First steps towards an intentional vision system

Julian Eggert, Sven Rebhan and Edgar Koerner

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

Abstract. Contrary to many standard vision systems which proceed in a cascaded feedforward manner, imposing a fixed order in the sequence of visual operations like detection preceding segmentation and classification, we develop here the idea of a vision system that flexibly controls the order and accessibility of visual processes during operation. Vision is hereby understood as the dynamic process of 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. In this paper we present the concept and the necessary main ingredients and show first steps towards the implementation of a real-time intentional vision system.

1 Motivation

The selective choice of sensory information and the corresponding tuning of sensor parameters, i.e., the focusing of visual processing resources, is a fundamental problem of any artificial or biological vision system with at least a claim for some minimal generality. 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.

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 sensory-near measurements.

To the contrary, in the first attempts to outline a computational theory of vision, Marr stated that the goal of e.g. 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" [8]. This

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

reconstruction/recovery paradigm did not make any explicit reference to internal modulation and focusing processes so that it suggests a task-independent passive observer. For underconstrained real-world vision processes, it was soon clear that an accurate description of the 3D-surrounding is largely impossible. Furthermore, the claim of a detailed internal world representation does not take into account system processing constraints, like e.g. those showing up when several competing visual object become involved and attention has to concentrated on one of them. All this together suggests that a consistent maintenance over time of such a accurate internal representation is computationally unfeasible.

As a consequence, behaviorist paradigms appeared, 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. [3]), 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* ([2]) regards vision processes always in combination with the context of some tasks that should be fulfilled. Common to both approaches is that they have concentrated on behaviors and actions that are directly observable from outside, and in how visual information can be extracted that supports particular behaviors.

In the intentional vision framework we present here, visual cognition is understood as a goal-driven mediation between an internal representation and the incoming sensory stimulation. The mediating control processes thus serve to gather visual information that could be potentially used for guiding overt behaviors, (without necessarily being tied to the behaviors). In fact, we interpret any internal modulation and attentional focusing as a (virtual) action. [1] 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. "pick 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".

In the next two sections, we will describe our first steps towards building an intentional vision system which uses an intermediate level of representation of the visual world that incorporates task and intention based components. We will elaborate the necessary ingredients that such a system should have, and then focus on a few visual subtasks that we have chosen for a first implementation basis. In a further section, we describe the model and results gained from the implementation of these subtasks.

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

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

2 Ingredients for intentional vision

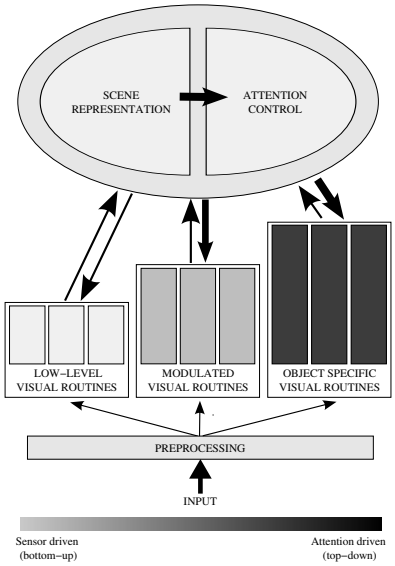

Fig. 1. Architecture overview of the intentional framework as targeted in this work. After a preprocessing step which may not be controlled by the task via attention, the visual routines are all driven as a combination of bottom-up and top-down factors (more sensory-driven routines are displayed on the left side, more attention-driven routines on the right). A strong parallelism of the visual routines avoids any fixed order of operations, rather, the scene representation combines and sequentializes the routines in a memory- and task-dependent way, controlling and routing the information flow.

Contrary to systems with strict hierarchical representational structures a la Marr, an intentional vision system as devised here, has hardly any really "passive" sensing components besides perhaps some rudimentary preprocessing that can be decoupled from the rest of the system. All the components that contribute to the process of acquisition of visual information are active in that they are selectively modulated, controlled and triggered resp. activated/deactivated. The idea is that e.g. visual object classification is not a result of a fixed preprocessing-segmentation-classification "path". Instead, we break up the fixed sequence and consider segmentation, masking and classification as possibly interdependent, but largely separate visual processes with each delivering its result to a higher-level representation, whose role is to drive them in the right order until the overall task is complete. This introduces a special kind of flexibility into the system, since for different contexts the higher level representation could decide to modulate the visual processes differently, e.g. by using a segmentation based on different cues, or to use a combined mask that includes some prior information about the object in question, etc.

