3

State of the Art of Computational Attention Systems

The increased interest on research on visual attention together with the increased power of computers and the resulting ability to realize complex computer vision systems has led to a wide variety of computational systems on visual attention. In this chapter, we will review the most influential work in this field. We already considered models of visual attention in the previous chapter. Although several of them are also implemented computationally, their focus is on the psychological aspect of visual attention more than on the technical aspect: the models of the previous chapter try to explain and better understand human perception whereas the systems in this chapter usually have the aim to improve vision systems for applications in computer vision and robotics. Of course, there is an overlap of the objectives and there are psychological models that might be useful in computational applications and technical systems well suited to explain psychophysical data.

In this chapter, we will first introduce several of the most important computational systems on visual attention (section 3.1). Then, we discuss several characteristics that distinguish the different approaches, for example which features are implemented or whether top-down cues are considered (section 3.2). Next, we present several applications of attentional systems in computer vision and robotics in section 3.3 and finally we conclude and discuss the limitations of current approaches (section 3.4).

3.1 Computational Models of Visual Attention

In this section, we will introduce some of the most important computational attention systems, especially those with the highest impact on our work. We start be introducing the model of Koch & Ullman, which laid the theoretical basis for many current attention systems [Koch and Ullman, 1985]. Next, we describe the system of Milanese, since it was one of the first implementations of an attention model and introduced several useful mechanisms that were later adopted by other approaches [Milanese, 1993]. Then, one of the currently

S. Frintrop: VOCUS: A Visual Attention System..., LNAI 3899, pp. 33-53, 2006.

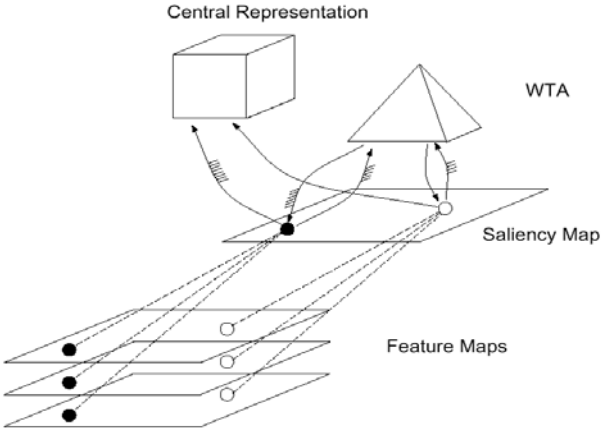

Fig. 3.1. The Koch-Ullman model. Different features are computed in parallel and their conspicuities are represented in several *feature maps*. A central *saliency map* combines the saliencies of the features and a *winner take all network (WTA)* determines the most salient location. This region is routed to the *central representation* where complex processing takes place (Fig. reprinted with permission from [Koch and Ullman, 1985]. ©1985 Springer)

best-known attention systems is presented: the *Neuromorphic Vision Toolkit (NVT)* of Itti et al. [Itti et al., 1998]. It will be described in some detail since it had the greatest impact on our work. Worth mentioning in this context is a derivative of the NVT that includes top-down information on target objects [Navalpakkam et al., 2005]. Another system that is able to cope with top-down information is the one of Hamker [Hamker, 2005]. After describing these attention systems explicitely, we mention in section 3.1.5 several additional approaches that emphasize other important aspects and are worth mentioning.

3.1.1 Koch & Ullman

The first approach for a computational architecture of visual attention was introduced by Koch and Ullman [Koch and Ullman, 1985] (see Fig. 3.1). When it was first published, the model was not yet implemented, but it provided the algorithmic reasoning serving as a foundation for later implementations and for many current computational models of visual attention. The idea is that several features are computed in parallel and their conspicuities are collected in a *saliency map*. A *Winner-Take-All network (WTA)* determines the most salient region in this map, which is finally routed to a *central representation*. Here, complex processing takes place restricted to the region of interest.

The model is based on the *Feature Integration Theory* by Treisman [Treisman and Gelade, 1980] (cf. chapter 2.3.1): the idea of feature maps that represent in parallel different features as well as the idea of a central map of

attention — Treisman's *master map of location* — are adopted. The saliency computations are also influenced by rules called *proximity* and *similarity preferences*, which favor regions that are close or similar to the last focused region. However, newer findings claim that distance has no effect on attentional shifts, that means there is no proximity effect [Remington and Pierce, 1984, Kröse and Julesz, 1989].

An important contribution of Koch and Ullman's work is the WTA network — a neural network that determines the most salient region in a topographical map — and a detailed description of its implementation. It may be noted that the WTA network shows how the selection of a maximum is implementable by neural networks, that means by single units which are only locally connected. This approach is strongly biological motivated and shows how such a mechanism might be realized in the human brain. However, for a technical system a WTA is certainly an overhead since there are much easier ways to compute a maximum from a saliency map. Nevertheless, many computational attention systems take over the idea of a WTA.

After selecting the most salient region by the WTA, this region is routed into a *central representation* which at any instant contains only the properties of a single location in the visual scene. Due to this routing, the approach is also referred to as *selective routing model*. How the routing is performed and what happens with the information in the central representation is not mentioned; the idea is that more complex vision tasks are restricted to the selected information. Finally, the authors suggest a mechanism for inhibiting the selected region causing an automatic shift towards the next most conspicuous location *(inhibition of return (IOR))*.

The idea of a central representation in this form is hardly plausible from a biologically point of view: simple and complex processing of visual information in the brain is thought to be more intertwined than suggested by this model. But from a computational point of view the method is suggestive since it enables a modular assembling of different systems: an attentional system for the detection of regions of interest and a recognition system for the detailed investigation of these regions.

The proposed architecture is merely bottom-up; it is not discussed how top-down influences from higher brain areas may contribute to the selection of salient regions.

