VRIJE UNIVERSITEIT BRUSSEL
FACULTY OF ENGINEERING
Department of Electronics and Informatics (ETRO)
Image Processing and Machine Vision Group (IRIS)

Visual Attention Framework: Application to Event Analysis

Thesis submitted in fulfilment of the requirements for the award of the degree of
Doctor in de ingenieurswetenschappen (Doctor in Engineering)
by

Geerinck Thomas

Examining committee
Prof. Hichem Sahli, Vrije Universiteit Brussel, promoter
Dr. Valentin Enescu, Vrije Universiteit Brussel, co-promoter
Prof. Ann Nowé, Vrije Universiteit Brussel
Prof. Rik Pintelon, Vrije Universiteit Brussel
Prof. Werner Verhelst, Vrije Universiteit Brussel
Prof. Eric Soetens, Vrije Universiteit Brussel
Prof. Jan Cornelis, Vrije Universiteit Brussel
Prof. Robert B. Fisher, University of Edinburgh
Dr. Lucas Paletta, Institute of Digital Image Processing

Address
Pleinlaan 2, B-1050 Brussels, Belgium
Tel: +32-2-6291300
Fax: +32-2-6292883
Email: tgeerinc@etro.vub.ac.be

Brussels, May 2009

Contents

List of Figures

List of Tables

Abstract

One of the monolithic goals of computer vision is to automatically interpret general digital images or videos of arbitrary scenes. However, the amount of visual information available to a vision system is enormous and in general it is computationally impossible to process all of this information bottom-up. To ensure that the process has tractable computational properties, visual attention plays a crucial role in terms of selection of visual information, allowing monitoring objects or regions of visual space and select information from them for report, recognition, etc. This dissertation discusses one small but critical slice of a cognitive computer vision system, that of visual attention. In contrast to the attention mechanisms used in most previous machine vision systems, which drive attention based on the spatial location hypothesis, in this work we propose a novel model of object-based visual attention, in which the mechanisms which direct visual attention are object-driven. Considering the temporal dynamics associated with attention-dependent motion, an attention-based visual motion framework is also proposed. Finally, since a vision system will always have a set of tasks that defines the purpose of the visual process, a top-down approach is proposed to define the competition of the visual attention occurring not only within an object but also between objects, and illustrated in the framework of a surveillance system.

Chapter 1

Attentional Selection of Objects of Interest

1.1 Space-based Models of Visual Attention

A wide variety of visual attention models, simulating human perception, exists in the field of psychology. In this section, we give an overview of visual attention models considering space as the unit of attentional selection.

1.1.1 Treisman's Feature Integration Theory [66]

The *Feature Integration Theory (FIT)*, introduced in 1980 [66], is considered as the seminal work for computational visual attention. The theory evolved towards current research findings. Figure 1.1 depicts the main ideas of the FIT scheme. The reader is referred to [68] for more details.

In [66], it is stated that "different visual features are registered automatically and in parallel across the visual field, while objects are identified separately and only thereafter at a later stage, which requires focused attention". Information from the resulting *feature maps* - topographical maps that highlight saliency according to the respective feature - is collected in a *master map of location*. This map specifies *where* (in the image) the entities (points, regions, objects) are situated, but not *what* they are. Scanning serially through this map directs the focus of attention towards selected scene entities and provides data useful for higher perception tasks. Information about the target entities is gathered into so called *object files* [68].

1.1.2 Wolfe's Guided Search [7]

Another very important work, in the field of visual attention, is the *Guided Search Model (GSM)* of Wolfe [7, 76, 74, 75]. Figure 1.2 depicts the model architecture. It shares many concepts with the FIT, moreover, it gives details allowing computational implementations. Like FIT, it models several feature maps. Unlike FIT it does not follow the idea that there are separate maps for each *feature type* (red, green, ...), it defines only one map for each *feature dimension*, and within each map different feature types are represented. However, Wolfe mentions that there is evidence for differences between features. For

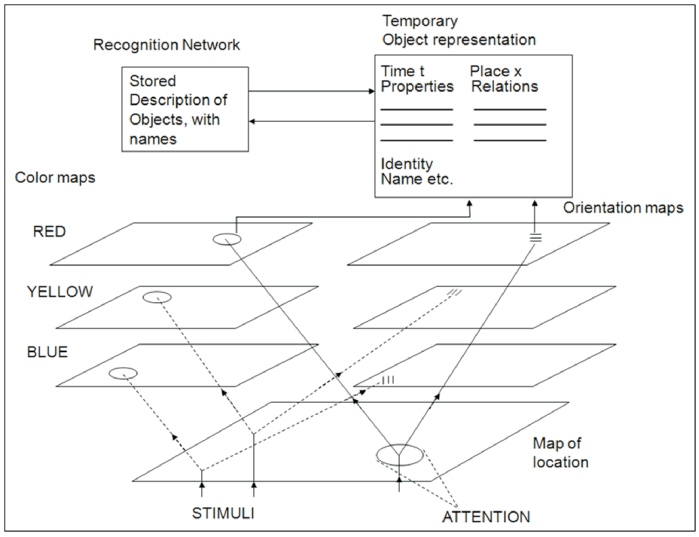

Figure 1.1: Model of *Feature Integration Theory (FIT)* [67]. Features such as color and orientation are coded automatically, pre-attentively and in parallel. Each feature dimension consists of several *feature maps* (red, yellow, blue for color). The saliency values of the feature are coded in the *master map of locations*. When attention is focused on one location in this map, it allows retrieval of the features that are currently active at that location and creates a temporary representation of the object in an *object file*.

example, there may be multiple color maps but only one orientation map [46]. The features considered in the implementation are color and orientation.

