Approaches and Challenges for Cognitive Vision Systems

Julian Eggert, Heiko Wersing

Honda Research Institute Europe GmbH,
Carl-Legien-Strasse 30,
63073 Offenbach/Main, Germany

Abstract. A cognitive visual system is generally intended to work robustly under varying environmental conditions, adapt to a broad range of unforeseen changes, and even exhibit prospective behavior like systematically anticipating possible visual events. These properties are unquestionably out of reach of currently available solutions. To analyze the reasons underlying this failure, in this paper we develop the idea of a vision system that flexibly controls the order and the accessibility of visual processes during operation. Vision is hereby understood as the dynamic process of selective adaptation of visual parameters and modules as a function of underlying goals or intentions. This perspective requires a specific architectural organization, since vision is then a continuous balance between the sensory stimulation and internally generated information. Furthermore, the consideration of intrinsic resource limitations and their organization by means of an appropriate control substrate become a centerpiece for the creation of truly cognitive vision systems. We outline the main concepts that are required for the development of such systems, and discuss modern approaches to a few selected vision subproblems like image segmentation, item tracking and visual object classification from the perspective of their integration and recruitment into a cognitive vision system.

1 Introduction

1.1 Motivation: The Quest for a Cognitive Vision System

Imagine a complex visual scene, like given by a working environment in a factory or a traffic situation, where several objects have to be analyzed and kept in mind for a meaningful, visually-guided way of operation. What happens in the mind of humans when interacting with such a scene is still largely a mystery. A plethora of questions immediately arises on how the brain copes with the large potential complexity of visual sensory analysis of complex scenes, in particular when they are not static (which is the case in nearly all situations in a real environment, with most exceptions being artificially generated like when observing a photograph). With *potential complexity* we denote the combinatorial way of choices that the brain has to deal with for the visual analysis: It e.g. has to decide on which visual properties to concentrate (dynamic properties like motion-induced displacements

and appearance changes, static properties like characteristic patterns, colors, shadings, textures, 3D aspects like depth or surface curvature, to name only a few), how to tune the system on these properties (usually the visual properties that the brain has access to are not capable of analyzing a large sensory spectrum in full detail, instead, sensory analysis has to focus on the relevant sensory ranges by a dynamic adaptation process), how to extract single parts and objects from the scene (deciding on what makes up the most relevant aspects on a scene), how to analyze these parts during the time-course of the scene analysis, how to detect and analyze properties that depend on the combined treatment of several parts or objects (like e.g. relational properties where different parts have to be considered as a conjunction, as it is the case when a distance or a relative position of parts is of interest, or the appearance of two similar objects is analyzed in detail for discrimination), and finally, how to combine the results with each other resp. how to bootstrap choices made in a particular domain of the visual analysis using results gained in another domain. This list of choices that a human visual system has to perform is of course non-exhaustive and could be extensively continued; but we can already notice the diversity and the complexity of operations that is involved in such a process. In short, multiple specialized analyses occur during vision, which have to be tuned, adapted, and selectively integrated over time.

To the contrary, when we speak of vision, we often only denote a particular, isolated aspect, like e.g. visual object classification, i.e., the attribution of a class or category label to a selected portion of the visual input. With the previous list in mind, we are able to understand that vision is a complex process within which object classification or any other specialized visual analysis is only one minor component among many. The main reason behind the complexity of visual operations is an inherent *resource limitation*. Now, given that many of the specialized analyses can probably be carried out in parallel, and taking into consideration that the brain is a device with a myriad of elements working concurrently, particularly specialized for parallel processing, why should there be any resource limitation at all? The reason is that the space of possible interactions among several parts and objects in a scene is too large. The combinatorial complexity explodes when visual analyses involve the integration of several cues and objects. In addition, several resources for visual analysis have to be recruited and adapted exclusively for a single visual subtask rendering them inaccessible for others, as can be easily understood in the case of the eyes, which during gazing concentrate on a particular portion of a scene and even on a particular depth. Even though in cases of higher visual processing that are further away from the sensory periphery the case of exclusive resource allocation is not as evident as for the eyes, the logical considerations are analogous. A case where this becomes evident is for objects defined by conjunctions of visual properties (e.g. form and color), which require attentional focusing for correct recognition, a phenomenon that has been hypothesized to work in analogy to an internal "zoom lens" [19, 18] which would allow the preferential but exclusive processing of only one object at a time. Besides some visual preprocessing steps like the extraction of local edges, patterns or velocities, which can be carried out in par-

allel over the entire visual field and which serve as a common sensorial basis, most subsequent visual processing steps suffer from similar resource limitations: They have to be specially tuned to a particular purpose, object or visual task so that they are exclusively specific, meaning that they cannot be used for the inspection of e.g. another object at the same time since this would require the same processing substrate. In other words, they have to be *controlled* by some higher level instances and the control strategies of specialized visual processes have to be orchestrated depending on higher level demands, as provided by a broader knowledge context or information about a task that a system has to perform.

1.2 Access to Visual Memory

A further important factor that introduces a resource constraint is visual memory. Although this is a term with a broad range of meanings, here we denote it as the capacity to retain information about aspects of a visual scene that can be recalled at later moments or used to reinspect parts of the scene. We can retain information about form, visual properties like color, texture, shading and reflectance, as well as positions and positional relations of visual objects. On a scene level, we can recall the experience of a particular scene impression, as well as the overall spatial arrangement and identity of visual objects. In the following, we use the term visual memory in the sense of a working memory for the current visual scenery [1].

While it is undisputed that selected results of visual analysis subprocesses are stored in visual memory, there is a diverging debate about how much information can be stored, what exactly is stored and to which degree of accuracy. The last point refers e.g. to the dispute whether the brain attempts a faithful internal reconstruction of the physical world that it inspects through its visual senses. The alternatives are that visual memory may be trying to construct an as-complete-as possible internal representation of the world as opposed to being partial and selective, in the sense that it only stores information about specific objects that are of interest at a given moment. Similarly, it is argued that the brain targets at an accurate representation of the true physical causes of a sensory input (e.g. representing the world as an accurate geometric environment with physical objects) vs. representing the world only up to the level of description that suffices for a given task or behavior in a situative context. The tendency is towards a partial and selective representation at a suitable level of description that is adjustable to the situation, with the main arguments supported by *change blindness* (the fact that changes of visual properties or parts of a scene are not noticed if they are not attended, e.g. [41, 6]) and memory capacity measurements (many psychophysical experiments suggest that the capacity

[1] A metaphor suggestive for the type of information that is stored in visual working memory is that of a theatre stage as introduced by [5], containing a context, a scenario, actors and objects; in addition to spotlights that highlight parts of the scene.

of visual short term memory is extremely small, about 4-5 items, see e.g. [45, 15], but this refers particularly to a very specific type of iconic memory). We will briefly return to this topic in section 6; the important bottomline here is that visual memory constitutes a resource bottleneck for visual processing.

Why should this be so? If we regard visual memory as more than a mere buffer for storing n memorized iconic items, but rather a sketchpad where information from the specialized visual analyses can converge, then it is *the* substrate where selective, object- and situation-specific integration of information occurs. As argued in the last section, such an integration most likely involves selective tuning of the underlying visual subprocesses, under consideration of resource constraints like exclusive recruitment and competition. Specific integration would therefore involve an active, selective choice of the system to inspect certain visual properties, based on the knowledge that the visual memory is able to provide. At the same time, such a selective choice would imply that an order for accessing visual subprocesses in terms of prioritization schemes is imposed (some visual objects or attributes are identified as being important to be inspected before others), which on its own implies that the proper sequentialization has to be cared for. Such a control scheme would require a large amount of prior knowledge about the visual subprocesses themselves, i.e., it would require knowledge of the system about its own sensory apparatus and its limitations.

The picture that emerges is that of a generalized visual memory working as the central executive for the control instances responsible for an active acquisition of visual information. It would be a visuospatial sketchpad (see [7] for an early proposal of a visuospatial sketchpad, however quite different from the specific one proposed here) where visual events and measurements are annotated, hypotheses about causes relating visual events are created (eventually leading to notions of rudimentary objects or interaction elements), corroborated and refuted, the entire visual presence (the knowledge about the current visual situation) is kept up-to-date and from which the processes for the underlying visual analyses are controlled. At the same time, such a sketchpad would be the ideal candidate for the integration of, and coupling with, information from other senses.

1.3 Overview

In this paper, we will put forward the idea that the combinatorial complexity of controlling several visual processes should be at the center of considerations when trying to understand a cognitive vision system. We will term this the "control view of cognitive vision", and, in the following, mean this view when we speak about "cognitive vision", if not especially denoted otherwise [2]. This is a view that differs considerably from most standard approaches to vision [3], and that has a number of deducible consequences that we should focus on. First of all, we

[2] Although the opinion of different authors on what cognitive vision is differ substantially in the literature.

[3] However, other work in comparable directions exists, see e.g. [9, 35].

can ask which high-level representational framework does allow best for a vision system operating in the described conditions. Second, we can ask which low level visual sensory subprocesses are a prerequisite for such a system, especially when operation in a sufficiently complex visual environment is demanded. One of the many pitfalls of modern vision systems is that the need for control processes does not become apparent if the conditions are too restrained or specific. Third, we can ask what particular characteristics should visual subprocesses have that operate in a cognitive vision system as described. How do they specialize to enable control, and how general should the results be that they deliver to other parts of the system? Forth, we can ask who mediates the control processes. This question has tight interactions with the quest for understanding one of the key ingredients for the power of visual processing in the human brain, *visual attention*. Fifth, we have to ask about the intrinsic properties of the control process (or, rather, the plural *control processes*) itself. What type of convergence of visual information does it need, where and how (and when!) is this information represented, what does the control process optimize, how does it subdivide and delegate control to visual subprocesses and tune them accordingly, and how does it finally evaluate and integrate the results of a visual analysis. In its majority, these questions have rarely been approached yet by current vision research (at least under the perspective of an integrated cognitive vision system), and the scientific research is not able to give a concluding answer to any of them at this point [4]. Nevertheless, they provide a starting point for a paradigm shift in the research of cognitive vision systems.