3.1.2 Milanese

One of the earliest implementations of a visual attention system was introduced by Milanese [Milanese, 1993, Milanese et al., 1994]. It is based on the Koch-Ullman model [Koch and Ullman, 1985] and uses filter operations for the computation of the feature maps. Hence, it is one of the first in the group of *filter-based models*. These models are especially well-suited to be applied to real-world scenes since the filter operations — used frequently in computer

vision — provide useful tools for the efficient detection of scene properties like contrasts or oriented edges.

The idea of the feature maps and the saliency map was taken over from the Koch-Ullman model. As features, Milanese considers two color opponencies — red-green and blue-yellow —, 16 different orientations, local curvature and, if no color information is available, intensity. To compute the feature-specific saliency, he proposes a *conspicuity operator* which compares the local values of the feature maps to their surround. This operator is motivated from the on-off and off-on cells in the cortex and is also a common technique for detecting contrasts in images; it is usually referred to as *center-surround mechanism* or *center-surround difference*. The resulting contrasts were collected in so called *conspicuity maps*, a term that was since then frequently used to denote feature-dependent saliency.

The conspicuity maps are integrated into the saliency map by a relaxation process that identifies a small number of convex regions of interest. The output of the system is the saliency map that shows a few regions of interest. A process determining the order in which to select regions from this map is not mentioned. A drawback of the system is its high computational complexity that results from the many filter operations on different scales and by the relaxation process which, as per Milanese, usually requires about a dozen iterations. Although this drawback is nowadays no longer as significant as when the system was developed, the approach is still too computationally demanding for real-world applications.

In a derivative [Milanese et al., 1994], Milanese includes top-down information from an object recognition system realized by *distributed associative memories (DAMs)*. The idea is that object recognition is applied to a small number of regions of interest that are provided by the bottom-up attention system. The results of the object recognition are displayed in a top-down map which highlights the regions of recognized objects. This map competes with the conspicuity maps for saliency resulting in a saliency map combining bottom-up and top-down cues. The effect is that known objects appear more salient than unknown ones. It may be doubted if this is consistent with human vision, on the contrary, humans tend to pay more attention to unknown objects [Wang et al., 1994]. Nevertheless, for a technical system this might be an interesting approach, the more so as it is possible to provide the DAM only with a single object and thus highlight this object in a scene. This would correspond to visual search. Not mentioned is if there is an advantage of this system over pure object recognition.

Note that the top-down information only influences the conspicuity maps (feature dimensions) and not the feature maps (feature types). Therefore, it is not possible to strengthen properties like “red” or “vertical”. Furthermore, the system depends strongly on the object recognition system. It is not able to learn the features of an object independently. Nevertheless, the system provides an interesting approach and has set benchmarks for several techniques

which are used in computational attention models until today. Unfortunately, this promising system was not further developed since 1994.

3.1.3 Itti et al.

One of the currently best known attention systems is the *Neuromorphic Vision Toolkit (NVT)*, a derivative of the Koch-Ullman model [Koch and Ullman, 1985], that is steadily kept up to date by the group around Laurent Itti [Itti et al., 1998,Itti and Koch, 2001a,Miau et al., 2001,Itti and Koch, 2001b,Navalpakkam et al., 2005]. Their model as well as their implementation serve as a basis for many research groups; one reason for this is the good documentation and the availability of the source code for download, allowing other researchers to experiment and further develop the system [URL, 05].

Fig. 3.2 shows the basic structure of the model. The ideas of the feature maps, the saliency map, the WTA and the IOR were adopted from the Koch-Ullman Model, the approaches of using linear filters for the computation of the features, of determining contrasts by center-surround differences and the idea of the conspicuity maps were probably adopted from Milanese [Milanese, 1993]. The main contributions of this work are detailed elaborations on the realization of theoretical concepts, a concrete implementation of the system and the application to artificial and real-world scenes. The authors describe in detail how the feature maps for intensity, orientation, and color are computed: all computations are performed on *image pyramids*, *Image pyramid* a common technique in computer vision that enables the detection of features on different scales. Additionally, they propose a weighting function for the weighted combination of the different feature maps by promoting maps with few peaks and suppressing those with many ones. This technique is computationally much faster than the relaxation process of Milanese and yields good results. Since the suggested weighting function still suffered from several drawbacks, they introduced an improved procedure in [Itti and Koch, 2001b].

The system contains several details that were chosen for efficiency reasons or because they represent a straight-forward solution to complex requirements. This approach may lead to some problems and inaccurate results in several cases. For example, the center-surround mechanism is realized by the subtraction of different scales of the image pyramid, a method that is fast but not very precise (cf. page 61). Then, the conspicuity of the feature intensity is collected in a single intensity map, although neuro-biological findings show that there are cells both for on-off and for off-on contrasts [Palmer, 1999] and psychological work suggests considering separate detectors for darker and lighter contrasts [Treisman, 1993]. This simplification leads to some non-plausible results in certain pop-out experiments and in the top-down guidance of attention (cf. page 60). The same is true for the computation of the color-opponency maps: one red-green and one blue-yellow map are computed instead considering red-green as well as green-red and blue-yellow as well as yellow-blue contrasts separately. Furthermore, the chosen color space RGB represents colors