Comparable to the *master map of location* in FIT, there is an *activation map* in GSM in which the feature maps are fused. But in contrast to at least the early versions of FIT, in GSM the attentive part profits from the results from the pre-attentive one. The fusion of the feature maps is done by summing them.

Additionally to this bottom-up behavior, the model also considers the influence of top-down information. To realize this, for each feature there is not only a bottom-up but also a top-down map. The latter map selects the feature type which distinguishes the target best from its distractors. This is not necessarily the feature with the highest activation for the target. Only one feature type is chosen.

1.1.3 Additional Psychophysical Models

Besides the FIT and the GSM models, there is a wide variety of psychophysical models on visual attention. The often used metaphor of attention is a *spotlight* coming from the *zoom lens model* [15]. In this model, the scene is investigated by a spotlight with varying size. Many attention models fall into the category of connectionist models, referring to models based on neural networks. They are composed of a large number of processing units connected by inhibitory and excitatory links. Examples are the *dynamic routing circuit* [47], SeLective Attention Model (SLAM) [51], SEarch via Recursive Rejection (SERR) [22], and Selective Attention for Identification Model (SAIM) [20].

A formal mathematical model is presented in [34]: the CODE Theory of Visual Attention (CTVA). It integrates the COntour DEtector (CODE) theory for perceptual grouping. The theory is based on

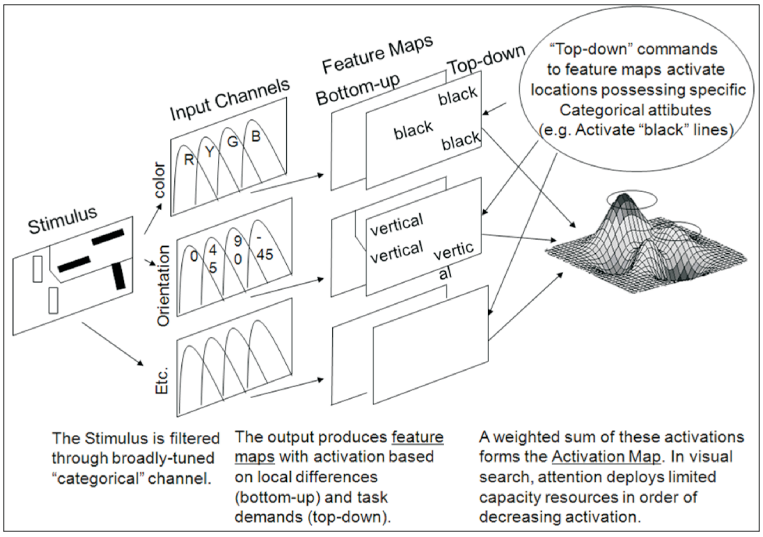

Figure 1.2: The *Guided Search model* of Wolfe [76]. One map for each feature dimension codes the properties of a scene concerning several feature types. Additionally to these bottom-up maps, top-down maps highlight the regions with task-specific attributes. A weighted sum of these activations forms the *activation map*.

a *race model* of selection. In these models, a scene is processed in parallel and the element that first finishes processing is selected (the winner of the race). That means, a target is processed faster than the distractors in a scene. Newer work concerning CTVA can be found, for example, in [5].

1.1.4 Koch & Ullman [32]

The first approach for a computational architecture of visual attention was introduced by Koch and Ullman [32] (see Figure 1.3). It served 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 (WTA)* network determines the most salient location in this map, which is routed to a *central representation*, where more complex processing might take place.

The model is based on the FIT of Treisman [66]. The feature maps, that represent in parallel different features, as well as the central map of attention (Treisman's *master map of location*) are adopted.

An important contribution of Koch and Ullman's work is the WTA network - a neural network that determines the most salient location in a topographical map - and a detailed description of its implementation. The WTA network shows how the selection of a maximum in neural networks is performed, by single units that are only locally connected. This approach is strongly biologically motivated and shows how such a mechanism might be realized in the human brain. However, from the implementation point of view, WTA brings a computational overload to the system.

The most salient location, selected by WTA, is then routed into a central representation which at any instant contains only the properties of a single location in the visual scene. The idea is that more complex vision processes are restricted to selected information. Due to this routing, the approach is also referred to as *selective routing*. Finally, a mechanism is suggested for inhibiting the selected region causing an

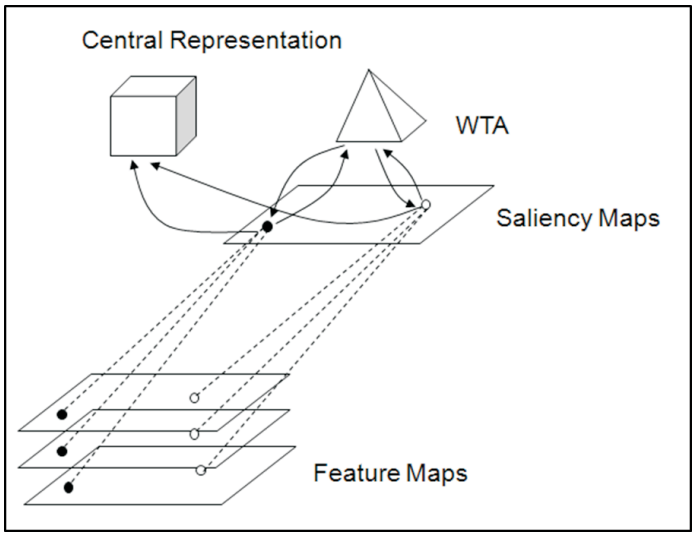

Figure 1.3: The Koch-Ullman model [32]. Different features are computed in parallel and their *conspicuities* are represented in several *feature maps*. A central *saliency map* combines the saliencies of the features and a *winner take all network (WTA)* determines the most salient location. This region is routed to the *central representation* where complex processing takes place.

automatic shift towards the next most conspicuous location (*inhibition of return (IOR)*).