The processes needed for an intentional vision system can be coarsely systematized into the *visual routines* responsible for providing measurements about the visual scene, *visual modulation* processes that allow for a task-dependent specialization of the visual routines, providing the capability to incorporate top-down assumptions, *visual control* mechanisms that decide on the visual task decomposition, driving the visual routines in an appropriate way, and the *visual*

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

representation that compresses the short-term memory of visual measurements delivered by the visual routines and the memory of the attended items.

The active character of all these processes has to be considered already in the design process of the system. This means that we have to provide means so that other visual routines and the acquired visual knowledge can modulate the behavior of a visual routine in a systematic way, since every routine is supposed to be used in more than a single configuration resp. more than a single visual task (usually composed of a concatenation of several visual routines). In addition, we seek for visual routines that are general enough to be used for multiple visual tasks and that can appropriately adapt their operation to a different visual context. From the technical side, we have to provide the interfaces so that the different visual routines can communicate with each other and with the visual short-term memory.

Figure 1 shows the main parts of an intentional vision system. After a preprocessing stage which may be common to all visual routines but not actively controlled by the visual task(s), there is a plethora of visual routines subserving different purposes, with rather direct connection to the visual short-term memory, representation and control components. The visual routines are neither exclusively dominated by bottom-up or top-down processing, rather they span a broad spectrum, with some of them being mainly sensory-driven (Fig. 1, left side) and only broad parameter modulation on a long time scale by the visual context, some being modulated on a short timescale by the visual short-term memory (middle), and some which are very object-, location- or feature-specific, depending on top-down request for operation (right side).

3 Specific visual routines

In this section we shortly outline the visual routines that we are using for first implementation tests. From a systematic point of view, we have used four broader classes of visual routines (in fact there may be many more with several distinct processing pathways for extraction of different properties of the visual stimulus, compare this with biologically inspired models which usually assume a "what" and a "where" pathway): Visual routines for *triggering point hypotheses* about item location in the visual field, for the *dynamic prediction and confirmation* of the items position over time, for the extraction of *area information and figure-background segmentation*, and for *area- and segmentation-based measurements*.

3.1 Hypothesis triggerering via tunable saliency

One of the main visual subtasks in a visual system is to discover and localize potentially interesting parts of a visual scene. Such parts may serve as initial hypotheses for more detailed visual inspections, e.g. to identify "objects". [2]

[2] With objects, we denote here not necessarily the visual appearance of physical objects present in the real world, but rather things of the visual scene that deserve a special internal representation because they can be characterized by a low-dimensional state descriptor like e.g. position, size and classification label.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Usually, the argument is that the selection of potentially interesting locations of a scene should be sparse, i.e., extracting only few interest points as compared to the input image resolution, and fast, e.g. by using simple, parallelizable operators. Computational costs of subsequently analyzing stages like object detection and recognition could then be kept at a more manageable level.

In our system, the selection of interesting locations is accomplished by a series of saliency-computing modules, which calculate spatiotemporal contrasts at various scales for different visual cues like motion, color and structure (see [7, 6] for a review). Circular spatial center-surround contrasts reveal locations of the visual input where a visual cue deviates from average. After the center-surround calculations for each visual cue, neural field relaxation dynamics (see e.g. [4]) provide a regularization of the results so that the saliency results from the different cues can be integrated into a single spatial saliency map from which interesting points are selected.

Although saliency is usually understood as being low-level and working mainly in a feedforward manner, it can be guided and modulated to a large extent by top-down factors (see e.g. [13, 9] for recent articles on this topic). In our system, top-down guidance enters at various levels. On the most general level, the contribution of each cue to the collector saliency map is weighted by a gain factor that is controlled by the scene representation. By this way, depending on the task context, different desired visual properties can be emphasized (e.g., for dynamical scenes, concentrate on salient points determined by motion). In addition, for each saliency cue spatial modulation maps are provided, which can be used to bias spatial areas of the visual input. In such a way we can e.g. specify a target search that uses color contrast for the upper and motion contrast for the lower half of the input. Furthermore, each saliency cue is composed of several "feature" subchannels, which can again be weighted using a gain factor. In such a way, the system can focus on selected features of a saliency cue, e.g. preferring horizontal over vertical motion for the motion saliency contribution. Last but not least, we are using a number of saliency contributions with cues that are completely object- and context-specific ("target color cue", "target depth cue", ...), by calculating the difference of the sensory input from an expected cue configuration, such as a specific hue and saturation.

