hand, and to give rise to lighter computation.

$Q4$ *How can humans attach semantic labels at pixel level?*
Humans provide scene interpretation thanks to linguistic descriptions. This requires a deep integration of visual and linguistic skills, that are required to come up with compact, yet effective visual descriptions. However, amongst these high level visual skills, it is worth mentioning that humans can attach semantic labels to a single pixel in the retina. While this decision process is inherently interwound with a certain degree of ambiguity, it is remarkably effective. The linguistic attributes that are extracted are related to the context of the pixel that is taken into account for label attachment, while the ambiguity is mostly a linguistic more than a visual issue. The theory proposed in this paper addresses directly this visual skill since the labels are extracted for a given pixel at different levels of abstraction. Unlike classic convolutional networks, there is no pooling; the connection between the single pixels and their corresponding features is kept also when the extracted features involve high degree of abstraction, that is due to the processing over large contexts. The focus on single pixels allows us to go beyond object segmentation based sliding windows, which somewhat reverses the pooling process. Instead of dealing with object proposals [21], we focus on the attachment of symbols at single pixels in the retina. The bottom line is that human-like linguistic descriptions of visual scenes is gained on top of pixel-based feature descriptions that, as a byproduct, must allow us to perform semantic labeling. Interestingly, there is more; as it will be shown in the following, there are in fact computational issues that lead us to promote the idea of carrying our the feature extraction process while focussing attention on salient pixels.

$Q5$ *Why are there two mainstream different systems in the visual cortex (ventral and dorsal mainstream)?*
It has been pointed out that the visual cortex of humans and other primates is composed of two main information pathways that are referred to as the ventral stream and dorsal stream [5]. The traditional distinction distinguishes the ventral "what" and the dorsal "where/how" visual pathways, so as the ventral stream is devoted to perceptual analysis of the visual input, such as object recognition, whereas the dorsal stream is concerned with providing motion ability in the interaction with the environment. The enforcement of motion invariance is clearly conceived for extracting features that are useful for object recognition to assolve the "what" task. Of course, neurons with built-in motion invariance are not adeguate to make spatial estimations. A good model for learning of the convolutional need to access to velocity estimation, which is consistent with neuroanatomical evidence.

$Q6$ *Why is the ventral mainstream organized according to a hierarchical architecture with receptive fields?*

Beginning from early studies by Hubel and Wiesel [9], neuroscientists have gradually gained evidence of that the visual cortex presents a hierarchical structure and that the neurons process the visual information on the basis of inputs restricted to receptive field. Is there a reason why this solution has been developed? We can promptly realize that, even though the neurons are restricted to compute over receptive fields, deep structures easily conquer the possibility of taking large contexts into account for their decision. Is this biological solution driven by computational laws of vision? In [3], the authors provide evidence of the fact that receptive fields do favor the acquisition of motion invariance which, as already stated, is the fundamental invariance of vision. Since hierarchical architectures is the natural solution for developing more abstract representations by using receptive fields, it turns out that motion invariance is in fact at the basis of the biological structure of the visual cortex. The computation at different layers yields features with progressive degree of abstraction, so as higher computational processes are expected to use all the information extracted in the layers.

$Q7$ *Why do animals focus attention?*
The retina of animals with well-developed visual system is organized in such a way that there are very high resolution receptors in a restricted area, whereas lower resolution receptors are present in the rest of the retina. Why is this convenient? One can easily argue that any action typically takes place in a relatively small zone in front of the animals, which suggests that the evolution has led to develop high resolution in a limited portion of the retina. On the other hand, this leads to the detriment of the peripheral vision, that is also very important. In addition, this could apply for the dorsal system whose neurons are expected to provide information that is useful to support movement and actions in the visual environment. The ventral mainstream, with neurons involved in the "what" function does not seem to benefit from foveal eyes. From the theory proposed in this paper, the need of foveal retinas is strongly supported for achieving efficient computation for the construction of visual features. However, it will be argued that the most important reason for focussing attention is that of dramatically simplifying the computation and limit the ambiguities that come from the need to sustaining a parallel computation over each frame.