In the following sections, we will proceed step-by-step to develop the idea of a control-based cognitive vision system. In a first section, we will give a brief account on current paradigms in vision research, shortly reviewing the main characteristics of the different approaches and trying to position the control view of visual processing within these ideas. We will see that the control view describes a regime that is not covered by the two prominent (admittedly extreme) paradigms of current vision - cognitivist and emergent - , rather, it can be identified as a third paradigm that poses important questions on its own right that are not explicitly covered otherwise. In addition, we will explain how the control view is related to many open themes reoccurring in visual research, as there are: Active vision, grounding, anchoring, binding, visual anticipation and prediction. In a second, more extensive section, we exemplarily zoom in on a few specialized visual processing "subsystems" that current vision research has identified as key ingredients of a general vision architecture. We will review the properties of such systems, concentrating on them from the perspective of the control view of cognitive vision. These subsystems would represent the (on one hand) fixed basic structure, since it determines which visual properties the system can in principle analyze; on the other hand they would have to be sufficiently

[4] In particular, the last question is central since it somehow codetermines the previous ones; i.e., we need an understanding of the nature of the control processes in vision to be able to design and understand better visual subprocesses dealing with specialized analyses.

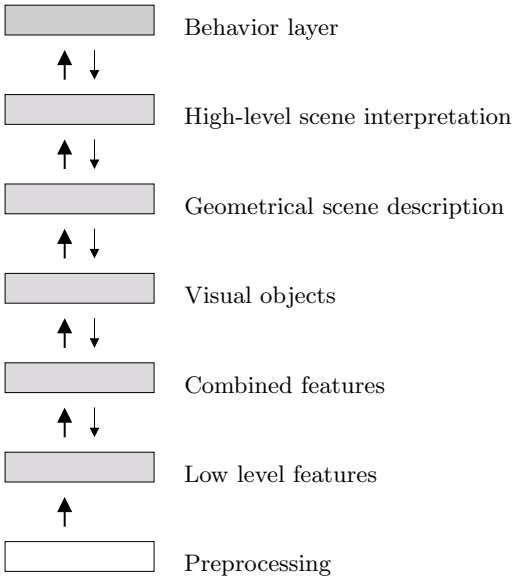

Fig. 1. A typical architecture of a cognitivist vision system. A cascade of representation layers serves to construct increasingly amodal abstractions that describe the objective external world. At the highest level, this is used to achieve intelligent behavior. Details of the representation have to be foreseen by a human designer.

flexible to be able to be recruited and tuned by control processes. Being an open topic, a short discussion about the type of required representation for cognitive vision and the interaction of this representation with the visual subprocesses concludes the paper.

2 Challenges for Cognitive Vision Systems

In the broadest sense, the term cognitive vision is used for the current state of research dealing with vision systems that incorporate a rich internal state representing accumulated visual information and which operate flexibly even under unforeseen changes of the visual environment. It has been introduced to separate these from previous attempts of visual systems that often were tailored to specific visual contexts and which exhibited little robustness and adaptivity.

2.1 The Range from Cognitivist to Emergent Vision System

The two main paradigms in cognitive vision are the cognitivist and the emergent systems approaches. The *cognitivist* approaches assume that the target of such systems are the faithful reconstruction, by terms of an appropriate explicit representation, of the external world from visual data. The representation is centered around the requirement that it should describe the objective external

world [49] as accurately as possible, including its geometrical and physical properties (already Marr stated that the goal of a computer vision system should be a "description of the three-dimensional world in terms of surfaces and objects present and their physical properties and spatial relationships" [32]).The process to achieve this is an abstraction chain, which starts at the perceptual level, abstracts from there using appropriate symbol sets, and reasons symbolically with the gained representations in order to achieve an intelligent behavior. For the cognitivist approach, the main job of a cognitive vision system is to provide the symbolic representation which then can be operated upon with more general logical frameworks. Since all vision systems operating in real world have to start at a quantitative level based on noisy and uncertain data, modern cognitivist systems are turning towards subsymbolic preprocessing levels based on probabilistic, machine learning, or connectionist techniques.

Figure 1 shows an example of a cognitivist system. We can see a cascade of stages that extracts increasingly complex components of a scene, ranging from signal to symbolic representations through several layers of processing. Each stage involves computations that abstract and generalize on the preceding stage. At the final stage, the representation is used to reason about high-level concepts such as spatial object configurations, and to generate behavior. The information flow between the layers may be bidirectional, expressing that there can be a modulatory influence from higher levels to improve the performance of lower level processing stages.

A major critique of cognitivist approaches is that they heavily depend on the designer to find a representation that is suited to the solution of a visual problem, and that the representational structures gained from human idealization exhibit a bias that is detrimental. In addition, purely symbolic representations and rule-based reasoning on such representations has proven to be insufficient to capture the variability of real-world sensory perception. Probabilistic and learning frameworks are being proposed as alternatives to this problem [36], relaxing the demand for an explicit representation and adapting a systems structure to empirically provided constraints.

The second paradigm is the *emergent systems view*. This view emphasizes that the system is embedded in a cognitive agent whose capabilities are determined and have been developed in interaction with an environment. The agent operates within the environment and constructs its representation of the world as a result of this operation [33]. This is enabled by a continuous and real-time interaction of the system with the environment, and leads to systems that can cope well with the specific environmental conditions and the variability of the system-environment interaction.

Figure 2 shows a sketch of an emergent vision system. The agent preprocesses the visual data and then passes it on to a flexible structure where the proper representations should emerge during system-environment interaction and co-determination. The importance here is on the coupling between the system and the environment through the behaviors of the agent. In the emergent paradigm, the purpose of vision is simply to provide the appropriate sensory data to enable

sensible actions. The richness of the developed vision system is to a large extent determined by the richness of the action interface. The work of the designer is to choose the developmental structure, the complexity of the environment (which may vary over time) and the action interface.

Emergent vision systems are usually implemented using parallel, real-time and distributed architectures. Dynamical systems models provide characteristics that allow, in principle, the development of cognitive visual functions, achieved by means of consequent self-organization. The vision system as such is a sort of black box, which adjusts its dynamics so as to achieve a desired behavior. The representation of the visual information is implicit, in the sense that there are no a-priori identifiable states or nodes representing entities of the visual world like objects, etc. In addition, no symbolic representations (especially human-designed) are required. Cognitive phenomena like visual memory should emerge from the developmental process; in addition, identifying these phenomena is a matter of interpretation of the systems dynamics by an external observer. The accuracy and completeness of the visual representation of an emergent vision system is optimally adapted to its interaction repertoire, meaning the it is just sufficient to enable the system to do certain things.

In practice, although systems could be developed that exhibit surprisingly non-trivial behaviors that would otherwise require considerable designing efforts (see e.g. [21] and [48] for a review), it remains to be shown that emergent visual systems can develop higher-order cognitive capabilities. Solutions evolved by emergent systems tend to specialize to a particular visual context (which often is representationally poor or at least does not require perceptually more abstract representations) and have problems to scale up and generalize to other domains. It is again the burden of the designer, this time not to choose the detailed representational structure but the teaching signals, the environmental richness and the necessary learning mechanisms. Furthermore, the capabilities to generalize are often already given by an appropriate preprocessing stage, which also has to be provided by the designer.

2.2 Grounding and Binding

In a symbolic cognitivist approach, an internal representation of the world from sensory signals is gained by increasing abstractions that could allow for a decoupling from the systems perceptual apparatus (they become *amodal* [8]). If this were so in a rigorous sense, representational parts of the system could be isolated that have no relation at all with the external world. The question then arises how these parts can have semantic content on their own right, i.e., a meaning that refers to real world entities and behaviors. In a more relaxed consideration, one could say that representations at higher processing levels of a cognitivist system loose the information about the original sensory objects that created them. This is called the so-called *symbol grounding* problem.

There is a very close analogy to a second hypothetical problem of cognitive science, the *binding problem*. Interestingly, this second problem is usually

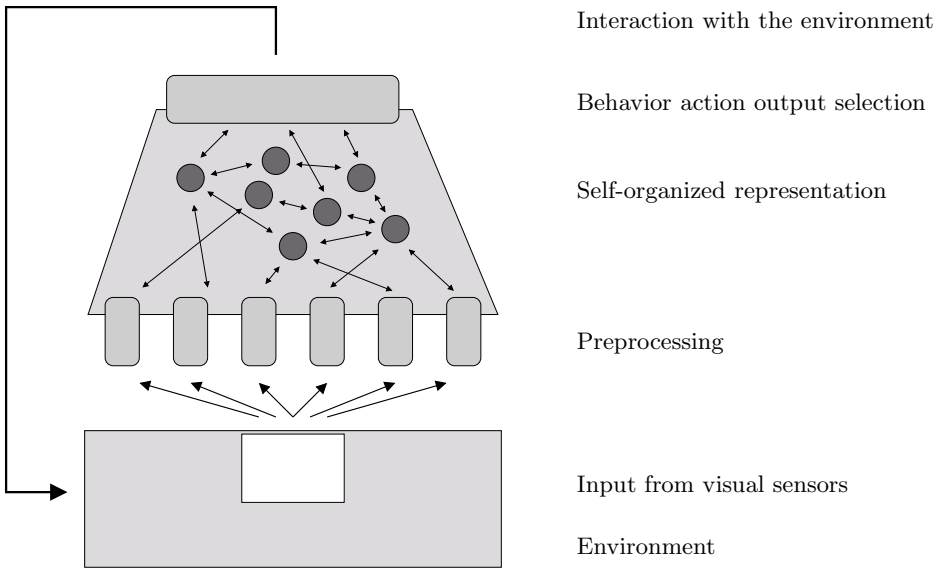

Fig. 2. The emergent systems view. In this case, the system is embedded into an environment and the representations self-organize by continuous interaction with the environment. No symbolic or human-designed representations are necessary, the expectation is that cognitive phenomena arise implicitly from the developmental process.

motivated by the connectionist and neural network background and not by cognitivist approaches. It denotes the loss of reference to the originally constituting cues as one moves along a processing hierarchy. An often cited example is given by a configuration of two stimuli, each defined by a particular conjunction of two different cues (e.g. form and color, with stimulus 1 having a T-form and red color, and stimulus 2 being a green cross). If form and color are processed independently of each other and generalize over position, all the information that arrives at a higher stage is that there are two forms and two colors present. The "binding" of its two constituting cues to an object is then lost, making it impossible to retrieve the information which colors and forms correspond to each other. In a sense, this is the same problem as for symbol grounding, only that we do not consider the relation between external sensory signals and internal symbolic representations, but between internal representations at different levels of abstraction.

In the brain, binding losses indeed seem to occur when inspecting scenes with several objects defined by conjunctions of cues, leading to *conjunction errors* [46], meaning that people make mistakes when forced to determine which form corresponds to which color for each object. These errors always appear in combination with attentional overload, i.e., when there are not sufficient attentional resources that can be devoted to each object, either because of too many objects present or too short presentation times.