1.1.5 Milanese [40]

The implementation of the visual attention system by Milanese [40, 41] is based on the Koch and Ullman model [32] and uses filter operations for the computation of the feature maps. Hence, it is one of the first *filter-based models*. These models are especially well-suited to be applied to real-world scenes since the filter operations provide useful tools for the efficient detection of scene properties like contrasts or edges' orientations.

As features, Milanese considers two color opponencies - *red-green* and *blue-yellow* -, 16 different orientations, local curvature and, intensity. To compute the feature-specific saliency, he proposes a *conspicuity operator*, referred to as *center-surround mechanism* or *center-surround difference*, which compares the local values of the feature maps to their surround. The resulting contrasts are collected in the 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 locations of interest, highlighted on the saliency map. A process determining the order in which to select the locations from this map is not proposed.

In [41], Milanese includes top-down information from an object recognition system. 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 top-down map competes with the conspicuity maps for saliency, resulting in a saliency map combining top-down and bottom-up cues. The effect is that known objects appear more salient than unknown ones. 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

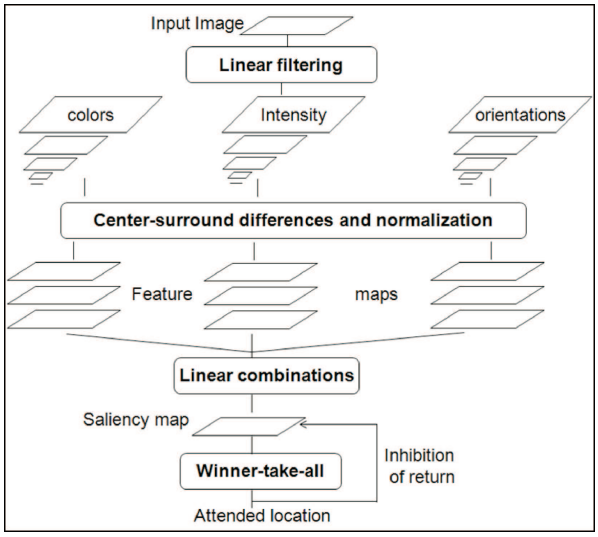

Figure 1.4: Model of the *Neuromorphic Vision Toolkit (NVT)* by Itti et al. [27] From an input image, three features are computed: color, intensity, and orientation. For each feature, an *image pyramid* is built to enable computations on different scales. *Center-surround mechanisms* determine the conspicuities 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.

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.

1.1.6 Itti et al. [27]

One of the currently most used attention systems is the *Neuromorphic Vision Toolkit (NVT)*, a derivative of the Koch-Ullman model [32], that is steadily kept up to date [27, 25, 39, 26, 43]. The good documentation and the availability of the source code [1] makes the model a very popular basis for many research groups. Figure 1.4, shows the basic structure of the model. The ideas of the feature maps, the saliency map, the WTA and the IOR mechanisms were adopted from the Koch-Ullman Model; the approaches of using linear filters for the feature computation, of determining the contrasts by center-surround differences, as well as the conspicuity maps were adopted from Milanese [40]. 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*, a common technique in computer vision that enables the estimation of features at different scales. Additionally, they propose a number of different techniques for combining different feature maps, including a weighting function promoting maps with fewer peaks and suppressing those with many ones; and a non-linear procedure introduced in [26].

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. Then, the conspicuity of the feature intensity is collected in a single intensity map, although neuro-biological findings show that there are cells for both on-off and off-on contrasts [48] and psychological work suggests considering separate detectors for darker and lighter contrasts [68]. The same is true for the computation of the color-opponency maps: one red-green and one blue-yellow map are computed instead of 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 differently to human perception, which seems not appropriate for a system simulating human behavior.

Some of these detailed drawbacks were pointed out by Draper and Lionelle [10] who showed that the NVT lacks robustness according to 2D similarity transformation like translations, rotations, and reflections. They pointed 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, called Selective Attention as a Front End (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 task - the task Draper had 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.

To evaluate the quality of the NVT, a comparison with human behavior was performed in [49]. 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 fractal images 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 et al. [38], [39] investigated the combination of the NVT with object recognition, considering the simple biologically plausible object recognition system HMAX, from MIT [59], and the recognition with support vector machines. Walther et al. [72] continued these investigations, also in combination with the HMAX object recognition model. In [73], they combine the system with the well-known recognition approach of Lowe [35] 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 [8], [23], [24]. It was shown how the processing can be distributed among different CPUs enabling a fast, parallel computation.

1.1.7 Hamker [18]

The attention system of Hamker [18, 17] aims mainly at modeling the visual attention mechanism of the human brain. Its objective is more on explaining human visual perception and gaining insight into its functioning than on providing a computational implementation. The model is based on current computer models [32], [27]. Hamker's model, shown in Figure 1.5, shares several aspects with the architecture of Itti et al. [27]: contrasts are computed 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 addition to this bottom-up behavior, the system belongs to the few existing ones that consider top-

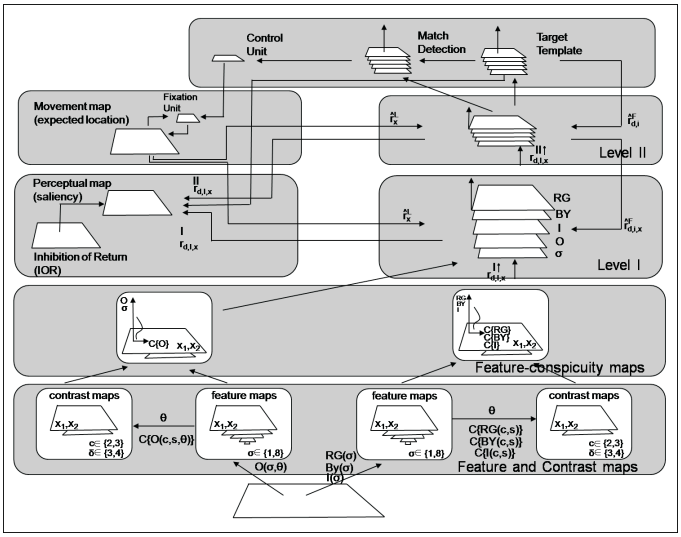

Figure 1.5: The attention system of Hamker [18]. From the input image, several feature and contrast maps are computed and fused into feature-conspicuity maps and finally into the perceptual map. Additionally, target information influences the processing. Match detection units determine whether a salient region in the perceptual map is a candidate for an eye movement.