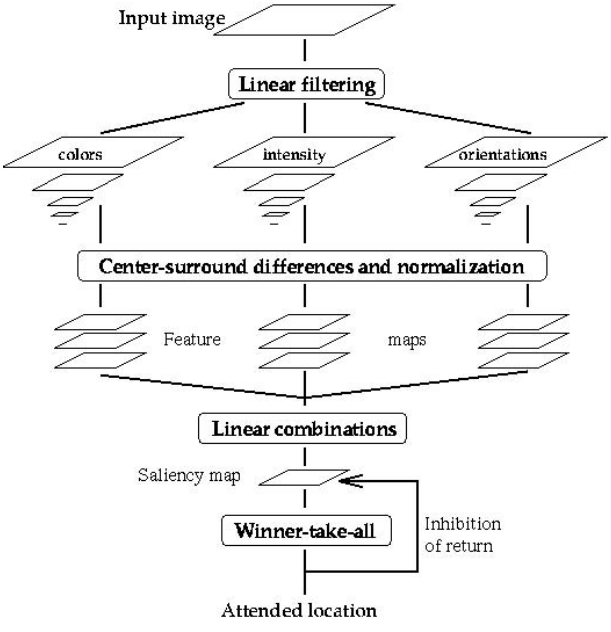

Fig. 3.2. Model of the *Neuromorphic Vision Toolkit (NVT)* by Itti et al. From an input image, three features are computed: color, intensity, and orientation. For each feature, an *image pyramid* is built to enable computations on different scales. *Center-surround mechanisms* determine the conspicuities concerning the features which are collected in a *central saliency map*. A *winner take all network* determines the most salient location in this map which yields the focus of attention. *Inhibition of return* inhibits this region in the saliency map and enables the computation of the next focus (Fig. reprinted with permission from [Itti et al., 1998]. ©1998 IEEE)

differently to human perception, which seems not appropriate for a system simulating human behavior and leads to implausible results, too. Although these are details, considering them in the implementation results in significant improvements in performance as will be shown in this work.

Some of these drawbacks were already pointed out by Draper and Lionelle [Draper and Lionelle, 2003] who showed that the NVT lacks robustness according to 2D similarity transformations like translations, rotations, and reflections. They point out that these drawbacks result from weaknesses in implementation rather than from the design of the model itself. To overcome these drawbacks, they introduced an improved version of the system, SAFE, which shows several differences and is more stable with respect to geometric transformations. It may be noted, that although these invariances are important for an object recognition system — the task Draper has in mind — they are not obviously required and maybe not even wanted for a system that aims

at simulating human perception since usually human eye movements are not invariant to these transformations, too. Nevertheless, it should be guaranteed that the computations are as correct as possible and that variances result only from the model and not from its implementation. On the other hand, if it is desired to achieve fast computations, time needs to be traded off against precision.

To evaluate the quality of the NVT, a comparison with human behavior was performed in [Parkhurst et al., 2002]. The authors compared how the saliency computed by the system matched with human fixations on the same scenes and found a significant coherence which was highest for the initial fixation. They also found that the coherence was dependent on the kind of scene: for fractals it was higher than for natural scenes. This was explained by the influence of top-down cues in the human processing of natural scenes, an aspect left out in the NVT.

Miau at al. investigated the combination of the NVT with object recognition, considering in [Miau and Itti, 2001, Miau et al., 2001] the simple biologically plausible recognition system HMAX and in [Miau et al., 2001] the recognition with support vector machines. Walther et al. continued these investigations, starting in [Walther et al., 2002] also with a combination with the HMAX model. In a current approach [Walther et al., 2004], they combine the system with the well-known recognition approach of Lowe [Lowe, 2004] and show how the detection results are improved by concentrating on regions of interest.

A test platform for the attention system — the robot platform *Beobot* — was presented in [Chung et al., 2002,Itti, 2002,Itti, 2003]. In [Itti, 2002], it was shown how the processing can be distributed among different CPUs enabling a fast, parallel computation.

Navalpakkam

The NVT in its basic version does concentrate on computing bottom-up attention. The need for top-down influences is mentioned but not realized. In a recent approach, Navalpakkam and Itti introduce a derivative of their bottom-up model which is able to deal with top-down cues [Navalpakkam et al., 2005]. The idea is to learn feature values of a target from a training image in which the target is indicated by a binary mask. Considering the target region as well as a region in the close surrounding — considering 9 locations from a 3×3 grid of fixed size centered at the salient location — the system learns the feature values from the different feature maps on different scales. This yields a 42 component feature vector (red/green, blue/yellow, intensity, and 4 orientations, each on 6 scales). However, it may be doubted if it is useful to learn the scale of a target since during visual search the target should be detected on different scales. During object detection, this feature vector is used to bias the feature maps by multiplying each map with the corresponding weight.

Thereby, exciting and inhibiting as well as bottom-up and top-down cues are mixed and directly fused into the resulting saliency map.

One difficulty with this approach is that it is not clear how bottom-up and top-down cues compete. Desirable for a technical system would be the possibility to adapt the strength of the respective influence according to the state of the system, similar to the approach of Milanese, that means a high or even exclusive concentration to the target's features in one case (task-oriented system state) and a higher influence of diverting bottom-up cues in another case (curious, explorative system state). Additionally, since there is evidence that two distinct brain areas are associated with bottom-up and top-down mechanisms in human perception [Corbetta and Shulman, 2002] (cf. chapter 2), it might be useful to separate the processing also in a computational system.

Unfortunately, a detailed analysis of the quality of the detection has not yet been published. Instead, the results in [Navalpakkam et al., 2005] concentrate on showing that the system detects a target faster when operating in top-down mode than the original bottom-up system. Also of interest would be investigations on how many fixations are needed in average in different visual search tasks and on the robustness of the system concerning changes in viewpoint and illumination. In chapter 5.4.5, we compare our attention system VOCUS in detail with the NVT, pointing out the differences of the models and showing results of comparative experiments.

So far, we have commented only on the aspects of Navalpakkam's approach that regard the main contributions of this monograph and therefore are of most interest here. However, it shall be mentioned that the system has several further aspects, only partially realized at the moment, which are interesting and promising. For example, the knowledge base in which the objects are stored is organized as a graph with entities as vertices and their relationships as edges. An object may be related to another for example by being similar or by being a part of the other object. This information might help in visual search: for example if a hand shall be found and a finger is detected, the knowledge that a finger is a part of a hand implies that the hand has been found.

Another interesting aspect is the idea of extending the model by additional information on the scene by computing the gist and the layout of the scene according to the psychological triadic architecture presented in [Rensink, 2002]. This is not yet realized but is, as per [Navalpakkam et al., 2005], subject for future work.