2.3 Anchoring and FINST's

Both the symbol grounding as well as the binding problem are rather artifical, idealized constructions, as we will argue in the following. They are based on the assumption that even when looking at a representational result in isolation one should get an automatic reference to the original objects or lower level representations that generated it. By automatic it is meant that this reference is a passive, or straightforwardly deducible property.

Proposals on how to do this exist, at least for the binding problem. They introduce mechanisms which code the references between representational items by additional states, which can be evaluated when accessing the representations. These states serve to relate (i.e., they "bind") the different representations with each other, or representations with parts of the original sensory input. One proposal is to use time labeling mechanisms, e.g. by the phase of an oscillatory activity that synchronizes by appropriate internal system dynamics.

Although this seems reasonable at a first glance, it is still a passive property, in the sense that it should be automatically present when some access to a part of the internal representation is required. Nevertheless, the capacity of such mechanisms is severely limited, since they can work only on a very small subset of the representational items at a time.

The important point here is that binding should not be considered as a representational property, but rather as an *effortful process*. Effortful meaning here that it requires active and selective focusing on a small subset of representational items because it allocates exclusive processing resources in a vision system. As such this process has to be controlled from within a larger knowledge context, including both context information about the visual scene as well as specific information about the visual subprocesses that mediate the binding.

A term that appears in the literature in a similar context is that of *grounded cognition* [8] and *anchoring* [14]. Grounded cognition emphasizes the role of modal (perception specific) representations, internal simulations, imagery and situated action for binding internal representations to external objects. Anchoring is the "process of creating and maintaining the correspondence between symbols and percepts that refer to the same physical objects" [14], and is presumed to be a necessary component in any physically embedded system that uses a symbolic internal representation.

For the special domain of simultaneous multiobject tracking and attentional selection, *FIngers of INSTantiation* (FINST's, [38]) have been proposed to solve the anchoring task in early vision. It is argued that the process of incrementally constructing perceptual representations, solving the binding problem as well as grounding perceptual representations in experience, arises from the capacity to select and keep track of a small number of sensory items. These items are identified to have a particular, consistent and enduring identity that can be maintained during the tracking process despite considerable changes in their properties. In a sense, FINST objects have been described as mental "rubber bands" between internal representations and external objects.

Extending the ideas of a cognitive vision system from section 1, we consider FINST's to be but one emanation of a deeper concept that emphasizes the process and control aspects of establishing temporary, selective correspondences between more abstract, higher-level representations and the input at the sensory periphery.

2.4 Anticipation and Prediction

What is the deeper purpose of anchoring? On one side anchoring and its manifestations, e.g. as proposed by FINST's or as experienced during tracking, represents a behaviorally useful capability in its own right: It keeps the attended item in focus so that it can be easily reaccessed when needed. It also stabilizes the sensory input, so that some variablities in the appearance change are compensated for, as can be seen in a straightforward way in the case of translational (in)variance during object tracking (this is particularly evident when combined with the overt behavior of gazing at an object during smooth pursuit).

But the more important point of anchoring is that it requires an active internal anticipation of the stimuli as they are expected in the near future. A good anticipation is necessary because it narrows down the search range and therefore decreases the processing resources for reaccessing the item in next timesteps (here we close the loop to the resource limitation arguments of section 1).

Anticipation is a hard generalization task for a cognitive vision system, because visual stimuli are highly variable due to two different reasons: First, the items themselves as they appear in the physical world (e.g. non-rigid objects) together with their projection by the sensory apparatus (e.g. 3D onto 2D, causing view changes as an object rotates) are highly variable, and second, nearly every behavior can have a severe effect on the sensory input (which is straightforward for direct interaction with objects like manual manipulation, but consider also indirect effects like egomotion of the system changing the visual appearance of an object). If a cognitive system wants to generalize its anticipatory capabilities, it has to be able to separate the two sources of variability. This has deep consequences: It means that the system has to acquire knowledge about the *process of its own sensory apparatus* on one side and about the *external causes of sensory inputs* on the other side [5]. Process knowledge on its own sensory apparatus will allow such a system to discount or compensate for changes caused by own behavior as well as internal adaptation and modulation, leading to a more robust

[5] This reminds us to the cognitivist idea of building an accurate and detailed model of the physical world, see section 2.1. In the current argumentation, however, we 1) emphasize the role of the dynamic *anticipatory and control process* instead of concentrating on the structure and content of the internal world representation, 2) we make no claims about the degree of accuracy, so that variably coarse descriptions of the physical causes may already be sufficient for reasonably good anticipation, depending on the demands of a task, and 3) the modal knowledge of the system about its own sensory capabilites is crucial, whereas purely cognitivist approaches prescind from this, targeting an amodal, abstract representation of the outer physical world only.

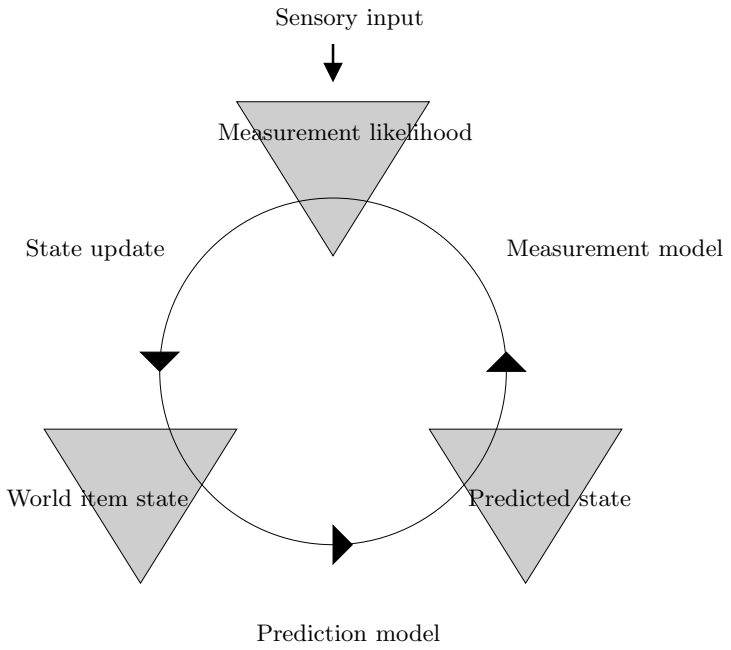

Fig. 3. Prediction as a fundamental process for active acquisition of visual information in a cognitive visual system. Shown is a prediction-measurement-update loop that makes use of two types of knowledge: 1) The prediction model that expresses how the state of a sensory item changes over time, representing knowledge about the external causes of sensory inputs, and 2) the measurement model that comprises knowledge about a systems own sensory processes, anticipating the expected sensory input for a given hypothetical state of a sensory item. In the control view of cognitive vision systems, several prediction-measurement-update loops interact, comprising process information about sensory items and the active coupling between internal representations and external sensory measurements.

and stable sensory analysis of sensory objects. Knowledge about the external causes leads to more stable and generalizable representations of the world, since many visual changes occur by own sensory behavior and not by changes in the state of the world objects themselves (as in the case when objects are static but an observer moves around them).

The two types of knowledge and their interaction during active sensory processing can be schematized in a perceptual cycle as shown in figure 3. The knowledge about a sensory item resp. the external causes of a sensory input is represented in the item state (bottom left) and a prediction model that indicates how the state is expected to change in the future (bottom). The knowledge about its own sensory processes is comprised in the measurement model, which is applied onto the sensory input (from top) to estimate how likely a sensory measurement is for the assumption of a predicted item state. This likelihood is used in an update step to adjust the items state. In the control view of cognitive

vision, perceptual cycles with prediction-measurement-update steps constitute a central element, both for low-level visual processes (for example in a tracking context), as well as operations far from the sensory periphery, working on more indirect representations.

2.5 A Modern Account of Active Vision Systems

In this paper we argue that the selective choice of sensory information and the corresponding tuning of sensor parameters, i.e., the effective focusing of visual processing resources, is a necessary property of any artificial or biological vision system with a claim for some minimal generality and flexibility. Such focusing capabilities largely make up for the flexibility of biological visions system that, depending on the (visual) task in question, the context of already acquired as well as prior information and the available processing resources, may deploy very differently even for identical sensory input [6].

The idea is that different visual tasks, motivated by internal goals, trigger different visual processes, and that these processes have to be organized in a systematic way because there is simply not enough capacity otherwise. Such a system would therefore continuously have to modulate and adapt itself, organize the cooccurrence or their temporal order of visual operations, and monitor their success. The processes referred here are mainly seen as internal operations, such as e.g. the selective enhancement of competition, the dynamic adjustment of filter parameters or the concentration on special feature channels, like edges, motion, color, etc. The means by which this could occur is via attention, combining *top-down* signals that provide expectations and measurement resp. confirmation requests with *bottom-up* signals that provide measurements coming from the sensors.

The task-dependence of internal organization processes in such a vision system is a view shared with behaviorist paradigms, which concentrate on "visual abilities which are tied to specific behaviors and which access the scene directly without intervening representations". One of them is *active vision* (see e.g. [2]), a term that is used for systems that control the image acquisition process e.g. by actively modulating camera parameters like gaze direction, focus or vergence in a task-dependent manner. Along a similar line, *purposive vision* ([1]) regards vision processes always in combination with the context of some tasks that should be fulfilled. Common to both active and purposive vision approaches is that they have concentrated on *overt* behaviors and actions that are directly observable from outside, and in how visual information can be extracted that supports particular behaviors.

To the contrary, in the framework put forward in this paper, a proper minimal representational structure on which the control and modulation processes can operate is crucial. Visual cognition is understood as any goal-driven mediation between an internal representation and the incoming sensory stimulation. The

[6] With the classical attentional phenomena being one notorious example for the focusing of visual resources during operation

mediating control processes serve to gather visual information that could be potentially used for guiding overt behaviors, (without necessarily being tied to the behaviors). In fact, we interprete any internal modulation and attentional focusing as a *virtual action*, in principle not different from overt actions [7]. The basic assumption is that, from a task-driven perspective, there is simply not enough processing capacity to cover all the different ways to operate on the visual input in a hard-wired manner, so that a vision system has to flexibly organize its internal visual processing during operation and this organization has to be controlled by the (visual) intentions of the system. The tasks and intentions we mention here are supposed to be of intermediate level, but still relatively close to the sensory domain, like e.g. "concentrate on an interesting moving object in the visual scene and keep its coordinates up-to-date", "compare the feature composition of two objects" or "track an object, use motion segmentation to separate it from the background".

So what is a cognitive vision system intended to do operationally? It should:

- Establish temporary or semi-continuous links from internal representations of sensory events to the incoming sensory information ("anchoring").
- Use this possibility actively when different / additional / not yet analyzed data is required or information has to be renewed.
- Work with the stored information to establish relations, discover regularities and analogies, i.e., explore and learn about the visual sensory world.
- Use the gained world knowledge to control active processes for the acquisition of visual information.