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 bars, 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.

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 few simple features and does not consider spatial configuration of features, and lacks rotation invariance.

1.1.8 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 above described approaches, for example, they consider additional features. Among them we can refer to the work of Backer [2], who presents a model of attention with two selection stages. The first stage resembles standard architectures like [32], but the result is not a single focus but a small number (usually four) 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

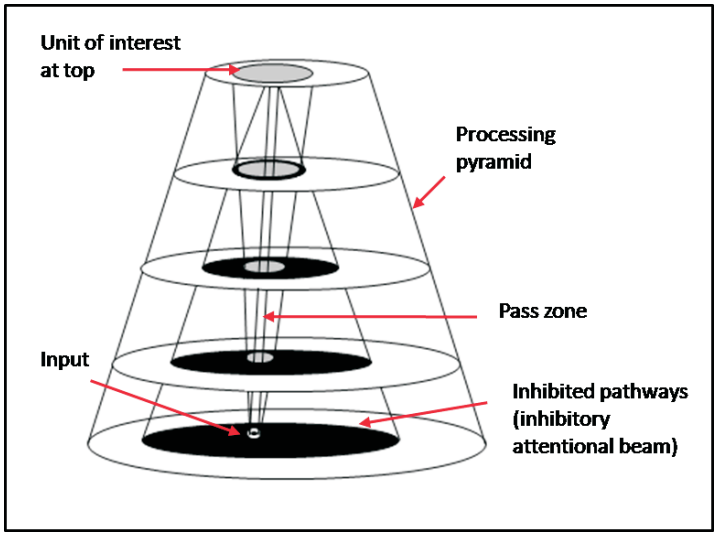

Figure 1.6: The *inhibitory attentional beam* of Tsotsos et al. [69] The selection process requires two traversals of the pyramid: first, the input traverses the pyramid in a feedforward manner. Second, the hierarchy of WTA processes is activated in a top-down manner to localize the strongest item in each layer while pruning parts of the pyramid that do not contribute to the most salient item.

experimental data on multiple object tracking and object-based inhibition of return.

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, they are mentioned only briefly.

One of the most known models in the field of *connectionist models* is the *Selective Tuning Model* of visual attention by Tsotsos et al. [69], [70] (Figure 1.6). It consists of a pyramidal architecture with an *inhibitory beam*. The 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 [42] is not considered. The model has been implemented for several features, for example luminance, orientation, or color opponency [69], and currently in a sophisticated approach also for motion, considering even the direction of movements [70]. Note that in each version only one feature dimension is processed; the binding of several feature dimensions has not yet been considered but is subject for future work, as per Tsotsos.

An unusual adaptation of Tsotsos's model is provided in [54]: 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 [6]. 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.

1.2 Objects and Attention

1.2.1 Space-based vs. Object-based Visual Attention

Visual Attention is a complex and extensive process, a set of processes that is difficult to define precisely. Some of the central aspects of our everyday notion of attention are reviewed by Pashler (1998) [61]. Intuitively, attention seems to be an extra processing capacity which can both intentionally and automatically select - and be effortfully sustained on - particular stimuli or activities. As such, we can roughly define visual attention as the mechanism that allocates limited visual resources for processing selected aspects of the retinal image more fully than non-selected aspects [48]. By using this intelligent visual selection, the visual system can flexibly explore the contents and layout of a complex visual field [36].

In the vast psychophysics literature concerning visual attention [61], the nature of the underlying units of attentional selection engenders two groups of theories. Traditional models characterize attention in spatial terms, as a spotlight (or a 'zoom lens') which can move in the visual field, focusing processing resources on whatever falls within that spatial region - be it an object, a group of objects, part of one object and part of another, or even nothing at all, as described in the previous section. Recent models of attention, in contrast, suggest that (in some cases) the underlying units of selection are discrete visual objects, and that the limits imposed by attention may then concern the number of objects which can be simultaneously attended.

In the following we present the most influential evidence for spatial selection, evidence for object-based attention, extracted from experimental paradigms. This includes selective looking, divided attention, attentional cuing, and multi-element tracking. With this short review, we emphasize on the major themes in the study of objects and attention, without exhaustively discussing empirical details. The reader can refer to the review papers of Driver and Baylis [11], Kanwisher and Driver [31], and Scholl [61].

Evidence for spatial selection

The contrast between objects and locations is the main motivation driving the study of object-based attention. Does attention always select spatial areas of the visual field, or may attention sometimes directly select discrete objects? The canonical evidence for spatial selection, which gave rise to the dominant 'spotlight' and 'zoom lens' models for spatial attention, comes from spatial cuing studies. Posner, Snyder, and Davison [52], for instance, showed that a partially valid cue to the location where a target would appear, speeded the response to that target, and slowed responses when the cue was invalid and the target appeared elsewhere.

These types of results suggested that attention was being deployed as a spatial gradient, centered on a particular location and becoming less effective as the distance from that location increased.

Early suggestions from 'selective looking'

Some of the earliest evidence for object-based selection, came from the work of Neisser [44, 45]. Subjects, given a 'selective looking' task (two spatially superimposed movies), failed to notice unexpected events which happened in the unattended scene. As such, this early work, though subject of methodological

flaws, provides evidence that attention does not simply consist of a single unitary region of spatial selection.