$Q8$ *Why do foveal animals perform eye movements?*
Human eyes make jerky saccadic movements during ordinary visual acquisition. One reason for these movements is that the fovea provides high-resolution in portions of about $1, 2$ degrees. Because of such a small high resolution portions, the overall sensing of a scene does require intensive movements of the fovea. Hence, the foveal movements do represent a good alternative to eyes with uniformly high resolution retina. On the other hand, the preference of the solution of foveal eyes with saccadic movements is arguable, since while a uniformly high resolution retina is more complex to achieve than foveal retina, saccadic movements are less important. The information-based theory presented in this paper makes it possible to conclude that foveal retina with saccadic movements is in fact a solution that is computationally sustainable and very effective.

$Q9$ *Why does it take 8–12 months for newborns to achieve adult visual acuity?* There are surprising results that come from developmental psychology on what a newborn see. Charles Darwin came up with the following remark:

> It was surprising how slowly he acquired the power of following with his eyes an object if swinging at all rapidly; for he could not do this well when seven and a half months old.

At the end of the seventies, this early remark was given a technically sound basis [20]. In the paper, three techniques,—optokinetic nystagmus (OKN), preferential looking (PL), and the visually evoked potential (VEP)—were used to assess visual acuity in infants between birth and 6 months of age. More recently, the survey by Braddick and Atkinson [14] provides an in-depth discussion on the state of the art in the field. It is clearly stated that for newborns to gain adult visual acuity, depending on the specific visual test, several months are required. Is the development of adult visual acuity a biological issue or does it come from higher level computational laws?

$Q10$ *Causality and Non Rapid Eye Movements (NREM) sleep phases* Computer vision is mostly based on huge training sets of images, whereas humans use video streams for learning visual skills. Notice that because of the alternation of the biological rhythm of sleep, humans somewhat process collections of visual streams pasted with relaxing segments composed of "null" video signal. This happens mostly during NREM phases of sleep, in which also eye movements and connection with visual memory are nearly absent. Interestingly, the Rapid Eye Movements (REM) phase is, on the opposite, similar to ordinary visual processing, the only difference being that the construction of visual features during the dream is based on the visual internal memory representations [18]. As a matter of fact, the process of learning the filters experiments an alternation of visual information with the reset of the signal. A good theory of learning visual features should provide evidence to claim that such a relaxation coming from the reset of the signal nicely fits the purpose of optimizing an overall optimization index based on the previously stated principles. In particular, in [3], the authors point out that periodic resetting of the visual information favors the optimization under causality requirements. Hence, the role of eye movement and of sleep seem to be important for the optimal development of visual features.

3 Hierarchical Description of Visual Tasks

In this section we discuss visual tasks and their intriguing connection with language. This analysis is motivated by the evidence provided in nature of excellent visual skills that arise regardless of language. At the light of the following analysis, one should consider to start go beyond the tradition of computer vision of emphasizing classification tasks. Visual perception drives different functional tasks in animals, so as the human intersection with language must properly be analyzed.

Let $\mathscr{T} = [t_0, t_1]$ be the *temporal domain* and let $\mathscr{X} \subset \mathbb{R}^2$ be the retina. We consider the video domain $\mathscr{D} := \mathscr{T} \times \mathscr{X}$ so as

$$v : \mathscr{D} \to \mathbb{R}^d : \ (t, x) \to [v_1(t, x), \ldots, v_d(t, x)]'$$

is the video signal on $\mathscr{D}$. In the classic case of RGB coding, we have $d = 3$. Throughout the paper, $v(\mathscr{D})$ denotes any video, while we use $\mathscr{V}$ to denote the universal set of videos, where any video belongs to. Likewise, $v(t, \mathscr{X})$ denotes the frame at t and $\mathscr{I}$ denotes the universal set of images with values $v(t, x) \in \mathbb{R}^d$. Clearly, we have $v(\mathscr{D}) \in \mathscr{V}$. Now, humans are capable of providing sophisticated linguistic representations from video $v(\mathscr{D}) \in \mathscr{V}$, which involve both local and global features. Clearly, abstract descriptions of a visual scene do require considerable linguistic skills, which emerge also at local level when specific words can also be attached to any pixel of a given visual frame. Basically, humans are capable of providing a linguistic description of $v(\mathscr{D})$ that goes well beyond object classification. The amount of visual information is typically so huge that for an appropriate cognitive transcription at linguistic level to take place one cannot rely on classification, but must necessarily involve the compositional structure of language. This kind of difficulty clearly emerges when trying to provide a linguistic description to blind people, a task which is quite difficult also for humans.