3.1.4 Hamker

The attention system of Hamker aims mainly at modeling the visual attention mechanism of the human brain [Hamker, 1998, Hamker, 2000, Hamker, 2005]. Its objective is more on explaining human visual perception and gaining insight into its functioning than on providing a technical system. Nev-

ertheless, this approach is discussed here and not in the previous chapter since it is based on current computer models [Koch and Ullman, 1985, Itti et al., 1998] and since it is often presented in the computer vision community. Hamker's model, shown in Fig. 3.3, shares several aspects with the architecture of Itti: he computes contrasts for several features — intensity, orientation, red-green, blue-yellow and additionally spatial resolution — and combines them in feature-conspicuity maps. The conspicuities of these maps are combined in a *perceptual map* that corresponds to the common saliency map. In earlier approaches, Hamker negates the existence of a saliency map in the human brain. But since new findings in neuro-science claim that there is a region in the brain fulfilling the function of collecting salient cues [Mazer and Gallant, 2003], he adopted his system accordingly [Hamker, 2004, Hamker, 2005].

In addition to this bottom-up behavior, the system belongs to the few existing ones that consider top-down influences. It is able to learn a target, that means it remembers the feature values of a presented stimulus. This stimulus is usually presented on a black background; hence, the system concentrates on the target's features but is not able to consider the background of the scene. This means a waste of important information since it is not possible to favor features that distinguish a target well from its background. When searching for a red, vertical bar among red, horizontal ones, the color red is not relevant; in this case, it would be useful to concentrate on orientation. To achieve a stable and robust system behavior, it would be necessary to learn the features of a target from several training images.

After determining the target's features, they are memorized in a *working memory*. From here, they influence the conspicuity of the features in a presented test scene and thus merge the conspicuities of bottom-up and top-down cues. It may be noted that the target information influences the processing of the conspicuity maps, but not the earlier processing of the feature maps. Bottom-up and top-down cues together determine the saliency in the perceptual map. A problem with this approach might be that it is not clear how bottom-up and top-down cues compete. As for the NVT, it might be useful to introduce a factor as the one by Milanese that allows the adaption of the influence of bottom-up and top-down cues.

Hamker distinguishes between covert and overt shifts of attention, the latter corresponding to eye movements. The covert focus of attention is directed to the most salient region in the perceptual map. Whether this region is also a candidate for an eye movement is determined by so called *match detection units* that compare the encoded pattern with the target template. If these patterns are similar, an eye movement is initiated towards this region and the target is said to be detected. The match detection units are an interesting approach in this system. However, it may be noted that this is a very rough kind of object recognition which is only based on a few simple features and does not consider spatial configuration of features. It also recognizes only patterns that are presented with the same orientation as during learning. Therefore, although at the moment this kind of recognition seems to be not sufficient in

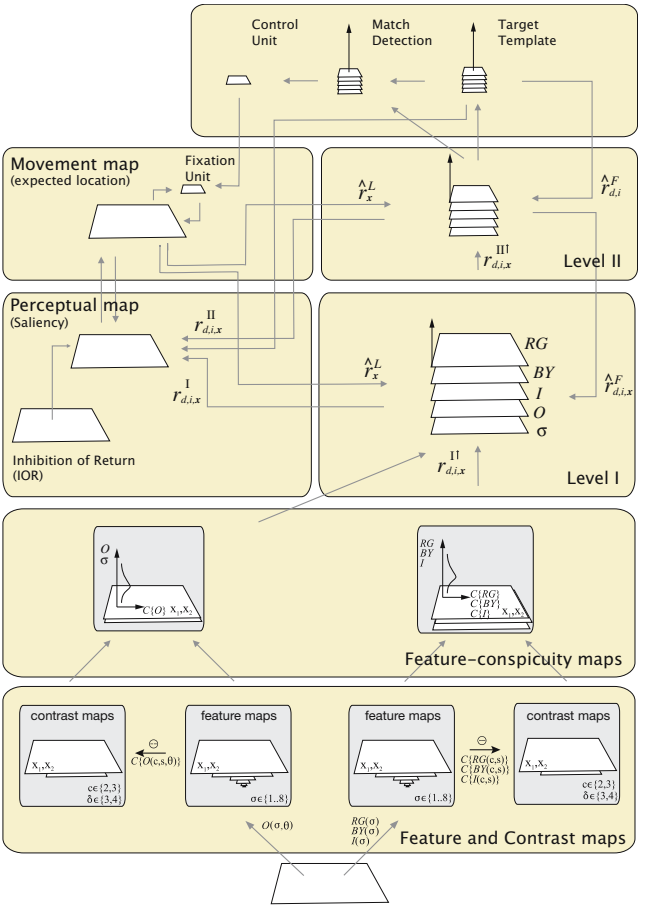

Fig. 3.3. The attention system of Hamker. From the input image, several feature and contrast maps are computed and fused into feature-conspicuity maps and finally into the perceptual map. Additionally, target information influences the processing. Match detection units determine whether a salient region in the perceptual map is a candidate for an eye movement. See text for details (Fig. reprinted from [Hamker, 2005], ©2005, with permission from Elsevier)

detection and false detection rates for a technical system, it is nevertheless an interesting approach and seems to be a step into the right direction.

3.1.5 Additional Attention Systems

Beside the mentioned attention models, there is a wide variety of models in the literature. Many differ only in minor changes from the already described

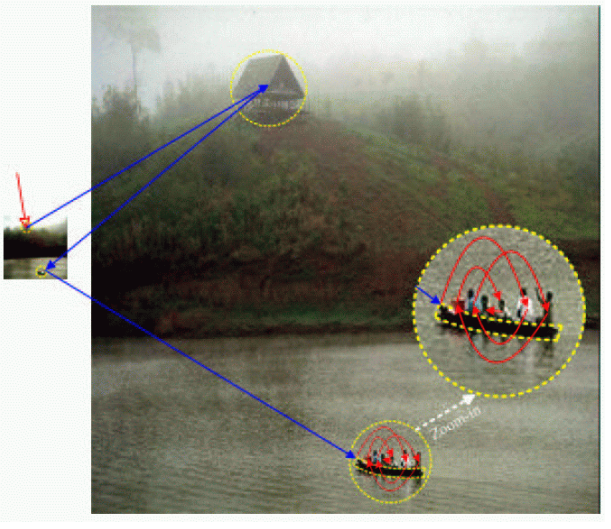

Fig. 3.4. The attentional shifts performed by the system of Sun and Fisher. First, the house and then the boat are focused on a coarse resolution (left, blue arrows). Second, the boat region is zoomed in and is investigated in more detail, resulting in fixations on the people (red arrows) (Fig. reprinted from [Sun and Fisher, 2003], ©2003, with permission from Elsevier)

approaches, for example, they consider additional features. Here, we mention some of the more important approaches in the field.