The establishment of temporary or semi-continuous links between internal representations and sensory information occurs by means of prediction-measurement-update loops as introduced in section 2.4. Ideally, the granularity of the information represented in the prediction-measurement-update loops need not be defined a priori, but may be developed in an self-organized way by visual investigation, resulting in the right level of abstraction and detail. In any case, it is assumed that several loop exist, that they interact and that they even organize in a hierarchical manner. One example is given by the multiple adaptation processes at different abstraction levels occurring during vision, as there are:

- Local modulation processes adapting to optimal sensory ranges, as e.g. given by local filter contrast adaptation and contour completion processes.
- Prediction-measurement-update loop of elementary sensory states of visual items, such as the position update of an item that is being tracked.
- Higher-level prediction-measurement-update loops dealing with engagement and loss of lower level loops, such as given by processes of finding suitable sensory items, engaging in a tracking loop, evaluate the success to detect

[7] We even explicitly disregard any overt actions like gaze or head orienting for the following argumentation, since we think that the more interesting aspects of visual cognition appear without the need to concentrate on the hardware specificities of sensory devices.

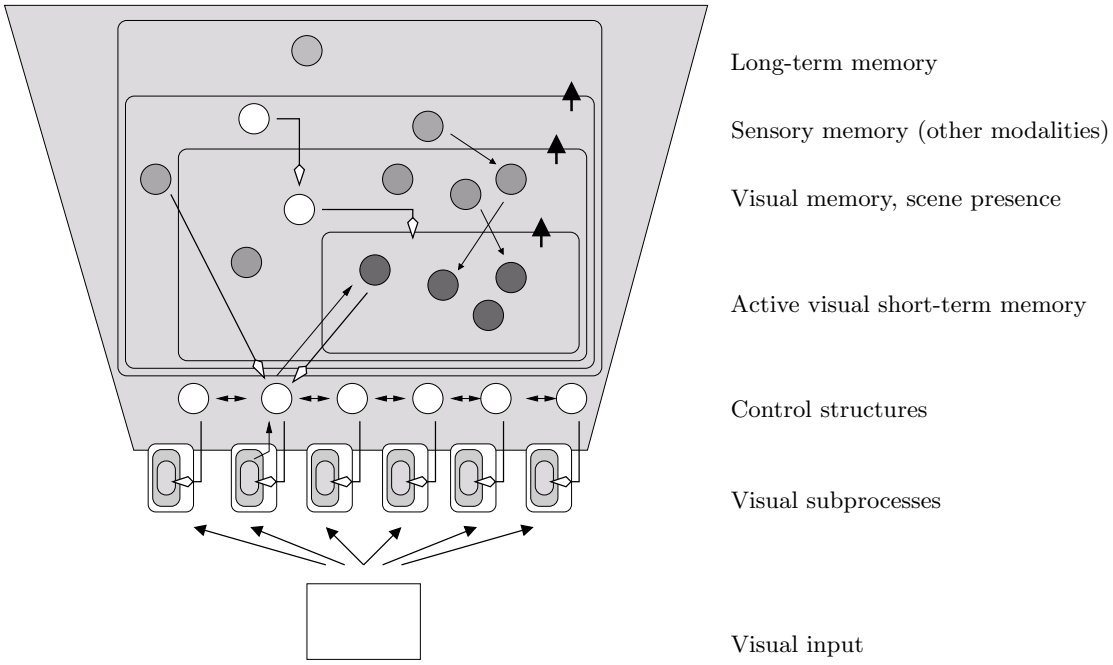

Fig. 4. Scheme of the necessary structures for a cognitive visual system with a focus on virtual visual actions and their control. White nodes denote control instances which mediate between sensory item representation and visual processes at all levels of abstraction. Sensory items denote state information which can be updated by temporarily activating prediction-measurement-update loops, as it occurs on a rapid timescale for the items in active visual short-term memory. The basic control framework as well as the visual subprocesses are assumed to be given, whereas the memory content and its representational structure can be developed in an emergent way.

e.g. when a prediction fails, reacquiring the item when it was lost and finding an alternative item if necessary.

– Serialization processes. Since the prediction-measurement-update loops of different visual events compete for resources, higher-level visual tasks requiring several of them have to organized temporally, e.g. establishing which precede others.

Figure 4 shows a scheme of the necessary structures for a cognitive visual system as they are proposed in this paper. On the lowest level, after some general-purpose preprocessing that is independent of the items in visual memory, multiple visual subprocesses in form of prediction-measurement-update loops apply. The recruitment and the modulation of these loops is organized by control structures (indicated by white nodes) acting as proxies between them and visual memory. The loops work largely in an object-specific mode, e.g. specialized to

search and find sensory objects with predefined visual properties, or segment a visual region starting with an already known position and approximate size.

Nodes indicate "representational compounds", comprising both states of sensory items as well as process knowledge on how to link the state with active visual subprocesses in a prediction-measurement-update loop to couple the state with sensory perception. Different levels of abstractions of memory items are indicated by boxes, ranging from short-term, purely visual to long-term, cross modal memory. The main difference between them is in the type of integration into the sensory control processes, reflecting a control hierarchy rather than an representation hierarchy as previously in the cognitivist paradigm of figure 1. As an example, active visual short-term memory is composed of those visual items in memory that are engaged in a short-term prediction-measurement-update loop, anchoring them temporarily but continuously with sensory events. In our view, this is a buffer of a small subset of task-relevant items which are dynamically selected by a responsible control structure (in figure 4 the white node in the "scene presence" frame) from a larger number of items that make up the visual scene memory. The items in active visual short-term memory provide predictions and modulation priors to visual subprocess control structures (indicated by the arrows with white heads), which steer the subprocesses on demand and update the information of the items. Other information from sensory memory can provide high-level modulation priors, as indicated in figure 4 by the additional arrow from top-left to a subprocess control structure. However, control processes mediating between memory and the active acquisition and update of information are assumed to work not only at the sensory periphery, but also between higher levels of representation.

A cognitive vision system as proposed is therefore composed of an intertwined hierarchy of representations providing item and control information, together with a corresponding control dynamics of the processes necessary for the active acquisition of information at any representational level. Control structures of this kind could in principle self-organize in an emergent way (see the emergent systems view from section 2.1), but are very hard to develop systematically, in a self-organized fashion. Therefore, we propose to provide a framework for control structures and their representations, but not the representation of sensory items themselves, which could develop incrementally and autonomously during interaction. Only at the lowest sensory level, at the interface to some well-defined visual subprocesses we would predefine the basic visual sensory events.

3 Ingredients of a Cognitive Vision System

In the following, we present some concrete descriptions of visual subprocesses that would be needed by a cognitive vision system. As suggested in the introduction, we are interested in the control aspects of such subprocesses, highlighting the multiple control loops that arise when being operated in combination with other subprocesses in conjunction with visual memory representations.

The particular choice of the subprocesses is not intended to be unambiguous or complete; rather, it is motivated by their suitability to be recruited into a general-purpose control framework for cognitive vision.

3.1 Preprocessing Example: Segmentation

One very important visual subprocess that precedes several other visual operations is image segmentation. In the following description we understand by image segmentation the segregation of the 2D visual input space into 2 regions, one characterizing the (generally more specific) region of interest corresponding to an object or a part of the scene, and the other one (generally unspecific) corresponding to the rest, i.e., the "background". We describe one image segmentation method of choice (there are numerous) that we are using for building a cognitive vision system, again with the focus of understanding it in relation with a superordinate control instance.

The segmentation occurs by means of level-set methods [37, 34, 57, 11, 28], which separate all image pixels into two disjoint regions [37] by favoring homogeneous image properties for pixels within the same region and dissimilar image properties for pixels belonging to different regions. The level-set formalism describes the region properties using an energy functional that implicitly contains the region description. Minimizing the energy functional leads to the segmentation of the image. The formulation of the energy functional dates back to e.g. Mumford and Shah [34] and to Zhu and Yuille [57]. Later on, the functionals were reformulated and minimized using the level-set framework e.g. by [11]. Among all segmentation algorithms from computer vision, level-set methods provide perhaps the closest link with the biologically motivated, connectionist models as represented e.g. by [24]. Similar to neural models, level-set methods work on a grid of nodes located in image/retinotopic space, interpreting the grid as having local connectivity, and using local rules for the propagation of activity in the grid. Time is included explicitly into the model by a formulation of the dynamics of the node activity. Furthermore, the external influence from other sources (feedback from other areas, inclusion of prior knowledge) can be readily integrated on a node-per-node basis, which makes level-sets appealing for the integration into biologically motivated system frameworks.

Level-set methods are front propagation methods. Starting with an initial contour, a figure-background segregation task is solved by iteratively moving the contour according to the solution of a partial differential equation (PDE). The PDE is often originated from the minimization of an energy functional [34, 57].

Compared to "active contours" (snakes) [27], that also constitute front propagation methods and explicitly represent a contour by supporting points, level-set methods represent contours implicitly by a level-set function that is defined over the complete image plane. The contour is defined as an iso-level in the level-set function, i.e. the contour is the set of all locations, where the level-set function has a specific value. This value is commonly chosen to be zero, thus the inside