3.1 Pixel-Wise and Abstract Visual Interpretations

One of the most challenging issues in vision is human ability to jump easily from pixel-wise to recognition processes and more abstract visual interpretations that involve frames as well as portions of a video. When focussing attention on a certain pixel in a picture, humans can easily make a list of "consistent objects" that reflects the visual information around that pixel. Interestingly, that process takes place by automatically adapting a sort of "virtual window" used for the decision. This results in the typical detection of objects with dimension which is growing as that virtual window gets larger and larger. More structured objects detected at a given pixel are clearly described by more categories than simple primitive objects, but, for humans, the resulting *pixel-wise* process is surprisingly well-posed from a pure cognitive point of view. However, such a pixel-wise process seems to emerge upon request; apparently, humans do not carry out such a massive computation over all the retina. In addition, there are abstract visual skills that are unlikely to be attacked by pixel-wise computation. Humans provide visual interpretations that goes beyond the truly visual pattern (see e.g. Kanizsa's illusions). This happens because of the focus of attention, which somehow locates the object to be processed. As the focus is on the pixel $f(t)$, the corresponding object can be given an abstract geometrical interpretation by its shape expressed in term of its contour. While pixel-based processes are based on all the visual information of the retina associated with a given pixel, shape-based recognition emerges when recognizing objects on the basis of their contour, once we focus attention of a point of the object.

Pixel-wise processes can only lead to the emergence of decisions on objects, which is fact a static concept. It cannot allow us to draw conclusions on actions, whose understanding does require to involve portions of video. However, like for objects, the detection of the "contour of actions" yields a very useful abstraction. The notion *object affordance* has a strict connection with that of action. We carry out many object recognition processes on the basis of actions in which they are involved, so as objects are detected because of their role in the scene. In other words, the affordance involves the *functional role* of objects, which is used for the emergence of abstract categories.

3.2 The Interwound Story of Vision and Language

In the previous section, we have discussed pixel-wise versus abstract computational processes aimed at generating labels to be attached to objects and actions. We can think of two different alphabets Σ_p and Σ_s which refer to words related to *pixel-wise and shape-based recognition processes*, respectively. For instance, while terms like `eye`, `mouth`, and `face` are typical elements of Σ_p, their geometrical description is based on terms in Σ_s. So we say that the `face` has an `oval` shape, where `oval` is a typical elements of Σ_s.

Overall, a visual agent performs cognitive tasks by working on $\Sigma_a = \Sigma_p \vee \Sigma_s$. It is important to point out that Σ_a is only the alphabet of *primitive terms*, since when dealing with structured objects and actions, visual agents play with concepts described by additional terms

Basically, the extraction of semantics from video requires linguistic descriptions, even at local level, where one is asked to select words from the alphabet $\omega \in \Sigma_s$. Here we regard any word ω as a symbol with attached semantics, like in the case of any natural language.

The most abstract task that humans are capable to face is that of constructing a function χ_0 as follows

$$\chi_0 : \mathscr{V} \to \mathcal{L}_0 : v(\mathscr{D}) \to \chi_0(v(\mathscr{D})), \tag{1}$$

where $\mathcal{L}_0 \subset \Sigma_s^\star$ is a type zero language in Chomsky's hierarchy. This embraces any linguistic report from visual scenes, like, for instance, movie review. In addition to the ability of extracting information from visual sources, a remarkable specific problem in the construction of χ_0 is that of properly handing the temporal flow of the frames and to provide a semantic representation of the movie actions. Clearly, a movie review does not only require the ability of extracting a visual representation, but also to properly understand the actions so as to produce a corresponding descriptions. While cats and eagles are commonly regarded as animals with very good visual skills, they cannot produce movie reports. Basically, the sentence $\chi_0(v(\mathscr{D})) \in \Sigma_s^\star$ is expected to be taken from a language $\mathcal{L}_0$ of highest level in Chomsky classification, which is denoted by $\mathcal{L}_0$.