'Same-object advantages' in divided attention

In studies of divided attention [13, 4, 3], it was concluded that observers were less accurate at reporting two properties from separate objects, but were able to judge two properties of a single object without extra cost. This has been termed a 'same-object advantage'.

Space-based theories cannot easily account for such results, since spatial location does not vary with the number of perceived objects. However, the interpretation of these divided attention tasks is still controversial. It has been argued that the results of these divided attention studies are due to the fact that automatic attentional spread must fill a greater area with two objects than with one [9]. The details of this interpretation still implicate object-based attention, but the mechanism responsible is seen to be automatic spread of attention.

Multiple Object Tracking

The object-based nature of attentional selection is also apparent in dynamic situations, in which object tokens must be maintained over time (multiple object tracking (MOT)). Experiments from Pylyshyn and Storm [53], and Sears and Pylyshyn [62] state that the observed tracking performance cannot be accounted for by a single spotlight of attention which cyclically visits each item in turn. Also, attention has been found to speed response times to attended objects, and this advantage appears to be target-specific in MOT. Third, it is indicated that attention is split between the targets rather than being spread among them.

Object-attention concerns objecthood and object-based selection in a spatiotemporal context. Unlike space-based theories, spatial locations that do not contain any object are not considered in attentional selection. The object-based hypothesis is based on the assumptions that perceptual (but pre-conceptual) organization of a visual scene into discrete units occurs before attention is allocated, and that attention then selects or enhances visual stimuli as organized into objects rather than undifferentiated regions of visual space [56]. The contrast with the spotlight metaphoric model is clear. Since, on the spotlight model, everything in the spotlight is assumed to be processed in parallel, features from two nearby or overlapping objects should be attended as easily as a single object, whereas on the object-based model this would not be the case. In literature (see [28] for a review), there is a well-established body of evidence in support of the idea that dividing attention between objects results in less efficient processing than attending to a single object.

It should be noted that spotlight and object-based attention theories are not contradictory but rather complementary [33, 34]. Nevertheless, the object-based theory accounts for many phenomena better than the spotlight model does. Form the above discussions on space-based and object-based attention, it seems clear, that these two notions should not be treated as mutually exclusive. Attention may well be object-based in some contexts, location-based in others, or even both at the same time. The 'units' of attention could vary depending on the experimental paradigm, the nature of the stimuli, or even the intentions of the observer. The relation between space-, feature-, and object-based attention is not yet

clear. Available evidence suggests that different, but interacting, systems may be involved (e.g. [33]).

1.2.2 Attention and Perceptual Grouping

Several research works have emphasized that scenes are organized into perceptual groups defined by the Gestalt principles of similarity (common attribute), continuity (form a completed shape), proximity (close to one another), common fate (move together), etc. [48]. In this section, we consider how the attended objects, serving as units of attention, relate to other units, including perceptual groups, parts, and visual surfaces.

Attention and Perceptual Groups

Driver and Baylis [11] (also [14]) combined perceptual grouping work with attention demonstrations, and replicated some evidence for object-based selection, when Gestalt groups are used as stimuli instead of single objects. Such evidence suggests that 'object-based' attention and 'group-based' attention may reflect the operation of the same underlying attentional circuits.

Attending to Parts

Just as multiple objects can be perceptually grouped together, so can individual visual objects be composed of multiple parts. In the study of attention, recent research has demonstrated 'same-part advantages' (section 1.2.1) for complex objects composed of hierarchical part arrangements. These studies suggest that it may be worthwhile in future work to bring the literatures on attention and perceptual part structure into closer contact [63, 71].

Attending to surfaces

The previous paragraphs considered both multi-object units such as groups, and intra-object units such as parts. Visual surfaces constitute another level of representation which can encompass both of these categories: complex objects can consist of multiple surfaces, while multiple objects can be arrayed along a single surface.

From their experiments, He and Nakayama [19] indicate that attention can efficiently select individual surfaces. In another experiment, using a cuing study (similar to that of [14]), they demonstrate that in some cases attention must spread along surfaces. He and Nakayama conclude that the visual system can direct selective attention efficiently to any well-formed, perceptually distinguishable surface. In this context 'well-formedness' must be seen as local co-planarity and collinearity of surface edges.

As the previous three paragraphs have emphasized, there may be a hierarchy of units of attention, ranging from intra-object surfaces and parts to multi-object surfaces and perceptual groups. It remains an open question whether attention to each of these levels reflects the operation of the same or distinct attentional circuits.

1.2.3 Visual Object

From the above discussion, it can be said with a fair degree of certainty, that under certain experimental conditions, the allocation of attention depends on spatial properties of visual stimuli, and that under other experimental conditions, factors of perceptual organization play a more dominant role in distributing attention [30]. At the heart of the concerns discussed below is the problem of what exactly is meant by 'visual object'. To understand this concept, we consult two approaches in the literature: the object taxonomy of Jarmasz [28]; and the coherence theory of Rensink [56]. We also consider how the visual system organizes visual stimuli into the objects used by attention and how to simulate this behavior. The simulation is achieved either by Gestalt principles, or by other low-level mechanisms that are independent of higher-level conceptual knowledge. Subsequently, we wonder if these mechanisms are sufficient or if an observer's background knowledge and current mental states are also required at this early level.

Object taxonomy of Jarmasz [28]

Jarmasz [28] proposes a four-way taxonomy of objects that can play a role in vision:

- *c-objects*: physical objects, or what philosophers call concrete particulars
- *p-objects*: mental representation of visual objects or objects of phenomenal experience
- *v-objects*: virtual objects; 2D devices that are perceived as c-objects
- *a-objects*: attentional objects; intentional objects involved in attentional selection

The object-based attention thesis can now be restated as: "the attentional system selects *a-objects* in order to create *p-objects*, which are supposed to allow a person to know about and act upon the *c-objects* and *v-objects* that gave rise to the *a-objects*." The 'objects' of object-based attention are thus a-objects.