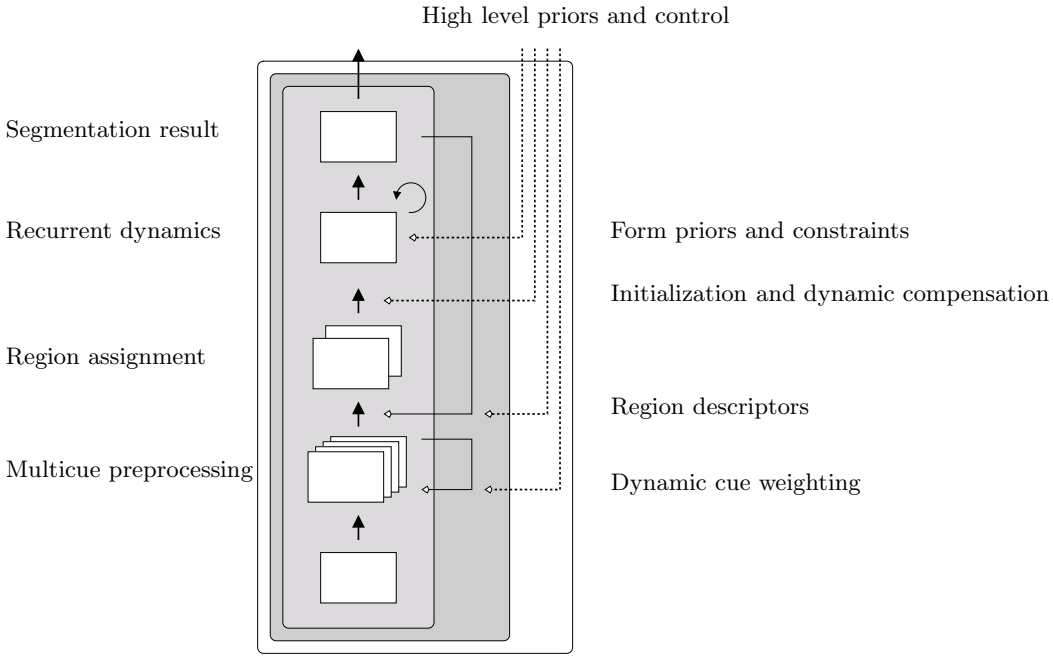

Fig. 5. Processing and control flow in a generalized segmentation process. The proper segmentation process occurs in a recurrent network dynamics indicated by the circular arrow. Local control loops involve the adjustment of the dynamic cue weightings as well as the compound of region assignment, recurrent dynamics and segmentation result evaluation. Furthermore, higher-level contexts requiring visual memory are involved in controlling the cue adaptation and selection process, the region descriptors, the initialization as well as other prior information on the item that should be segmented (arrows with white heads). All these processes are assumed to be covered by corresponding perceptual prediction-measurement-update cycles.

and outside regions can easily be determined by the Heaviside function $H(x)$ [8]. A level-set function $\phi \in \Omega \mapsto \mathbb{R}$ is used to divide the image plane Ω into two disjoint regions, Ω_1 (background) and Ω_2 (object), where $\phi(x) > 0$ if $x \in \Omega_1$ and $\phi(x) < 0$ if $x \in \Omega_2$. A functional of the level-set function ϕ can be formulated that incorporates the following constraints:

- Segmentation constraint: the data within each region Ω_i should be as similar as possible to the corresponding region descriptor ρ_i.
- Smoothness constraint: the length of the contour separating the regions Ω_i should be as short as possible.

[8] $H(x) = 1 \,\text{for}\, x > 0$ and $H(x) = 0 \,\text{for}\, x \leq 0$.

This leads to the expression [9]

$$E(\phi) = \nu \int_{\Omega} |\nabla H(\phi)| dx - \sum_{i=1}^{2} \int_{\Omega} \chi_i(\phi) \log p_i \, dx \tag{1}$$

with the Heaviside function $H(\phi)$ and $\chi_1 = H(\phi)$ and $\chi_2 = 1 - H(\phi)$. That is, the χ_i's act as region masks, since $\chi_i = 1$ for $x \in \Omega_i$ and 0 otherwise. The first term acts as a smoothness term, that favors few large regions as well as smooth region boundaries, whereas the second term contains assignment probabilities $p_1(x)$ and $p_2(x)$ that a pixel at position x belongs to the inner and outer regions Ω_1 and Ω_2, respectively, favoring a unique region assignment. Additional terms can be very easily appended, expressing e.g. prior knowledge about the expected form of the region.

Minimization of this functional with respect to the level-set function ϕ using gradient descent leads to

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \left[\nu \, \mathrm{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) + \log \frac{p_1}{p_2} \right] \quad . \tag{2}$$

A region descriptor $\rho_i(\mathbf{f})$ that depends on the image feature vector $\mathbf{f}$ serves to describe the characteristic properties of the outer vs. the inner regions. Examples are statistical averages, variances or histograms. The assignment probabilities $p_i(x)$ for each image position are calculated based on an image feature vector via $p_i(x) := \rho_i(\mathbf{f}(x))$. The parameters of the region descriptor $\rho_i(\mathbf{f})$ are gained in a separate step using the measured feature vectors $\mathbf{f}(x)$ at all positions $x \in \Omega_i$ of a region i. This occurs alternatingly, updating in a first step the level set function, characterizing the segmented region, and in a second step the region descriptors of the inner and outer regions. In [16, 51], probabilistic and histogram-based region descriptors are combined with level-set methods for an application in a multicue setting, required for general-purpose segmentation tasks (see below).

Figure 5 shows a block diagram of a more generalized segmentation framework. A visual input is first preprocessed by analyzing different but arbitrary visual cues and properties, like colors, textures and gabor structures at different orientations (but also more sophisticated cues can be incorporated, like disparity from binocular vision, or motion estimates from a sequence of images). The only prerequisite is that they all operate in the same spatial image space. For each target that should be segmented, the cues are combined with their proper region descriptors to get the assignment probability maps. These are fed into a recurrent dynamics as described to minimize the level-set functional. In a post-processing step, the segmentation result can be used for other purposes, like image classification, or the extraction of statistics resp. new region descriptors for the gained region.

The link with the control view of cognitive vision systems appears when we regard the numerous possibilities to control and tune the segmentation process

[9] Remark that ϕ, χ_i and p_i are functions over the image position x.

using prior knowledge to specialize it to a given task. Figure 5 indicates these control influences by the vertical arrows from the top. First of all, the multicue preprocessing demands for a cue adaptation and selection process, since some cues provide correlated, or, alternatively, irrelevant information. Second, specific region descriptors can be provided by previously gathered prior knowledge, e.g. about an objects' color, texture, etc. Third, the recurrent level-set dynamics can incorporate explicit information about a regions expected form, spatial preferences, and dynamic compensations (e.g. if during the segmentation process the visual input changes systematically so that it can be predicted). Sitting on top of this, modules have to decide on all these control alternatives, deciding on a first place that there may be a candidate area of the scene which may be worthwhile to look at or that more detailed information about an object with assumed properties that should be highlighted by the segmentation process is required. Similarly, modules have to decide on the success of the segmentation, detecting a failure and reengaging into the segmentation process if necessary.

The segmentation process itself (i.e., the candidate locations for segmentation, the prior assumptions for the segmentation tuning, the results and all the control issues) would have to be captured by an appropriate representation at visual working memory level, as was suggested in sections 1.2 and 2.5.

3.2 Multicue Tracking

As suggested in the first and second sections, tracking objects or other parts of a scene is a fundamental property of a cognitive visual system to temporally establish and maintain a link between an internal representation and sensory measurements originated by external causes. In short, tracking is a constrained feature and property search, dedicated to a object that can be described by specific, but rather arbitrary visual features (e.g. a visual pattern or statistical properties like certain cue combinations), together with an iterative estimation of dynamic object properties like its position, velocity and visual changes.

Humans are generally able to track arbitrary objects even if they have not seen or learned them before (i.e., without long-term memory), i.e., they can start tracking immediately from any region with characteristic properties. In addition, objects can be tracked even if their visual configurations change considerably (even if these changes can sometimes not be reported, [47]), it seems to be sufficient if certain dynamic assumptions are fulfilled (in easy cases, smoothness and continuity in one of the cues that make up the tracked object suffice). And even better, humans can track simultaneously several objects at once, although this capacity is limited to a handful of objects [39], a number that reminds to the capacity limits of visual short-term memory from section 1.2. Taken altogether, visual target tracking is remarkably robust and flexible, being able to deal with all sorts of target property changes and dynamics, uncertainties in the measurement and even periods of occlusion.

For a cognitive vision system, we target at a similarly flexible visual object tracking process, with the purpose to lock and maintain attention on an object or a part of a visual scene for a short time period. It should be able to deal

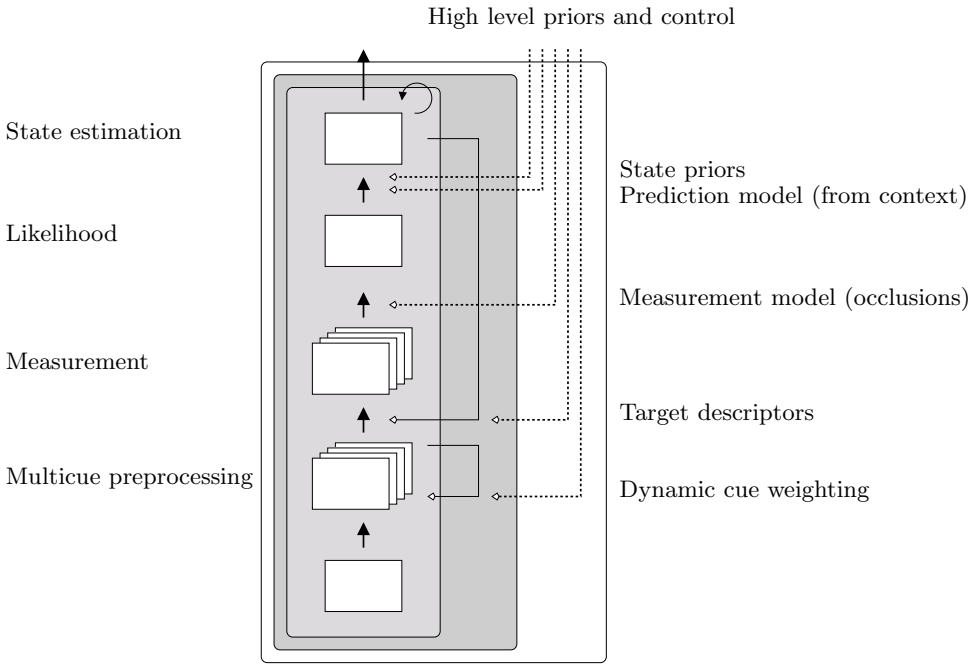

Fig. 6. Processing and control flow in a generalized item tracker. The core of a tracker is a probabilistic formulation of the prediction-measurement-update loop (circular arrow in the state estimation). To cope with a changing visual appearance of the target, an adaptation loop spans the measurement, likelihood and state estimation modules adjusting the target descriptors. Higher level priors may also influence the target descriptors, as well as the measurement and prediction models. On top of all, an item finding, engagement, tracking, evaluation and release or and reengagement loop actively binds items from visual short-term memory to the tracking process, see fig. 4 and section 2.5.

with varying visual conditions as well as asynchronous, nonregular update at low frame rates. In addition, for varying visual conditions, no single cue will be sufficiently robust to provide reliable tracking information over time, so that we have to use multiple cues for the tracking process (with a preprocessing as described for the multicue segmentation process in section 3.1). The idea is that if the cues are sufficiently complementary, there will always be at least one which can provide a tracking signal that can be exploited. For varying visual conditions, the reliability of the cues varies and some cues undergo signal jumps, but some of the remaining cue channels exhibit predictable signals that can be used for tracking.