Another fundamental visual task is that of *query answer*, that can be regarded as

$$\chi_0 : \mathscr{V} \times \mathcal{L}_0 \to \mathcal{L}_0 : v(\mathscr{D}) \to \chi_0(v(\mathscr{D})), \tag{2}$$

Table 1. Hierarchical structure of semantic labeling.

input $\leadsto$ **semantic description**	**Remarks**
$\chi_0(v(\mathscr{D})) \leadsto \mathcal{L}_0$	The language involves ordinary human scene descriptions. Spatial and knowledge levels are both involved.
$\chi_1(v(t,\mathscr{X})) \leadsto \mathcal{L}_0$	The language involves ordinary human picture descriptions. Spatial knowledge is only involved.
$\chi_2(t,x,v(t,\mathscr{X})) \leadsto \mathcal{L}_{-1}$	The language consists of a list of words (language degeneration, no ordering), that is $\mathcal{L}_{-1} \subset \Sigma_s^\star$ only.
$\cdot\chi_3(t,x,v(t,\mathscr{X})) \leadsto \mathcal{L}_{-2}$	The language consists of a vector of words (language degeneration, no order). Unlike $\mathcal{L}_{-1}$, the number of symbols is known in advance.

A simplified version and more realistic formulation of semantic labeling, when actions are not the target, is the one in which

$$\chi_1 : \mathscr{I} \to \mathcal{L}_0 : (v(t,\mathscr{X})) \to \chi_1(v(t,\mathscr{X})). \tag{3}$$

This tasks still requires $\mathcal{L}_0$ for linguistic description, but only spatial knowledge $\mathcal{K}_s$ is needed, since, unlike the previous case, there is no temporal processing required (Table 1).

A dramatic drop of complexity arises when asking the agent to provide visual skills on $v(t,\mathscr{X})$ while focussing attention to (t,x). This is described by

$$\chi_2 : \mathscr{D} \times \mathscr{I} \to \Sigma_s : (t,x,v(t,\mathscr{X})) \to \chi_2(t,x,v(t,\mathscr{D})), \tag{4}$$

Basically, while the decision is based on $u(t,x) = (t,x,v(t,\mathscr{X})) \in \mathscr{U}$, which represents quite an unusual granule of information with respect to what is typically processed in machine learning and pattern recognition, this time there is no linguistic description, since we only expected the agent to return a list of symbols of Σ_s. This simplifies dramatically the overall problem, thus contributing to decoupling visual and semantic processes. It is worth mentioning that the dramatic reduction of complexity in the semantic processes is paired with the emergence of focus of attention, namely with decisions based on $u(t,x) \in \mathscr{U}$. In principle, one can expect semantic labeling of (t,x) by means of a single $\omega \in \Sigma_s$, but in some cases dozens of words might be associated with $u(t,x)$. While the linguistic structure degenerates, we are still in presence of a compositional structure, so as the agent might generate remarkable lengthy sentences of pertinent words $\Sigma_s^\star$.

3.3 When Vision Collapses to Classification

An additional simplification on the semantic level arises when considering that the process of generating the words $\omega \in \Sigma^\star$ can be thought of as a compositional process based on a set Σ_d of "dummy symbols", so as

$$\chi_3 : \mathscr{D} \times \mathscr{I} \to \Sigma_d : (t,x,v(t,\mathscr{X})) \to \chi_3(t,x,v(t,\mathscr{D})), \tag{5}$$

Basically, the transition from $\chi_2(\cdot)$ to $\chi_3(\cdot)$ involves a further definitive linguistic simplification, which restricts the symbolic description from $\Sigma_s^\star$ to Σ_d. In so doing, all the complexity is now on the visual side, which requires decisions based on $u(t,x)$, so as we are finally in front of a *classification problem.* This description of visual tasks makes it clear that in order to conquer abstract computer vision skills, any agent does require to address both issues of input representation and linguistic descriptions. We claim that any systematic approach to vision cannot avoid to face the issue of decoupling the classification of visual features, with symbols in Σ_d, and the appropriate linguistic description.