Sun and Fisher present in [Sun and Fisher, 2003] a sophisticated approach to hierarchical object-based selection of regions of interest. Regions of interest are computed on different scales, first on a coarse scale and then, if the region is sufficiently interesting, it is investigated on a finer scale. This yields foci of attention of different extents, for example in a landscape image showing a lake, a boat is focused on a coarse scale, then the boat region is further investigated on a finer scale and the people in the boat are focused one after the other (see Fig. 3.4).

Backer presents an interesting model of attention with two selection stages [Backer, 2004, Backer et al., 2001]. The first stage resembles standard architectures like [Koch and Ullman, 1985], but the result is not a single focus but a small number, usually 4, of salient locations. In the second selection stage, one of these locations is selected and yields a single focus of attention. The model explains some of the more unregarded experimental data on multiple object tracking and object-based inhibition of return.

The attention model of Ouerhani et al. is implemented on a highly parallel architecture that allows to meet real-time requirements [Ouerhani, 2003, Ouerhani and Hügli, 2003c]. They have also integrated the rarely considered features depth and motion into their system [Ouerhani and Hügli, 2000, Ouerhani

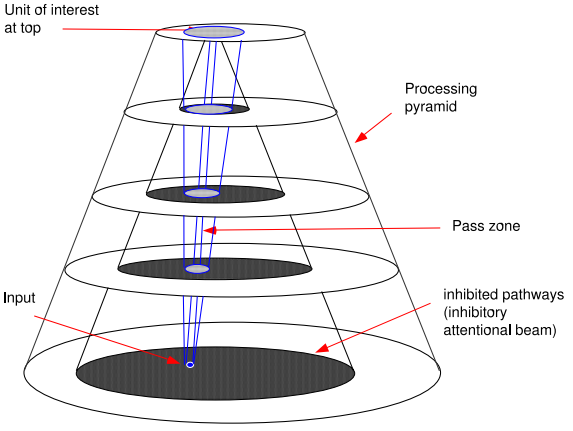

Fig. 3.5. The *inhibitory attentional beam* of Tsotsos et al. The selection process requires two traversals of the pyramid: first, the input traverses the pyramid in a feedforward manner. Second, the hierarchy of WTA processes is activated in a top-down manner to localize the strongest item in each layer while pruning parts of the pyramid that do not contribute to the most salient item (Fig. kindly provided by J. Tsotsos)

and Hügli, 2003b]. Another model which integrates these features is presented by the group of Eklundh [Maki et al., 1996, Maki et al., 2000].

Beside the mentioned models that are based on feature computations with linear filters, there is another important class of attention models: the *connectionist models*. These models process the input data mainly with neural networks. Usually, these models claim to be more biologically plausible than the filter models. Since this approach differs strongly from the approach presented in this thesis, these models will be mentioned only briefly here.

One of the most famous models in the field of *connectionist models* is the *selective tuning model* of visual attention by Tsotsos et al. [Tsotsos, 1990, Tsotsos, 1993, Tsotsos et al., 1995, Tsotsos et al., 2005]. It consists of a pyramidal architecture with an *inhibitory beam* (see Fig. 3.5). This beam is rooted at the selected item at the top of the hierarchy and has a *pass zone* and an *inhibit zone*. The pass zone is the pathway that is selected for further processing; in the inhibit zone, all locations are inhibited that do not belong to the selected item. It is also possible to include target-specific top-down cues into the processing. This is done by either inhibiting all regions with features different from the target features or regions of a specified location. Additional excitation of target features as proposed by [Navalpakkam et al., 2004] is not considered. The model has been implemented for several features, for example luminance, orientation, or color opponency [Tsotsos et al., 1995], and currently in a sophisticated approach also for motion, considering even the direction of

movements [Tsotsos et al., 2005]. Note that in each version only one feature dimension is processed; the binding of several feature dimensions has not yet been considered but is, as per Tsotsos, subject for future work.

An unusual adaptation of Tsotsos's model is provided in [Ramström and Christensen, 2002]: the distributed control of the attention system is performed by game theory concepts. The nodes of the pyramid are subject to trading on a market, the features are the goods, rare goods are expensive (the features are salient), and the outcome of the trading represents the saliency.

Another model based on neural networks is the *FeatureGate* Model described in [Cave, 1999]. Beside bottom-up cues it also considers top-down cues by comparing pixel values with the values of a target object; but since the operations only work on single pixels and so are highly sensitive to noise, it seems to be not applicable to real-world scenes.

3.2 Characteristics of Attention Systems

After introducing some of the most influential computational attention systems, we summarize in this section several characteristics of attention systems that distinguish the respective approaches. We start by distinguishing the objectives of the systems concerning psychological or technical issues and continue by discussing which features are computed in the different approaches. Next, we distinguish connectionist and filter models and finally, we examine what kinds of top-down influences exist and how they are realized in several computational attention systems.

3.2.1 Objective

Computational attention systems might be categorized by their objective. As already mentioned, the systems may be firstly designed to simulate and understand human perception or, secondly, to technically improve vision systems. Although systems of both classes may be very similar, this distinction usually has a high impact on the visibility of the systems: whereas the first class of systems is usually well known by the psychological and cognitive science community, the latter class is more familiar in areas like computer vision and robotics. Since each side may highly profit from the knowledge of the other, a better interchange between communities would be desirable.