After cue preprocessing, the fundamental problem that a tracking system has to solve is that of iterative, dynamic target state estimation. This means that it has to estimate continuously the state of a dynamic system using a series of measurements gained from an observable that can be put in relation with

the state. Fortunately, this type of problems has been extensively studied in the domain of dynamic nonlinear filtering, see e.g. [42] for a review.

For noisy states and measurements, the dynamic filtering problem can be formulated as an optimal recursive Bayesian estimator. Well-known estimators are e.g. the Kalman filter used for linear Gaussian problems (and its variants), but also techniques for the approximate numerical handling of the estimation problem, as given e.g. by the family of particle filter approaches (see e.g. [4] for an overview). For the Bayesian estimator, one attempts to construct for each timestep the posterior probability density function (pdf) of the state, taking into consideration the whole series of past measurements. The pdf contains the complete solution to the estimation problem (in a statistical sense), which means that from it, we can extract any relevant statistical estimate of the state.

During tracking, an estimate of the state is calculated every time a new measurement $\mathbf{Z}^t$ is received. This means that the filter is applied sequentially every time a new measurement becomes available, hopefully converging over time towards the solution. At every time step, only the most recent measurements are used to refine the estimate, so that the computational cost remains within acceptable bounds.

The posterior of the state given all past measurements reads

$$\rho(\mathbf{X}^t|\mathbf{Z}^t, ..., \mathbf{Z}^1) \tag{3}$$

with the present state $\mathbf{X}^t$ and the measurements $\mathbf{Z}^t, ..., \mathbf{Z}^1$ for all discrete, past timesteps $t, t-1, ..., 1$ including t.

Let us start from timestep $t-1$. We assume that the last posterior

$$\rho(\mathbf{X}^{t-1}|\mathbf{Z}^{t-1}, ..., \mathbf{Z}^1) \tag{4}$$

is known. The target is now to estimate the new, present posterior Eq. 3 by taking into account

- some additional knowledge about how the state $\mathbf{X}$ evolves over time from $t-1$ to t and
- knowledge about the measurement that is expected at time t if the system is in a state $\mathbf{X}$
- the real, new measurement $\mathbf{Z}^t$ taken at time t.

These points express formally in two stages of the filtering process, usually termed *prediction* and *update* stages. The prediction stage uses the knowledge about the systems state deployment over time to predict the expected posterior for the timestep t, i.e., it propagates the posterior from one timestep to the next without consideration of the new measurement. This type of prediction is usually coupled with uncertainty, so that it will generally spread and broaden the pdf. To the contrary, the update step uses the measurement $\mathbf{Z}^t$ to confirm and narrow the prediction. The two steps are then combined via the Bayes theorem, the prediction corresponding to the Bayesian prior and the measurement to the Bayesian likelihood used for adjusting the prior when extra information is

available. All this is a probabilistic concretization of the prediction-measurement-update steps as introduced in section 2.4.

Using knowledge about how the state $\mathbf{X}$ evolves over time from $t-1$ to t means, in a probabilistic sense, knowing

$$\rho(\mathbf{X}^t|\mathbf{X}^{t-1}) \tag{5}$$

if we restrict to a Markovian process of order one. Note that there is no dependency on the measurements/observables here, since we assume the measurement to not have any impact on the state itself. Then, Eq. 5 can be used to get (see e.g. [42])

$$\rho(\mathbf{X}^t|\mathbf{Z}^{t-1},...,\mathbf{Z}^1) = \int \rho(\mathbf{X}^t|\mathbf{X}^{t-1})\,\rho(\mathbf{X}^{t-1}|\mathbf{Z}^{t-1},...,\mathbf{Z}^1)\,d\mathbf{X}^{t-1}\ , \tag{6}$$

which is the expected posterior for time t by taking into consideration all past measurements $\mathbf{Z}^{t-1},...,\mathbf{Z}^1$, but not yet including the most up-to-date measurement $\mathbf{Z}^t$.

Similarly, using knowledge about the expected measurement for time t means to know

$$\rho(\mathbf{Z}^t|\mathbf{X}^t,\mathbf{Z}^{t-1},...,\mathbf{Z}^1)\ . \tag{7}$$

Bayes then gives us

$$\rho(\mathbf{X}^t|\mathbf{Z}^t,...,\mathbf{Z}^1) \sim \underbrace{\rho(\mathbf{Z}^t|\mathbf{X}^t,\mathbf{Z}^{t-1},...,\mathbf{Z}^1)}_{\text{Measurement likelihood}}\ \underbrace{\rho(\mathbf{X}^t|\mathbf{Z}^{t-1},...,\mathbf{Z}^1)}_{\text{Predictive prior}} \tag{8}$$

which combines the two equations Eqs. 6 and 7 to get the estimation of the new, updated posterior. (The proportionalities indicate that all the pdf's always have to be normalized.)

Ideally, $\mathbf{Z}^t$ is the complete multicue measurement. In practice, it is often assumed that the measurements are independent for each cue, so that the formalism applies for each cue likelihood independently and afterwards these can be combined. The probabilistic approach then automatically decreases the weight of the contributions of more "uncertain" cues (in terms of noisy, fluctuating). A probabilistic multicue tracking method that is robust against changes sudden changes in single cues is presented by Eggert et al in [17].

A nice property of the fully probabilistic approach is that it takes multiple simultaneous hypotheses into consideration. This implies that testing the different hypotheses is cheap - and therefore does not apply to more specialized scenarios, where a dedicated machinery has to be specialized and adapted in order to test each single hypothesis. The probabilistic framework for tracking is therefore subject to severe resource constraints, as stated in section 1, this time in terms of prediction range. In practice, the probabilistic approach only works for simple predictive models and has to be extended by further non-probabilistic adaptation loops.

Figure 6 shows the block diagram of tracking from a more general perspective. After the preprocessing of multiple cues, knowledge about the particular target is

incorporated, e.g. as a multicue template or other indication of visual properties, similarly to the region descriptors from section 3.1. This gives us the target-specific measurements which are used for the state estimation.

Control from high-level processing can be exerted at the level of the multicue preprocessing step, similarly to the segmentation case. Second, the target descriptor has to be provided and adjusted depending on an objects' appearance change. Third, the state adjustment relies on a predictive estimation that can be influenced by other visual subprocesses, e.g. including context knowledge about preferred location, velocity, rotation, etc. Forth, the dynamic state prediction model may be subject to change (consider as an example the case of a ball that is being tracked, rolling on a table surface and then reaching the border, falling down and bouncing away). Fifth, scene context information is crucial for the measurement part of the estimation, since object occlusions could be bridged by changing the state adjustment process if knowledge about occluder objects is available, by this way "explaining" situations in which the tracking fails. Context information is also necessary for the case of arising correlations between different object dynamics (e.g. as present in a hand-object coordination scene during grasping), which can be captured by modification of the prediction models (the prediction models of the interacting objects becoming entangled). Finally, higher-level modules have to either start the tracking engagement by presenting object hypotheses and starting conditions, finish the engagement when tracking fails, or organize for reengagement if the object should be kept in the processing focus of the system.

As argued before, we postulate that a suited representation in visual working memory that has access to the multiple adjustment loops and serves to control the tracking processes has to be established. This representation would then couple to other processes that demand a sustained focus on an object or the estimation of an objects' parameters as delivered by the tracking process, as well as to superordinate processes that organize the distribution of tracking resources on hypothetical objects of interest and the creation and destruction of tracking subprocesses.

It remains to be stated that an entire plethora of visual tracking approaches exist, depending on one hand on the types of representations that are used for representing the objects and on the other hand on the complexity of the appearance changes to be captured. For technical systems, many tracker work in constrained environments, like high input frame rates (resulting in very simple or nearly linear appearance changes, as assumed by KLT or Mean-Shift based differential tracking systems, see e.g. [31, 44, 13]) or a stationary background against which changes can be easily extracted. Here, we wanted to highlight control issues in trackers that work with large appearance changes, low frame rates, asynchronous update or measurement and sporadic and selective tracker engagement; controlled for a dedicated visual task within a specific visual scene and visual item memory context.

4 Learning in cognitive vision systems

The human visual system is not fixed and static over time, but a rather flexible and adaptive system that is subject to a wide variety of learning mechanisms. Learning occurs at different time scales, ranging from minutes up to life-long time spans, required to obtain proficiency in specialized, visually dominated tasks. Flexible and autonomous learning is also a key ability that distinguishes human visual capabilities from current state-of-the-art technological solutions in machine vision and object recognition.

In the following we first describe our approach to the main general principles that underlie the buildup of a task-driven behaviourally relevant visual knowledge representation. We then concentrate on the two issues of perceptual learning, operating on rather short time scales, and already realized concepts of integrated visual attention and object category learning.

4.1 Self-referential Buildup of Visual Knowledge Representations

The learning processes starting from the initial perceptual capabilities of an infant after birth up to the later development of specialized visual proficiency are an example for a complex knowledge acquisition process that requires the coordinated interaction of several areas of the brain. As an approach to understand this, the concept of self-referential learning [29] has been proposed, emphasizing the autonomous and active character of any brain-like knowledge acquisition process and the prevalence of task-driven and behaviorally relevant learning. Both aspects together ensure the consistent buildup of a visual knowledge representation that is useful for a living being, that has to survive in a dynamically changing and unpredictable environment.

Any higher-level biological cognitive system faces the challenge, that its development is to a large part determined by the interaction with its surroundings. Here the feedback from the environment rarely provides explicit teaching signals that have the quality of the supervised learning paradigm of neural networks or machine learning. The system rather has to rely on its own internal dynamics in determining the buildup of meaningful visual categories and evaluating their success in the interaction with the environment. Körner and Matsumoto [29] have emphasized the importance of a subjective stance towards this acquisition process, defining a "self", determined by a value system and guiding the learning process by emotions and resulting attentional biases. This value system strongly relates to phylogenetically older substructures of the brain, where especially the limbic system plays a crucial role. An important concept in self-reference means that the already acquired representational system strongly constrains the perception, evaluation, and thus value-driven acquisition of new knowledge. This filtering of new knowledge based on existing representations ensures the overall consistence of newly stored information in relation to the prior knowledge.

The reference to a value-system for guiding the acquisition of meaningful representations provides a direct link to the importance of behavior and task-related concepts for the learning of visual representations. According to this approach,

the formation of visual categories is generally done in reference to a task and behavioral constraints. This provides a strong coupling between action and perception, being a key ability of biological intelligent systems that has proven notoriously difficult to achieve in technical systems. A good example for such a representation in the field of vision are action-related object representations of manipulable man-made objects, that have been found in the dorsal brain areas [12].