What, then, are *a-objects*? The standard answer given by several researchers is that *a-objects* are perceptual groupings whose formation is governed by the Gestalt principles of perceptual organization [4, 33]. Until the advent of cognitive psychology, the Gestalt principles constituted the only available theory of perceptual organization, and were thus integrated into cognitive psychology by Neisser [44]. The choice of Gestalt groupings for *a-objects*, the objects that are selected by attention, was thus a natural one for object-based attention.

Coherence theory of Rensink [56]

On the other hand, the attention theory of Rensink [56, 57], based on a study of change-blindness phenomena, suggests that attention may endow structures with a coherence lasting only as long as attention is directed to it. These thoughts are formulated in Rensink's coherence theory of attention, stating:

- Prior to focused attention, low-level '*proto-objects*' are continually formed rapidly and in parallel across the visual field. These *proto-objects* can be fairly complex, but have limited coherence in space and time. Consequently, they are volatile, being replaced when any new stimulus appears at their retinal location.

- Focused attention acts as a metaphorical hand that grasps a small number of *proto-objects* from this constantly regenerating flux. While held, these form a stable object, with a much higher degree of coherence over space and time. Because of temporal continuity, any new stimulus at that location is treated as the change of an existing structure rather than the appearance of a new one.
- After focused attention is released, the object loses its coherence and dissolves back into its constituent *proto-objects*. There is little or no "after-effect" of having been attended.

According to the coherence theory, a change in a stimulus can be seen only if it is given focused attention at the time the change occurs. Since only a small number of items can be attended at any time [50, 53], most items in a scene will not have a stable representation. Thus, if attention cannot be automatically directed to the change, the changing item is unlikely to be attended, and change-blindness will likely follow. Moreover, unattended objects have limited spatiotemporal coherence. From visual search experiments, described in [57], proof is provided for the limited spatial coherence of *proto-objects*, relatively complex assemblies (by rapid grouping/segmentation) of fragments that correspond to localized structures in the world. [55] indicates that *proto-objects* are the lowest level structures directly accessible to attention, with much of their underlying detail being accessed only by deliberate effort. As such, *proto-objects* serve as the highest outputs of low-level vision, but also the lowest level operands upon which higher level attentional processes can act.

The proof for limited temporal coherence of *proto-objects* comes largely from studies on visual integration and change-blindness. Early level structures are either overwritten by subsequent stimuli or else fade away within a few hundred milliseconds, making them inherently volatile [58, 57]. Given that unattended structures have only limited spatial and temporal coherence, it follows that focused attention must provide the coherence that knits proto-objects into larger-scale objects and allows them to retain their continuity over time.

Discussion

Based on the presented evidence on object-based attention, we can state correctly that focused attention is intimately involved with the perception of objects (*c-objects* according to Jarmasz' taxonomy). Essential properties of an object include the requirement that it be discrete, be differentiated from its background, and have a coherent unity across space and time. Attention makes use of surface representations generated by early visual perceptual processes. These surface representations serve as units of object-based attention. Following Jaramsz' nomination these surface representations are called attentional objects or *a-objects*. *A-objects* are believed to be perceptual groupings whose formation is governed by perceptual organization processes. Being the only available theory of perceptual organization, the Gestalt principles are put forward as theoretical answer to the formation of *a-objects* as perceptual groupings. Once *a-objects* are formed, attention selects them, and creates a mental representation of the visual objects, the so called *p-objects*. Rensink, on the other hand, claims that the so called *proto-objects* are the units of attention. *Proto-objects* are fairly complex, rapidly formed surface representations generated by early visual perception. Attention selects a number of *proto-objects*, with limited coherence, and forms a (mental representation of a) stable object, with high coherence over space and time.

Comparing Jarmasz' view on objects as unit of attention with Rensink's thoughts, we conclude that both *a-objects* and *proto-objects* refer to the perceptual entities (groups) formed by early visual perception. From this point on, if we refer to objects as the unit of attention, we will use the term *proto-object*.

Reasoning about the simulation mechanism behind the formation of the *proto-objects*, brings us, following Jarmasz, to the Gestalt principles of perceptual organization. However, other low-level mechanisms, independent of higher-level conceptual knowledge, can also be considered to simulate the rapid, pre-attentive formation of *proto-objects*. We also emphasize the distinction between perceptual groups and objects, which could themselves be comprised of many perceptual groups.

To my opinion, the formulation by Jarmasz [28] describes best the relationship between object-based and space-based attention. Attention is not a reflexive mechanism based on any particular property or set of visual stimuli. Instead of being space-based or object-based, attention is space- and object-mediated. Spatial and object features are concepts used by the visual system to deploy attention.

1.2.4 Segmentation, Perceptual Grouping, and Attention

Without segmentation and grouping, object-based attention may lose its selection units. In general, segmentation processes - the processes that bundle parts of the visual field together as units - probably exist at all levels of visual processing. Some of these processes are early, using 'quick and dirty' heuristics to identify likely units for further processing. This results in a visual field which has been segmented into *proto-objects*, which are thought to be volatile in the sense that they are constantly regenerated [56]. In this scheme, the *proto-objects* serve as the potential units of attention. Once a *proto-object* is actually attended, additional object-based processes come into play. In Rensink's coherence theory [56], deploying attention to a *proto-object* gives rise to a more coherent representation of that object. It seems likely, however, that this attentional processing could in some cases override the earlier parsing characterized by the *proto-objects*. For instance, the additional attentional processing on a set of *proto-objects* may result in a higher-level representation of that portion of the visual field as a pair of intertwined objects, or as only a part of a more global object or group of objects. In general, since such processes can occur at multiple levels, 'segmentation' cannot be considered as synonymous with object-based attention. In conclusion we state that the units of some rapid and rough segmentation processes (*proto-objects*) may serve as the focus of attention, while the units of other perceptual grouping processes may be in part the result of (proto-)object-based attention.