3.2.2 The Choice of Features

Many computational attention systems focus on the computation of mainly three features: intensity, orientation, and color [Itti et al., 1998,Draper and Lionelle, 2003,Sun and Fisher, 2003,Ramström and Christensen, 2004]. Reasons for this choice are that these features belong to the basic features proposed in

psychological and biological work [Treisman, 1993, Wolfe, 1994, Palmer, 1999] and that they are relatively easy to compute. A special case of color computation is the separate computation of skin color [Rae, 2000, Heidemann et al., 2004, Lee et al., 2003]. This is often useful if faces or hand gestures have to be detected. Other features that are considered are for example curvature [Milanese, 1993], spatial resolution [Hamker, 2005], optical flow [Tsotsos et al., 1995, Vijayakumar et al., 2001], or corners [Ouerhani and Hügli, 2004, Fraundorfer and Bischof, 2003, Heidemann et al., 2004]. Several systems compute also higher level features that use approved techniques of computer vision to extract useful image information. Examples for such features are entropy [Heidemann et al., 2004], ellipses [Lee et al., 2003], eccentricity [Backer et al., 2001], or symmetry [Backer et al., 2001, Heidemann et al., 2004, Lee et al., 2003].

Motion is definitively an important feature in human perception (there is a large brain area (MT) mainly concerned with processing motion!). Nevertheless, it is rarely considered in computational models, probably because of the difficulties arising when dealing with dynamics. Some approaches that consider motion as a feature are [Backer and Mertsching, 2000, Maki et al., 2000, Ouerhani and Hügli, 2003b, Itti, 2002, Rae, 2000]. All of these approaches only implement a very simple kind of motion detection: usually, two subsequent images in a video stream are subtracted and the difference codes the feature conspicuity. The most sophisticated approach concerning motion was recently proposed in [Tsotsos et al., 2005]. This approach is highly biologically motivated, it considers the direction of movements, and processes motion on several levels similar to the processing in the brain regions V1, MT, and MST.

Another important aspect in human perception that is rarely considered is depth. In the literature it is not clear whether depth is simply a feature or something else; definitely, it has some unusual properties distinguishing it from other features: if one of the dimensions in a conjunctive search is depth, a second feature can be searched in parallel [Nakayama and Silverman, 1986], a property that does not exist for the other features. Computing depth for an attention system is usually solved with stereo vision [Backer and Mertsching, 2000, Maki et al., 2000]. The data obtained from stereo vision has the drawback that it is usually not very accurate and contains large regions without depth information. Another approach is to use special 3D sensors, as for example the lately appearing 3D cameras [Ouerhani and Hügli, 2000].

Finally, it may be noted that considering more features usually results in more accurate and biologically plausible detection results but also reduces the speed since the parallel architectures are usually implemented sequentially. Furthermore, the concept of the models is the same regardless of the number of features, therefore, most effects can already be shown with a small number of features.

3.2.3 Connectionist Versus Filter Models

As mentioned before, there is a distinction between connectionist models that are based on neural networks and filter models that use classical linear filters to compute features. Usually, the connectionist models claim to be more biologically plausible than the filter models since they have single units corresponding to neurons in the human brain, but it has to be noted that they are still a high abstraction from the processes in the brain. Usually, a single neuron is more complex than a complete computational system. Furthermore, also filter models may be strongly biologically motivated, as the system of Hamker shows [Hamker, 2005].

However, the advantage of connectionist models is that they are — at least theoretically — able to show a different behavior for each neuron whereas in filter models usually each pixel in a map is treated equally. In practice, treating each unit differently is usually too costly and so a group of units shows the same behavior. The advantage of filter models is that they may profit from approved techniques in computer vision and that they are especially well suited for the application to real-world images.

Examples of connectionist systems of visual attention are presented for instance in [Olshausen et al., 1993, Postma, 1994, Tsotsos et al., 1995, Baluja and Pomerleau, 1995, Cave, 1999]. As mentioned in chapter 2, many psychophysical models fall into this category, too, for example [Mozer, 1987, Phaf et al., 1990, Humphreys and Müller, 1993, Heinke et al., 2002]. Examples of linear filter systems of visual attention are presented for instance in [Milanese, 1993, Itti et al., 1998, Rae, 2000, Backer et al., 2001, Ouerhani, 2003, Sun and Fisher, 2003, Heidemann et al., 2004, Hamker, 2005].

3.2.4 Top-Down Cues

The distinction of bottom-up and top-down cues and their significance in human perception was already outlined in section 2.1.3. For a technical attention system, top-down cues are equally important: most systems are not only designed to detect bottom-up salient regions but there are goals to achieve and targets to detect. Although the importance of top-down cues is well known and even mentioned in many articles, most systems consider only bottom-up computations.

Before we discuss which systems consider top-down information, we will first distinguish between different kinds of top-down influences. Top-down information includes all kinds of information that exist at one moment in time concerning the mental state of the subject (or the inner state of the system) and knowledge of the outer world. This includes aspects like prior knowledge of the target, pre-knowledge of the scene or of the objects that might occur in the environment, but also emotions, desires, intentions, and motivations. The latter four aspects are hard to conceptualize and are not realized in any computer system we know about. The interaction of attention,

emotions, motivations, and goals is discussed in [Balkenius, 2000, Balkenius, 2002], but in his computer simulation these aspects are not considered.

Top-down information that refers to knowledge of the outer world, that means of the background scene or of the objects that might occur, is considered in several systems. In these approaches, for example all objects of a data base that might occur in a scene are investigated in advance and their most discriminative regions are determined, i.e., the regions that distinguish an object best from all others in the data base [Fritz et al., 2004, Pessoa and Exel, 1999]. Another approach is to regard context information, that means searching for a person in a street scene is restricted to the street region and the sky region is ignored. The contextual information is obtained from past search experiences in similar environments [Oliva et al., 2003, Torralba, 2003].