4.2 Perceptual Learning

Perceptual learning has been defined as an acquired change to a perceptual system to improve its ability to respond to the environment [23]. The corresponding time spans range from minutes up to days and weeks and the effects of perceptual learning can be quite long lasting. This type of learning adaptation has been contrasted against cognitive learning processes in the way that it applies to the perceptual or pre-attentive processes, that are beyond conscious accessibility. In an attempt to categorize different mechanisms of perceptual learning, Goldstone [23] has distinguished the following mechanisms:

- *Attentional weighting* modifies the relevance of visual features in a particular task context
- *Imprinting* introduces new, specialized features in a perceptual situation
- *Differentiation* separates previously indistinguishable features
- *Unitization* merges separate features into greater units to ease perception of such compounds

Perceptual learning is generally assumed to modify rather the early stages of cognition and is thus prior to high-level reasoning. The perceptual effect can, however, deeply influence higher areas by influencing the feature representations that are the basis of higher level concepts. This low-level property is highlighted by the limits of generality of this form of learning. Training on simple visual discriminations often does not transfer to different eyes, to different spatial locations, or to different tasks involving the same stimuli [20].

Although perceptual learning is an ubiquitous phenomenon in biological vision systems, almost all current computer vision systems lack this basic capability. The reason is that there is no general concept available that could deal with the resulting plasticity in such a system. A simple example is an object classifier that is trained by supervised learning on the output of a feature detection stage. Most current classification models assume a static feature representation and cannot handle an incremental and dynamic input stage. In a recent contribution Wersing et al. [53] have investigated a model of coupled learning between the "what" and "where" pathways for the bootstrapping of the representations for localizing and classifying objects. This establishes first steps towards modular cognitive vision systems where parallel learning in different modules can occur without destabilizing the robustness of the system.

4.3 Visual Category Learning and Object Attention

Main problems. The processes involved in visual categorization are generally considered more on the high-level or cognitive side of perception. Nevertheless it is obvious, that sensing and learning of object classes is strongly dependent on phenomena of attention, expectation, and task-driven utility. In creating a visual system with an autonomous strategy for learning visual concepts, the following questions have to be answered:

- *What and where* do we have to learn ?
- *When* do we have to learn ?

The first question is related to the ability of a learning visual system to attend to particular parts in the scene that offer some saliency that can be both bottom-up and top-down driven. In general an autonomously learning system requires an active strategy for selecting elements within a visual scene that are both interesting and can be robustly separated from the distracting surroundings. It should be one of the main targets of a cognitive vision systems to relax the strong segmentation constraints that are currently necessary for many computer vision approaches to supervised learning of object categories. Segmentation in this framework should be rather a form of attention that is mainly top-down driven by the prior knowledge of the system. In the human visual system there exists a clear functional differentiation in the processing of object identity ("what") and object positions ("where") [56].

The second question is related to the temporal coherence and stability of the learning process. An autonomously learning system cannot rely on an explicit teacher signal that triggers the start and end of a learning phase. It rather needs intrinsic signals that characterize points in time where learning is feasible, based on an internally driven detection of learning success. Prediction is one of the main concepts that can be used to estimate the feasibility of learning in a particular scene context. For prediction of the sensory input, it is necessary to produce generative models, that are capable of reproducing the relevant visual structures of real objects. We are here, however, not mainly concerned with the temporal aspect of prediction, but with prediction in the sense of the ability of the system to effectively represent an externally changing input using its internal representations. This is normally referred to the concept of deriving a generative model for the considered stimulus domain. To make this autonomous learning feasible, apriori information on relevant structural constituents of objects can be useful.

Related Work. The questions of attention-based learning and object isolation in real-world scenes have been investigated by a number of recent contributions:

Shams & von der Malsburg [43] considered the autonomous learning of visual shape primitives in an artificially generated setting with rendered scenes containing geon components. Using a correlation measure based on Gabor jet feature representations they manage to derive simple constituents of the scene. The scaling to more complex real-world scenes, however, was not yet considered.

Williams & Titsias [54] have proposed a greedy learning approach of multiple objects in images using statistical learning in a generative model setup. Their approach is based on a predefined sequence of learning steps. First the background is learned, then the first object, and subsequently more objects. The representation is based on a mask and a transformable template. A limitation is that the method can only register a single pose of a particular object. In a similar Bayesian generative framework, Winn & Jojic [55] use their LOCUS model for the learning of object classes with unsupervised segmentation. Additionally they can handle stronger appearance variation among the members of the learned class, i.e. color and texture.

Walther et al. [50] investigate the usage of bottom-up saliency for determining candidate objects in an unsupervised way in outdoor and indoor scenes. For each frame of a video sequence such candidate objects are determined offline, and represented using the SIFT feature approach developed by Lowe [30]. Matching objects are determined between pairs of frames, and compared to a human labeling of the objects in the scene. The saliency-based segmentation improves the matching performance and the system is robust with regard to scaling and translation, but not very good at representing 3D rotation and multiple poses of objects.

An interesting approach to supervised online learning for object recognition was proposed by Bekel et al.[10]. Their VPL classifier consists of feature extraction based on vector quantization and PCA and supervised classification using a local linear map architecture. They use a bottom-up saliency coupled with pointing gestures in a table setting to isolate objects in the scene. The segmentation method is similar to the one in [50].

Arsenio [3] uses an active perception model for object learning that is using motion-based segmentation, sometimes even induced by robot actions. The object representation is based on hashing techniques that offer fast processing, but only limited representational and discriminatory capacity.

Itti [25] develop a general theory of attention as Bayesian surprise. In their approach, surprise is quantified by measuring the difference between posterior and prior beliefs of the observer.

Attention for Online Object Learning. Wersing et al. [52] have presented a biologically motivated architecture for the online learning of objects and people in direct interaction with a human teacher. Their system combines a flexible neural object recognition architecture with an attention system for gaze control, and a speech understanding and synthesis system for intuitive interaction. A high level of interactivity is achieved by avoiding an artificial separation into training and testing phase, which is still the state-of-the-art for most current trainable object recognition architectures. They do this by using an incremental learning approach that consists of a two-stage memory architecture of a context-dependent working or sensory memory and a persistent object memory that can also be trained online.

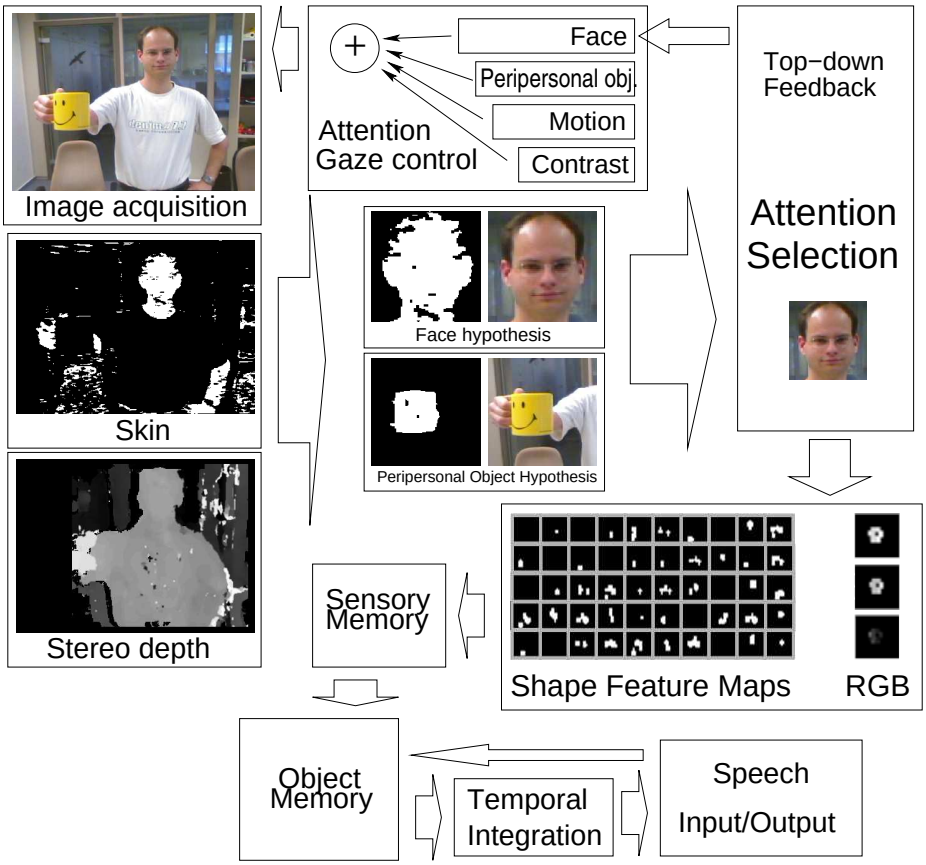

Fig. 7. Overview of the visual online learning architecture by Wersing et al. The system combines feature-based bottom-up saliency with top-down expectations on faces to be learned. Objects and faces are then incrementally learned in a unified shape feature map representation using short-term and long-term memory.

They use a stereo camera head mounted on a pan-tilt unit that delivers a left and right image pair for visual input (see Fig.7). The gaze is controlled by an attention system using bottom-up cues like edge/color/intensity contrast, motion, and depth, presented in more detail in [22]. Additionally top-down information on face targets is provided to be followed with a peaked map at the detected face position. Each cue is represented as a retinotopic activation or saliency map. A simple addition of the different cues is used, where clear priorities are induced by weighting the cues in the following sequence: contrast $<$ motion $<$ depth $<$ face. This simple model enables a quite complex interaction with the system to guide the attention appropriately.

The default state of the gaze selection system is an exploratory gazing around that focuses on strong color and intensity contrasts. Moving objects attract

more attention. An even stronger cue is generated by bringing an object into the peripersonal space, that is the near-range space in front of the camera that corresponds to the manipulation space of a humanoid robot [22]. However, also the weaker cues of contrast give a contribution and stabilize the attention. The strongest cue is the presence of a detected face, generating a strong task-specific attention peak at the detected position.

To trigger the online learning and recognition, two parallelly computed object hypotheses are used. Firstly, objects are learned and recognized, if they are presented within the peripersonal space. The object is attended, as long as it resides within this interaction space. Secondly, using skin color segmentation, candidate region segments are classified according to their face similarity. An accepted face region is then selected and processed using the same online learning and recognition pathway as for objects. The attention is retracted from the face, if no valid face-like segment was detected near the image center for two input frames.

The system of Wersing is capable of learning and robust recognition of several objects and face categories, as was shown in [52]. The interaction between the attention system and the object learning, however, is manually designed and not dynamic with regard to the selected feature modalities like stereo or face shapes. The implementation of dynamic mechanisms and learning principles also for this part of the system will be an important future step to ensure stronger autonomy of such online learning visual systems.