After the calculation of the collector saliency map, the extracted points of interest are compared with the visual short-term memory to see if they correspond to already stored "object hypotheses". If not, they are stored in memory together with their cue activities and current cue weighting factors, so that the knowledge about which cues triggered a hypothesis is retained. Such knowledge can then be used in later stages to direct a top-down driven search for the object if its location has become uncertain, or to specify cues for e.g. segmentation processes.

Summarizing, the results from the dynamically tunable saliency serve as a basis for the generation of entry-level target location hypotheses in short-term visual memory. These hypotheses are then used for other search and refinement processes that recruit further visual routines.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

3.2 Multitarget tracking

We also implemented a view-based, 2D Bayesian multicue tracker ([1]) to be able to lock and maintain attention on an object or a part of a visual scene over time. This is important for an intentional vision system with an internal object representation, since it allows to dynamically focus and do processing on the object. E.g., while tracking an object we can call classifier routines from time-to-time and check the objects identity, or use visual routines that require an integration over time, etc.

Design issues here were that we do not have to learn or train the tracker specifically for each object. Since we want the system to be sufficiently general to work with as many types of objects as possible, and since the tracker has to be able to get rapidly initialized by hypotheses provided either by the saliency selection (3.1) or by the visual memory, we used an appearance-based tracker that extracts multicue templates using a coarse initial hypothesis about the objects position and extension. The tracker itself is based on the assumption of temporal continuity in one or several of the cues. For contexts in which the tracked objects undergo considerable variations in their visual aspect (e.g., changes in color, shading, reflection and form), such a multicue approach is beneficial in terms of robustness and flexibility.

In our system, we included several tracker working in parallel so that multiple targets can be tracked independently and simultaneously. The tracked objects are represented in visual memory together with their positions and further tracking parameters (templates, cues). Similarly to the saliency-driven hypotheses, they provide a target location measurement, with the difference that they are very object-specific and that their state (position, velocity) and their link to the sensory input is continously updated.

3.3 Attention-mediated segmentation

In addition to the triggering of new hypotheses via saliency (Sec. 3.1) and the attending of objects via multicue tracker (Sec. 3.2) which both provide spatial point information (i.e., *where* an object is located), it is necessary to also get surface-based information, as e.g. provided by a figure-ground segregation process. Surface-based information available in the visual memory can be used by several complementary visual routines, e.g. using it as a mask for an object classifier, or by computing mask-based measurements on cues to calculate constrained statistics such as average color, structure or motion.

As in the tracking routines, we want segmentation to be applicable using multiple cues which can be weighted dynamically and selectively using top-down signals from the scene memory and the task context. Furthermore, since segmentation implies computationally expensive iterative optimization algorithms we have to cope with the resource constraints given by a dynamical scene.

In our system, object segmentation is based on level-set methods that sit on back of the tracker routines. This implies that a subset of all tracked objects (with the subset selection taking place in object memory) is chosen for segmentation,

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

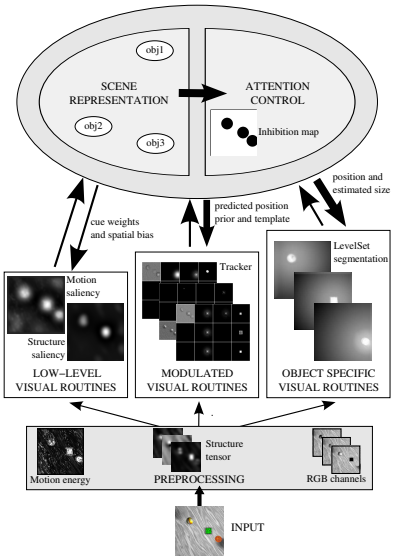

Fig. 2. Simulated system with 3 saliency routines (left side, structure, motion and color saliency), tracker for multicue tracking, and level-set based segmentation for figure-ground segregation of the tracked objects. The scene representation acts as a "blackboard" that combines the result of the visual routines and organizes the information flow. In our system, the saliency routines trigger hypotheses which can then be combined dynamically with their corresponding tracker and segmentation routines. An inhibition map (top right) prevents the triggering of new hypotheses at the locations of tracked objects.

that the segmentation occurs iteratively over time, and that it is calculated along with the moving objects, incorporating the predictions of the tracker.