Indeed, Mack et al. [37] and Rock et al. [60] have presented results that suggest that perceptual organization does not occur without attention [28]. Attention and visual perception are mutually dependent, likely interactive and concurrent processes. As such, perceptual grouping/organization is deeply intertwined with object-based attention and hierarchical selectivity (or multiple selective levels) by features, objects, or their hierarchically structured groupings.

Duncan [13] states, "The study of visual attention and perceptual organization must proceed together". However, one of the remaining questions is, when, where, and how the properties of an object, or elements of a grouping become a perceptive object or a grouping? Another question is, how the mutual impact between perceptual grouping and attention is evaluated or measured?

Trying to answer these types of questions, the interactions between the different processes need to be modeled. Rensink [57] posits that attentional interaction with lower-level structures is taking place via a nexus, a single structure containing a summary description of the attended object, for example, its size, overall shape, and dominant color. When a *proto-object* is attended, a link is established between it and the nexus, enabling a two-way transmission of information between these structures. Information going up the link allows the nexus to obtain descriptions of selected properties from the attended *proto-object*.

Information going down the link can in turn provide stability to the volatile *proto-object*, allowing it to be maintained or to rapidly regenerate. The nexus and its *proto-objects* form a local hierarchy, with two levels of description (object- and part-level).

In conclusion, attention influences, and is influenced by, perceptual organization in a way that favors information that is relevant to the actions and intentions of an agent.

1.2.5 Modeling Perceptual and Relevance-based Influence on Attention

The literature ([28]) argues that attention is best understood as cognitively mediated, and not a mere reflexive response to putative salient properties of visual stimuli. The central claim of this account is that, in order to do what attention is generally said to do (filter information, enhance processing, integrate visual features into unified percepts), the visual system uses basic visual features (color, shape, location, surfaces, motion, depth information, and so on) as tools in order to direct attention according to an observer's background knowledge, goals, intentions, and particular task demands. The deployment of attention itself then organizes and transforms these basic visual stimuli into a meaningful parsing of ones visual environment, through the effects noted above.

Visual saliency can attract visual attention if the current top-down attentional setting is not fully loaded (or in other words, the current attention can be gained without top-down control) [77]. In this regard, we wonder, what is the visual salience of a feature, an object, or a grouping? And what is the neural substrate to execute the saliency computation and judgement? How does visual saliency drive visual attention? The most important requirement to model visual attention in practice is how visual salience of a perceptual unit (whether a perceptual object or grouping) can be quantitatively measured, so that the saliency mapping of a visual field truly reflects its competitive situation for visual attention.

As concluded above, perceptual organization and attention are deeply intertwined favoring information relevant to the task at hand. Therefore, perceptual- (bottom-up) and relevance-based (top-down) influences on attention need to be modeled. Insight into modeling the interaction between bottom-up and top-down influences on attention, is acquired by studying existing models of object-based attention: the triadic architecture of Rensink [56], and the conative model of attention of Jarmasz [29].

The *triadic architecture* of Rensink [56]

The triadic architecture [56, 57] is depicted in Figure 1.7, and consists of three subsystems. First, the low-level vision system, which produces *proto-objects* rapidly and in parallel. The *proto-objects* result from linear and non-linear processing of the input scene and are "quick and dirty" representations of objects or object parts that are limited in space and time. Second, a limited capacity attentional system forms these structures into stable object representations. Finally, a non-attentional system provides setting information, for example, on the *gist* - the abstract meaning of a scene, e.g., beach scene, city scene, etc. - and on the *layout* - the spatial arrangement of the objects in a scene. This information influences the selection of the attentional system, for example, by restricting the search for a person on the sand region of a beach scene and ignoring the sky region. Whereas the two first modules resemble the traditional approaches of pre-attentive and attentive processing stages, the third part of the model provides some relevance-based information about the scene at hand and extends existing models in this way. As such, this model integrates low-level, rapid, rough perceptual organization; intertwined attention

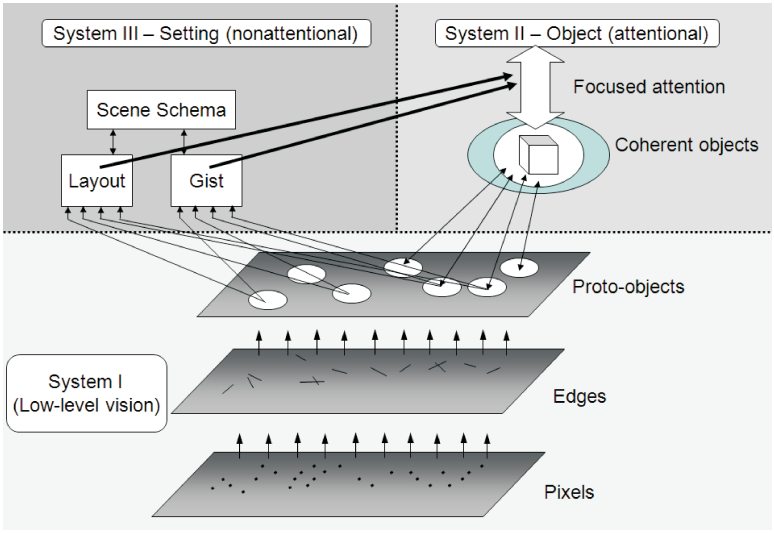

Figure 1.7: The *triadic architecture* of Rensink [56] suggests that visual perception is carried out via the interaction of three different systems: in the low level system, early level processes produce volatile *proto-objects* rapidly and in parallel. In system II, focused attention grabs these objects and in system III, setting information guides the attention to various parts of the scene.

and perceptual organization; and top-down relevance-based influences on the attentional selection task. However, in this model, these three modules are modeled as rather largely independent systems. Also, the detailed interaction between perceptual organization and attentional selection is not fully described.