5

6 Conclusions

During the last years, considerable progress has been made for single visual processes, as it is also the case for the examples presented here: Segmentation, tracking and appearance-based object classification and learning. Nevertheless, in a real-world scenario these processes have to be constrained by sensible assumptions to make the problems tractable. It is likely that no general-purpose solution exists for any of them without severe constraints, a dilemma that may be shared with biological vision systems.

This means that we are confronted with the principled problem that a number of visual subprocesses has to be organized, constrained, adapted and arbited *dynamically*, even for simple, brief visual tasks. As a consecuence, visual subprocesses have to be approached and designed in a substantially different way than in classical computer vision. This paper presents a proposal on how the organization of visual subprocesses could be achieved.

Where should information about how to constrain and adapt visual subprocesses come from in a first place? In essence, visual and long-term memory can store a large amount of specific priors and context knowledge which may be recalled to tune processes to particular scenarios, object categories and objects.

The main role of the control processes is to bring together different types of internal knowledge - long-term assumptions about the world and its items, short term scene and object memory, and process knowledge about its own internal adaptation processes, limitations and capabilities - to actively steer the acquisition of information. Because of limited processing resources, this occurs selectively, and on demand. It is about anchoring in the broadest sense, but with dynamically changing configurations of the memory representations which are being bound to sensory events and only when required for a visual task.

So far we have not discussed concrete realizations of the memory structure itself. What would be a minimal set of representational properties of a memory that is capable to serve as knowledge basis for control processes, and that can be temporarily entangled with selected sensory measurements? How feasible is the idea of a visual memory that gathers information about items, objects and scene properties?

The experimental evidence about peoples inability to detect severe visual changes does not seem to support the idea of a persistent dedicated visual memory. It rather suggests that "visual representations may be sparse and volatile, providing no cumulative record of the items in a scene" [6]. However, most of the studies do not take special care of attention, so that it may be that the visual system still builds a cumulative record of all attended stimuli and still miss all changes involving items that were not attended. Here we reencounter the resource limitation argument, both in terms of memory access and a bottleneck in attentional resources, since attended items require exclusive resource allocation. Visual memory may therefore store just what is necessary and what was accessible with limited access resources to visual subprocesses, rendering control processes deciding on what to focus the visual resources on even more important. This applies both to visual short-term memory as well as for consolidation processes during visual exploration, as introduced in section 4.3: What, where and when to learn.

Visual scene representations must therefore provide a substrate on which these issues can be taken into account. It is selectively impoverished (accumulating only sparse and incomplete visual information) and volatile (referring to short-term visual memory with its limited and temporary anchoring capabilities to sensory events), but it has to provide interfaces to control structures and control processes and to the different types of information extracted by the different visual subprocesses and modalities. First attempts to couple a sparse, relational visual memory with a simple visual system are presented in [26, 40]. A principled approach to integrate memory in form of priors, contextual and process information with dedicated control structures that tune visual subprocesses is however an open - yet fundamental - research topic for cognitive vision.

References

1. J. Aloimonos. Purposive and qualitative active vision. In *Proc. 10th Int. Conf. Patt. Recog.*, pages 345–360, June 1990.

2. Y. Aloimonos. Active vision revisited. In *Active Perception*, 1993.
3. A. Arsenio. Developmental learning on a humanoid robot. In *Proc. Int. Joint Conf. Neur. Netw. 2004, Budapest*, pages 3167–3172, 2004.
4. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. *IEEE Trans. Signal Processing, pages 100 107,2001*, 2001.
5. B. J. Baars. Metaphors of consciousness and attention in the brain. *Trends in Neuroscience*, 21(2):58–62, 1998.
6. M. Backer and H. Pashler. Volatile visual representations: Failing to detect changes in recently processed information. *Psychonomic Bulletin and Review*, 9:744–750, 2002.
7. A. D. Baddeley and G. J. Hitch. Working memory. In G.A. Bower, editor, *Recent Advances in Learning and Motivation*, volume 8, page 47. New York, Academic Press, 1974.
8. L. W. Barsalu. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.
9. C. Bauckhage, S. Wachsmuth, M. Hanheide, S. Wrede, G. Sagerer amd G. Heidemann, and H. Ritter. The visual active memory perspective on integrated recognition systems. *Image and Vision Computing*, 26(1), 2008.
10. H. Bekel, I. Bax, G. Heidemann, and H. Ritter. Adaptive computer vision: Online learning for object recognition. In *German Pattern Recognition Symposium*, pages 447–454, 2004.
11. T. Chan and L. Vese. Active contours without edges. 10(2):266–277, February 2001.
12. L. L. Chao and A. Martin. Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12(4):478–484, 2000.
13. D. Comaniciu and P. Meer. Mean shift analysis and applications. In *International Conference on Computer Vision*, pages 1197–1203, 1999.
14. Silvia Coradeschi and Alessandro Saffiotti. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96, 2003.
15. N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–185, 2001.
16. J. Eggert D. Weiler, V. Willert and E. Koerner. A probabilistic method for motion pattern segmentation. In *Proceedings of the IJCNN 2007*, 2007.
17. J. Eggert, N. Einecke, and E. Koerner. Tracking in a temporally varying context. In H. Tsujino, K. Fujimura, and B. Sendhoff, editors, *Proceedings of the 3rd HRI International Workshop on Advances in Computational Intelligence*. Honda Research Institute, Wako, Japan, 2005.
18. C. W. Eriksen. Attentional search of the visual field. In B. David, editor, *International Conference on Visual Search*, pages 3–19, 4 John St., London, WC1N 2ET, 1988. Taylor and Francis Ltd.
19. C. W. Eriksen and J. D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Percept. Psychophys.*, 40:225–240, 1986.
20. M. Fahle and M. Morgan. No transfer of perceptual learning between similar stmuli in the same retinal position. *Current Biology*, 6:292–297, 1996.
21. G.Metta, G. Sandini, and J. Konczak. A developmental approach to visually-guided reaching in artificial systems. *Neural Networks*, 12(10):1413–1427, 1999.
22. C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn. Peripersonal space and object recognition for humanoids. In *Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2005), Tsukuba, Japan*, 2005.
23. R. L. Goldstone. Perceptual learning. *Annual Review of Psychology*, 49:585–612, 1998.

24. Grossberg, Stephen, Hong, and Simon. A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision*, 19(2-4):263–321, 2006.
25. L. Itti. Models of bottom-up attention and saliency. In L. Itti, G. Rees, and J. K. Tsotsos, editors, *Neurobiology of Attention*, pages 576–582. Elsevier, San Diego, CA, Jan 2005.
26. S. Rebhan J. Eggert and E. Koerner. First steps towards an intentional vision system. In *Proceedings of the International Conference on Computer Vision (ICVS) 2007*, 2007.
27. M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal for Computer Vision*, 1(4):321–331, January 1988.
28. J. Kim, J. W. Fisher, A. J. Yezzi, M. Çetin, and A. S. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. In *International Conference on Image Processing, Rochester, New York*, volume 3, pages 797–800, September 2002.
29. E. Körner and G. Matsumoto. Cortical architecture and self-referential control for brain-like computation. *IEEE Engineering in Medicine and Biology*, 21(5):121–133, 2002.
30. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
31. Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
32. D. Marr. *Vision*. Freeman, San Francisco, 1982.
33. H. Maturana and F. Varela. *The Tree of Knowledge - The Biological Roots of Human Understanding*. New Science Library, Boston, 1987.
34. D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math*, 42:577–685, 1989.
35. V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages II: 2049–2056, 2006.
36. B. Neumann and R. Moller. On scene interpretation with description logics. *Image and Vision Computing*, 26(1):82–101, January 2008.
37. S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *J. Cmpt. Phys.*, 79:12–49, 1988.
38. Z. W. Pylyshyn. The role of location indexes in spatial perception: A sketch of the FINST spatial index model. *Cognition*, 32(1):65–97, June 1989.
39. Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3:179–197, 1988.
40. S. Rebhan, F. Roehrbein, J. Eggert, and E. Koerner. Attention modulation using short- and long-term knowledge. In *Proceedings of the International Conference on Computer Vision (ICVS) 2008*, 2007.
41. R. A. Rensink, J. O'Regan, J. Kevin, and J. J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8(5):368–373, 1997.
42. Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter*. Artech House, 2004.
43. L. Shams and C. von der Malsburg. Acquisition of visual shape primitives. *Vision Research*, 42 (17):2105–2122, 2002.
44. Jianbo Shi and Carlo Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, June 1994.

45. G. Sperling. The information available in brief visual presentations. *Psychological Monographs: General and APllied*, 74(11):1–30, 1960.
46. A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14:107–141, 1982.
47. J. Triesch, D. H. Ballard, M. M. Hayhoe, and B. T. Sullivan. What you see is what you need. *Journal of Vision*, 3(1):86–94, 2003.
48. T. van Gelder and R. F. Port. It's about time: An overview of the dynamical approach to cognition. In R. F. Port and T. van Gelder, editors, *Mind as Motion - Exploration in the Dynamics of Cognition*, pages 1–43. Bradford Books, MIT Press, Cambridge, MA, 1995.
49. D. Vernon. Cognitive vision: The case for embodied perception. *Image and Vision Computing*, 26:127–140, 2006.
50. D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, 100(1-2):41–63, October 2005.
51. D. Weiler and J. Eggert. Segmentation using level-sets and histograms. In Springer, editor, *Proceedings of the International Joint Conference on Neural Networks (ICONIP) 2007*, 2007.
52. H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J.J. Steil, H. Ritter, and E. Körner. Online learning of objects and faces in an integrated biologically motivated architecture. In *Proc. ICVS, Bielefeld*, 2007.
53. H. Wersing, S. Kirstein, B. Schneiders, U. Bauer-Wersing, and Edgar Körner. Online learning for boostrapping of object recognition and localization in a biologically motivated architecture. In *Proc. Int. Conf. Computer Vision Systems ICVS. Santorini, Greece.*, 2008.
54. C. K. I. Williams and M. K. Titsias. Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5):1039–1062, May 2004.
55. J. Winn and N. Jojic. LOCUS: Learning object classes with unsupervised segmentation. In *ICCV05*, pages I: 756–763, 2005.
56. S. Zeki. Localization and globalization in conscious vision. *Annual Review Neuroscience*, 24:57–86, 2001.
57. S. C. Zhu and A. L. Yuille. Region competition: Unifying snakes, region growing, and bayes/MDL for multiband image segmentation. *PAMI*, 18(9):884–900, 1996.