Level-set methods for segmentation (see e.g. [11, 5, 10] for a survey) are based on a functional that explicitly describes the segmentation criteria (cue specificity and homogeneity of inside and outside areas, contour properties like length and curvature, etc.) of the searched area. The functional optimization occurs by deriving dynamics for a level-set surface function, which implicitly describes the area and contour of the segmented region. The dynamics serve to modify the level-set surface function, until a local optimum is found. Important for level-set based segmentation is the choice of the initial conditions, which have to be sufficiently close to the desired result.

To couple the segmentation with multicue tracking, we take the hypothetical object position from a tracker and use it to create a surface prior as initial condition, e.g. using a circular segmentation mask. Then we start to iterate the level-set functional for a few steps. The cues on which the functional is calculated can be extracted from the saliency cue information of the object that served to trigger the tracking target. While the target is being tracked, we shift the level-set surface along with the predictions of the tracked object, and iterate a few steps at each position. This allows the system to preform a figure-ground segregation under realistic timing constraints in a dynamic scene configuration.

4 Results

Here we show simulation results of a system that we set up as a proof-of-principle for testing our ideas about intentional vision. We concentrated on dynamical

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

scenes with appearing and disappearing objects, since then we have to deal with instantiation and destruction of object hypotheses in visual short-term memory.

4.1 Saliency-tracking-segmentation loop

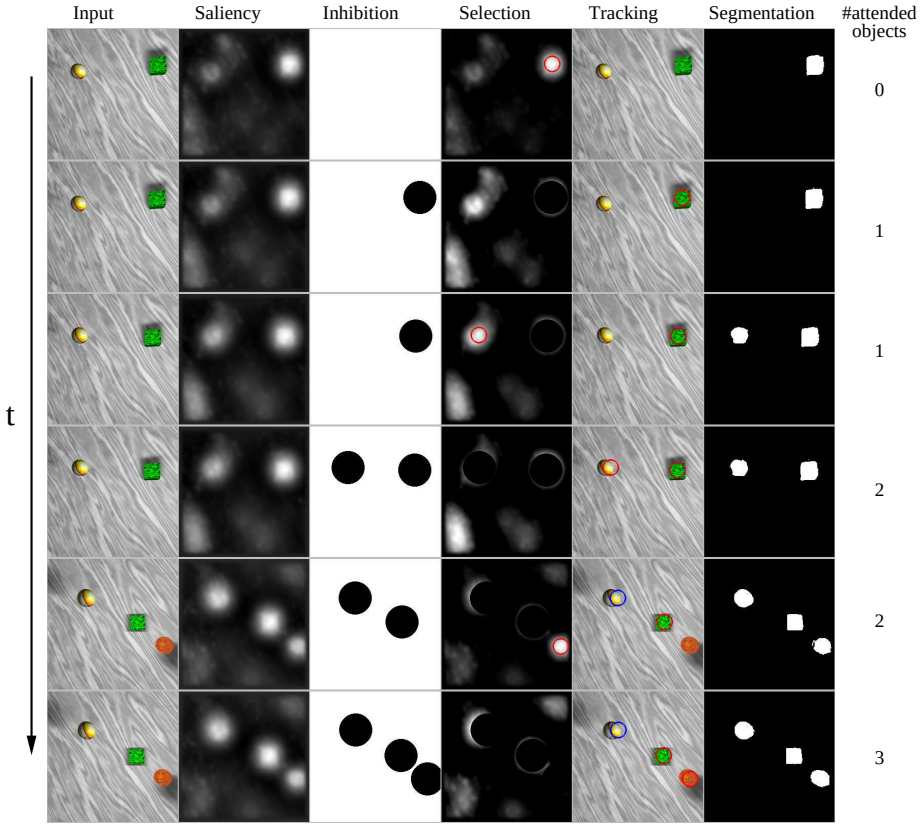

Fig. 3. Our system working on a synthetic image sequence. Time runs from top to bottom. Shown are from left to right the original input, the contrast-based saliency, the inhibition of attended objects, the newly selected objects for tracking, the tracker estimations for the objects positions, and the segmentation masks for tracked objects. New tracker are instantiated and "glued" onto objects depending on the already tracked objects in short-term memory and the saliency of a new object appearing in the scene.