Let us analyze the problems connected with the construction of χ_0 and χ_1, which both operate on global input representation, thus disregarding any focus of attention mechanism. The complexity can be promptly appreciated also in the simplest task χ_1. Clearly, it cannot be regarded as a classification since the agent is expected to provide truly linguistic descriptions. On top of that, when dealing with unrestricted visual environments, the interpretation of $v(t,\mathscr{X})$ is trapped into the chicken-egg dilemma on whether classification of objects or segmentation must take place first. This is due to the absence of any focus of attention, which necessarily leads to holistic mechanisms of information extraction. Unfortunately, while holistic mechanisms are required at a certain level of abstraction, the direct process of $v(t,\mathscr{X})$ do not offer the right source for their activation. Basically, there is no decoupling between the visual source and its linguistic description.

Interestingly, this decoupling takes place when separating $\chi_3(\cdot)$ with respect to the others. The development of abstract levels of description can follow the chaining process

$$\boxed{\mathscr{U} \xrightarrow{\chi_3} \Sigma_d} \xrightarrow{\chi_2} \Sigma_s^\star \xrightarrow{\chi_1} (\Sigma_s^\star, \mathcal{L}_0, \mathcal{K}_s) \xrightarrow{\chi_0} (\Sigma_s^\star, \mathcal{L}_0, \mathcal{K}_s, \mathcal{K}_t), \tag{6}$$

where $\chi_3(\cdot)$ is the only one which deals with the visual signal. All the other functions involve symbolic processes at different levels of abstraction. From one side, $\chi_3(\cdot)$ exploits the focus of attention on $(t,x) \in \mathscr{D}$ to better process the visual information, and, from the other side, it gets rid of any linguistic structure by relying on the classification of dummy symbols.

4 Conclusions

By and large, there is a lot of excitement around computer vision that is definitely motivated by the successful results obtained in the last few years by deep learning. While recognizing the fundamental progress gained under this new wave of connectionist models, this paper claims that the bullish sentiment behind these achievements might not be fully motivated and that the time has come to address a number of fundamental questions that, once properly addressed, could dramatically improve nowadays technology. The discussion is stimulated by the remark that the construction of learning theories of vision properly conceived for intelligent agents working on video instead of large image collections simplifies any visual task. In particular, the paper promotes the principle of developing visual features invariant under motion, which is claimed to be the only significant invariance that is required to gain the "what" function typical of the ventral mainstream.

References

1. Anderson, J.A., Rosenfeld, E. (eds.): Neurocomputing: Foundations of Research. MIT Press, Cambridge (1988)
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
3. Betti, A., Gori, M.: Convolutional networks in visual environments. Arxiv preprint arXiv:1801.07110v1 (2018)
4. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. InL CVPR 2009 (2009)
5. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. Trends Neurosci. **15**(1), 20–25 (1992)
6. Gori, M., Lippi, M., Maggini, M., Melacci, S.: Semantic video labeling by developmental visual agents. Comput. Vis. Image Underst. **146**, 9–26 (2016)
7. Goroshin, R., Brun, J., Tompson, J., Eigen, D., LeCun, Y.: Unsupervised learning of spatiotemporally coherent metrics. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015, pp. 4086–4093 (2015)
8. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
9. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. J. Physiol. (Lond.) **160**, 106–154 (1962)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates Inc., (2012)
11. Lee, H., Gross, R., Ranganat, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 609–616. ACM. New York (2009)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
13. Marr, D.: Vision. Freeman, San Francisco (1982). Partially reprinted in [1]
14. Braddick, O., Atkinson, J.: Development of human visual function. Vis. Res. **51**, 1588–1609 (2011)
15. Poggio, T.A., Anselmi, F.: Visual Cortex and Deep Networks: Learning Invariant Representations, 1st edn. The MIT Press, Cambridge (2016)
16. Ranzato, M., Huang, F.J., Boureau, Y.-L., LeCun, Y.: Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18–23 June 2007, Minneapolis, Minnesota, USA (2007)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)

18. Andrillon, T.N., Yuval, N., Cirelli, C., Tononi, G., Itzhak, F.: Single-neuron activity and eye movements during human REM sleep and awake vision. Nature (2014)
19. Tavanaei, A., Masquelier, T., Maida, A.S.: Acquisition of visual features through probabilistic spike-timing-dependent plasticity. CoRR, abs/1606.01102 (2016)
20. Dobson, V., Teller, D.Y.: Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. Vis. Res. **18**, 1469–1483 (1978)
21. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_26