The kind of top-down information that will be most relevant in this thesis is the prior knowledge of a target that is used to perform visual search. Systems regarding this kind of top-down information use knowledge of the target to influence the computation of the most salient region. This knowledge is usually learned in a preceding training phase but might in simpler approaches also be provided manually by the user.

In the existing systems, the target information influences the processing at different stages: some systems already influence the feature types (usually the feature maps) [Navalpakkam et al., 2005, Tsotsos et al., 1995], some systems influence the feature dimensions (usually the conspicuity maps) [Milanese et al., 1994, Hamker, 2005], and some influence the processing not before the computation of the saliency map [Rao et al., 2002, Lee et al., 2003, Navalpakkam and Itti, 2002]. The latter approach is a very simple one: the bottom-up saliency map is computed and the most salient regions are investigated for target similarity. It can be hardly called top-down influence of processing at all. Only targets that are most salient in a scene can be found with this approach. More elaborated is the tuning of the conspicuity maps, but biologically most plausible and also technically most useful is the approach to already bias the feature types as for example red or horizontal.

There are also different methods for influencing the maps with the target information. Some approaches inhibit the target-irrelevant regions [Tsotsos et al., 1995], whereas others prefer exciting target-relevant regions [Hamker, 2005, Navalpakkam and Itti, 2003]. New findings suggest that inhibition and excitation both play an important rule [Navalpakkam et al., 2004]; this is implemented in [Navalpakkam et al., 2005].

The processing of target-relevant top-down cues in computational attention systems is not yet well investigated. Even the systems that consider top-down cues are seldomly tested on natural scenes or only on hand-picked examples [Hamker, 2005]. The currently best tested system also including natural scenes is presented in [Navalpakkam et al., 2005]. Unfortunately, the quality of the detection results has not yet been published; the mentioned paper focuses on comparing the top-down approach with the previous bottom-up system (merely the improvement factor is indicated not the absolute detec-

tion results). Currently, there exists no complete, robust, and well investigated system of top-down visual attention which analyzes the influence of top-down cues systematically for different targets, with changing viewpoints, on different backgrounds, and under changing illumination conditions.

3.3 Applications in Computer Vision and Robotics

While psychological models of visual attention usually aim at describing and better understanding human perception, computational attention systems usually intend to improve technical systems. In this section, we discuss several application scenarios in the field of computer vision and robotics and introduce the approaches that currently exist in this field.

3.3.1 Object Recognition

Probably the most suggesting application of an attention system is object recognition since the two-stage approach of a preprocessing attention system and a classifying recognizer is adapted to human perception [Neisser, 1967]. It is worth mentioning that object recognition may be a subtask of more complex applications like object manipulation in robotics, which will be described later.

One example of a combination of an attentional front-end with a classifying object recognizer is shown in [Miau and Itti, 2001, Miau et al., 2001]. The recognizer is the biologically motivated system HMAX [Riesenhuber and Poggio, 1999]. Since this system focuses on simulating processes in human cortex, it is rather restricted in its capabilities and it is only possible to recognize simple artificial objects like circles or rectangles. In [Miau et al., 2001], the authors replace the HMAX system by a support vector machine algorithm to detect pedestrians in natural images. This approach is much more powerful with respect to the recognition rate but still computationally very expensive and lacks real-time abilities. Walther and colleagues combine in [Walther et al., 2004] an attention system with an object recognizer based on SIFT features [Lowe, 2004] and show that the recognition results are improved by the attentional front-end. In [Salah et al., 2002] an attention system is combined with neural networks and an observable Markov model to do handwritten digit recognition and face recognition. In [Ouerhani, 2003], an attention-based traffic sign recognition system is presented.

All of these systems rely only on bottom-up information and therefore on the assumption that the objects of interest are sufficiently salient by themselves. Non-salient objects are not detected and so they are missed. For some object classes like traffic signs which are intentionally designed salient, this works quite well; for other applications, top-down information would be needed to enable the system to focus on the desired objects.

It may also be mentioned that when combining object recognition with attention, the advantage over pure classification is usually the time saving

and not the quality improvement: most classifiers show no improvement if restricted to a region of interest (an exception is the work of Walther et al. [Walther et al., 2004] since the Lowe detector improves if restricted to a region of interest). Since most attention systems are still rather slow and the recognition systems not powerful enough to deal with a wide variety of objects, the advantage of such a combination of attention and classification does usually not yet show of to its best. Currently, there is no existing approach that exhibits a time saving resulting from the combination of attention and classification. However, in future, with more powerful recognition systems and more complex requirements concerning vision systems, an attentional front-end is a promising approach.

A different view on attention for object recognition is presented in [Fritz et al., 2004]: an information-theoretic saliency measure is used to determine discriminative regions of interest in objects. The saliency measure is computed by the conditional entropy of estimated posteriors of the local appearance patterns. That means, regions of an object are considered as salient if they discriminate the object well from other objects in an object data base. A similar approach is presented in [Pessoa and Exel, 1999].

3.3.2 Image Compression

A new and interesting application scenario is presented in [Ouerhani et al., 2001]: *focused image compression*. Here, a color image compression method adaptively determines the number of bits to be allocated for coding image regions according to their saliency. Regions with high saliency have a higher reconstruction quality with respect to the rest of the image.

3.3.3 Image Matching

Image matching is the task to redetect a scene, or part of a scene, in a newly presented image. This is often done by matching relevant key points. An approach that uses foci of attention computed by a saliency operator for image matching is presented in [Fraundorfer and Bischof, 2003].

3.3.4 Image Segmentation

The automatic segmentation of images into regions usually deals with two major problems: first, setting the starting points for segmentation (seeds) and second, choosing the similarity criterion to segment regions (cf. appendix A.3). Ouerhani et al. present an approach that supports both aspects by visual attention [Ouerhani et al., 2002, Ouerhani and Hügli, 2003a]: the saliency spots of the attention system serve as natural candidates for the seeds and the homogeneity criterion is adapted according to the features that discriminate the region to be segmented from its surroundings.