We coupled multiple saliency modules, target tracker and level-set segmenting routines with a visual short-term memory module in a dynamical way. The memory module keeps track of saliency-triggered hypotheses, tracked objects and the segmentation information. The lowest level object hypotheses are delivered by the saliency modules. A subset of these is dynamically selected for

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

tracking; in the simulations, the decision is based on the motion contribution to the saliency (corresponding e.g. to an implicit visual task "detect, follow, segment and classify all moving objects in the scene"). From the tracked objects, again a subset is selected and attached to level-set-based segmentation algorithms that calculate a mask that encompasses the object. Segmentation occurs while the objects move and are being tracked, as explained in section 3.2.

We used saliency cues to extract motion, color and intensity-based structure contrast. The local motion estimations were extracted using a windowed, normalized correlation-based probabilistic measurement technique according to [12]. The local intensity-based structure was determined calculating the spatially averaged structure tensor previous to the saliency contrast calculation at various spatial scales. The structure tensor saliency turned out to be quantitatively as good as but more effective than using various Gabor filters. The color saliency estimations were computed directly on the RGB channels of the input. In all three cases, the "subfeatures" (motion estimation for the different velocities, structure tensor components, RGB components) are regarded as collected into a feature vector / feature histogram, and the saliency contrast measures the difference of the spatially averaged feature vectors between center and surround.

The objects in visual memory interact with the triggering of new saliency-based hypotheses using a spatial inhibition map that is build at each timestep using the scene representation; see top right of figure 2. The inhibition map modulates the saliency results and suppresses saliency hypotheses at the locations of tracked objects. In figure 3, the inhibition map is represented in the third column. The 4th column shows the selection of new targets (small circles on white regions), which are scheduled for tracking at the subsequent timestep. The 5th column shows the original input with the tracked objects (colored circles).

5 Outlook

We regard the system presented in section 4 as a first step towards an intentional vision system. It is still very simple since it couples only a few visual routines via bottom-up and top-down information flow, and many open questions remain.

One of the major points is the type of representation required at the visual memory level. It is clear from our model that a detailed, complete and up-to-date internal representation of the visual input according to Marr is not sensible, since the system can only make use of a small portion of the data at any time. In our case, the basic representation was tied to the type of data going from/to the visual routines, comprising sensory blobs, tracked objects and segmentation masks, together with their properties like activations, positions, velocities and extensions. Nevertheless, a data-driven representation does not mean that there is no room for more complex, hierarchical representations of e.g. real "objects" which could act as collectors that have saliency blobs, tracker functionality and segmentation masks as properties and extend them by additional information.

At the representational level, we made the distinction between items that are only memorized (stored together with some time information, for possible active

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

access later in time) and those that are actively represented and maintained using attention and top-down information flow. These attended items should be those that are relevant for a certain visual task, since their number is severely limited by processing constraints. This brings us to the next big open issue: How to represent visual tasks within the intentional framework and how to select and drive the optimal visual routines for a task on a need-to-know basis?

In an extension of the presented model, we are working on a large-scale implementation with further saliency routines (measuring contrast in 3D depth, combination of hue and saturation and saturation only), tracking capabilities that include 3D information and enhanced histogram-based level-set segmentation. The scene representation is in this case a relational network that allows to build hierarchical representations of "objects" starting from the measured data, together with relationships and dependencies such as the inheritance of properties, with target of the research being the representation of the procedural, task-dependent parts that allow to concentrate on the active and focused processing of visual information as proposed by the intentional framework.

References

1. *Beyond the Kalman Filter*. Artech House, 2004.
2. J. Aloimonos. Purposive and qualitative active vision. In *Proc. 10th Int. Conf. Patt. Recog.*, pages 345–360, June 1990.
3. Y. Aloimonos. Active vision revisited. In *Active Perception*, 1993.
4. S. A. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.
5. T. F. Chan and L. A. Vese. An active contour model without edges. In *Scale Space*, pages 141–151, 1999.
6. 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.
7. L. Itti and C. Koch. Saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.
8. D. Marr. *Vision*. Freeman, San Francisco, 1982.
9. 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.
10. M. Rousson and R. Deriche. A variational framework for active and adaptative segmentation of vector valued images. In *IEEE Workshop on Motion and Video Computing*, pages 56–61, 2002.
11. J. Sethian. *Level Sets Methods and Fast Marching Methods*. Cambridge University Press, 2nd edition, 1999.
12. V. Willert, J. Eggert, J. Adamy, and E. Körner. Non-gaussian velocity distributions integrated over space, time and scales. *IEEE Transactions on Systems, Man and Cybernetics B*, 2005.
13. J. M. Wolfe, T. S. Horowitz, N. Kenner, and M. Hyle N. Hasan. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, 44:1411–1426, 2004.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de