3.3.5 Object Tracking

Tracking objects in dynamic environments is important in applications such as video surveillance or robotics. In [Ouerhani and Hügli, 2003b], the authors present an approach in which the salient spots are tracked over time; however, the tracking is only done by feature matching instead of using a proper tracking method as for example Kalman filters. In [Ouerhani and Hügli, 2004] the authors suggest to use this approach for robot localization. The localization itself has not yet been done.

3.3.6 Active Vision

Active vision represents the technical equivalent for overt attention by directing a camera to interesting scene regions and/or zooming these regions. The goal is to acquire data that is as suitable as possible to the current task and to reduce the processing complexity by actively guiding the sensors (usually the camera) to reasonable regions [Aloimonos et al., 1988]. In several cases, active vision is a subtask for applications like human-robot interaction and object manipulation, which will be discussed in the next sections.

In [Mertsching et al., 1999, Bollmann, 1999], the active vision system NAVIS is presented that uses an attention system to guide the gaze. It is evaluated on a fixed stereo camera head as well as on a mobile robot with a monocular camera head. In [Vijayakumar et al., 2001] an attention system is used to guide the gaze of a humanoid robot. The authors consider only one feature, visual flow, which enables the system to attend to moving objects. To simulate the different resolutions of the human eye, two cameras per eye are used: one wide-angle camera for peripheral vision and one narrow-angle camera for foveal vision. Other approaches which use attention systems to direct the gaze of an active vision system are described in [Clark and Ferrier, 1989] and [Driscoll et al., 1998].

3.3.7 Human-Robot Interaction

If robots shall interact with humans, it is important that both agree on a current object or region of interest. A computational attention system similar to the human one can help to focus on the same region. Breazeal introduces a robot that shall look at people or toys [Breazeal, 1999]. Although top-down information would be necessary to focus on an object relevant for a certain task, bottom-up information can be useful too if it is combined with other cues. For example, Heidemann et al. combine an attention system with a system that follows the direction of a pointing finger and so can adjust to the region that is pointed at [Heidemann et al., 2004]. In [Rae, 2000] this approach is used to guide a robot arm to an object and grasp it.

3.3.8 Object Manipulation in Robotics

A robot that has to grasp and manipulate objects first has to detect and possibly also to recognize the object. Attentional mechanisms can be used to support these tasks. For example, Tsotsos et al. present a robot for disabled children that detects toys by the help of attention, moves to a toy and grasps it [Tsotsos et al., 1998]. In another approach, Bollmann et al. present a robot that uses the active vision system NAVIS to play at dominoes [Bollmann et al., 1999]. The above mentioned approach of Rae in which a robot arm has to grasp an object a human has pointed at, falls also into this category [Rae, 2000].

3.3.9 Robot Navigation

In [Scheier and Egner, 1997] a mobile robot is presented that uses an attention system for navigation. The task was to approach large objects. Since larger objects have a higher saliency, only the regions with the highest saliency have to be approached. The task gives the impression to be rather artificially made up.

In [Baluja and Pomerleau, 1995,Baluja and Pomerleau, 1997], an attention system is used to support autonomous road following by highlighting relevant regions in a saliency map. These are obtained by computing the expectation of the contents of the inputs at the next time step.

3.3.10 Robot Localization

Another application scenario of an attention system in robotics is the detection of landmarks for localization. Especially in outdoor environments and open areas, the standard methods for localization like matching 2D laser range and sonar scans are likely to fail. Instead, localization by detection of visual landmarks with a known position can be used. Attentional mechanisms can facilitate the search of landmarks during operation by selecting interesting regions in the sensor data. By focusing on these regions and comparing the candidates with trained landmarks, the most probable location can be determined. A project that follows this approach is the ARK project [Nickerson et al., 1998]. It relies on hand-coded maps, including the locations of known static obstacles as well as the locations of natural visual landmarks.

As already mentioned, [Ouerhani and Hügli, 2004] suggest to use matching and tracking of salient regions for robot localization but a realization of the localization itself has not yet been done.

3.4 Discussion

In this chapter, we have introduced several of the best known computational systems of visual attention in the field of computer vision. Their objective is to

profit from findings on human perception to improve technical computer vision systems. We first presented some of the most influential systems in detail; after we discussed several characteristics of current systems, for example the kind of features that are computed. Finally, we presented several application scenarios in computer vision and robotics in which attention systems are applied.

The modeling of visual attention is a wide field and it is hardly possible for one group to address all of the issues that arise. Therefore, each system emphasizes and specializes on a different aspect. However, there are aspects that are hardly considered due to costly realization or to missing evidence from the field of human perception. Let us summarize some of the limitations of current computational attention systems and some issues that are seldomly addressed.

First, features like depth and motion are seldomly considered in computational attention system. When changing from static 2D images to dynamical 3D applications, both provide useful information in natural environments. Second, there are few systems which integrate top-down influences and enable visual search. The few systems that do show hardly any evaluation of their approach and usually present only some isolated examples of the functionality of their system. A robust, well-evaluated approach does not yet exist. Third, since most systems focus on bottom-up computations, the evaluation of the systems is hard because there is usually no ground truth. The decision whether a computed focus of attention is reasonable, is usually left to the observer. Fourth, the computations usually focus on camera data although human attention operates for all senses. Especially in robotics, the consideration of additional sensors would be desirable. Finally, although there are several approaches that combine their attention system with object recognition, these approaches usually do not evaluate this combination and do not show its advantage. Neither the improvement in time performance nor a change in detection quality is discussed. Furthermore, since most systems operate merely in a bottom-up mode, the combination of top-down attention with object recognition has not yet been done. This results in recognition systems that are only able to recognize the most salient regions in a scene but not a target of current interest.

In the following chapters, we will present the computational attention system VOCUS that overcomes most of the discussed limitations of existing approaches.