Conative model of attention by Jarmasz [29]

A different model, accounting better for the integration of perceptual and relevance-based influences on attention is described in [29]. It is argued that in this model, attention uses visual objects (the products of early perceptual organization) as tools to direct and guide action, and that in doing so, attention shapes its tools. As action depends heavily on an agent's goals, motivations, and needs; themes that have traditionally been grouped under the heading of conation in psychology. The model is called a conative model of attention. As such, visual attention is assumed to interface perception and conation. A three-part architecture for attention is proposed that integrates conative, conceptual, and stimulus-driven factors in a single attentional system. Two of the constituents are the two types of determinants of attentional allocation: the stimulus-driven products of early visual perception on the one hand, and the higher-order factors, such as memory and conation, on the other. The third constituent is the mechanism that combines the two types of determinants of attentional allocation into some structure that can direct attention, corresponding to the unit of object-based attention: the *proto-object*.

Figure 1.8 illustrates the different types of determinants in the architecture [29]. The stimulus-driven determinants provide the *proto-objects* of the visual field inferred from various low-level visual cues. The organism driven determinants provide information that facilitates perceptual organization and direct attention to elements or groups of elements relevant with an organism's intentions and needs.

In conclusion, early perceptual organization provides a rough initial parsing of a visual scene into elements that the visual perception system uses to organize attention into an attentional structure,

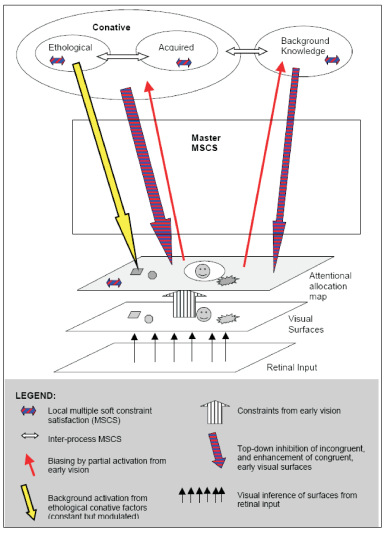

Figure 1.8: Functional architecture of a conative model of attention, as proposed by Jarmasz [29].

namely, a hierarchical organization of *proto-objects*.

In the absence of organism-driven factors, attention will be guided exclusively by the stimulus-driven cues, from which the first *proto-objects* are inferred. The stronger the cues of a particular *proto-object*, the higher the *proto-object* is placed in the attentional hierarchy. When organism-driven factors are strong, stimuli most relevant to an organism's needs and goals will receive most attention. When organism-driven and stimulus-driven constraints on attention are compatible, visual attention picks out stimuli that are relevant to an organism's intentions and actions in an effortless fashion. When the congruence between organism-driven and stimulus-driven constraints is reduced, the deployment of attention is more effortful and less efficient.

What remains as open issues at this point are the mechanisms, both neural and computational, which underlie the interrelation of organism-driven and stimulus-driven determinants of attentional deployment. These are outside the scope of this PhD.

1.2.6 Conclusion

Given the space-based attentional models reviewed in section 1.1, we can conclude that the space-based models of attention made good contributions to the implementation of location-based visual selection. However, the presented study on object-based attention identifies several limitations of space-based visual selection models, and defines primordial requirements for modeling attention to overcome these shortcomings. Research into object-based attention has received increasing interest, but research into useful systematic theories is still open research, especially computational models of object-based attention for real-world applications. We summarize the issues involved in developing biologically plausible object-based attention models as:

1. A recent study [16] shows that object-based and space-based attention share common neural mechanisms in the brain. Object-based and space-based attention are not exclusive but operate at multiple selection levels in the visual system depending on visual tasks. They achieve the coherent selection by objects, features, locations, and their groupings. Grouping is not a simple equivalent of

segmentation, but a key means to integrate both object-based and space-based attention together.

2. Object-based attention holds that the underlying unit of attentional selection is an object or a grouping of objects, features, locations, called *proto-object*. These perceptual *proto-objects* are identified and segmented early.

3. Segmentation (perceptual organization) and attention are mutually constrained and influenced [12]. Without segmentation and perceptual grouping, attention may lose its selection units.

4. Attention is controlled by bottom-up and top-down influences. This interaction, especially the top-down influence on attention, biases competition for attentional selection towards objects which are relevant to the current behavior.

5. In order to simulate human-like visual (re-)exploration behavior, an object's visual saliency should be evaluated in a spatiotemporal context. This way, visual saliency of an object varies with multiple resolutions and over time.

6. Grouping-based competition for attention should be performed by integration of object features, location features, and their distribution over the visual field. Grouping-based competition considers object-based hierarchical selectivity, involving object-based selection between objects and within an object, as required for real-world scenes [64]. Attention can work at multiple processing levels to execute selectivity by features, objects, locations, or their groupings.

7. Saliency mapping, grouping-based competition, and inhibition of return mechanism for control of attention must operate in a spatiotemporal context, for achieving human-like visual behavior in machine vision.

One issue concerning visual attention modeling has been omitted in this work, namely, the discussion about covert attentional selection and overt foveal eye movements. Visual attention covertly shifts in the visual field to select interesting objects when the fovea is fixated. Visual attention can perform visual selection without eye movements but eye movements require visual attention to function so as to assist attention to scrutinize the potential objects of visual selection in the periphery of the field of view [21]. Therefore, the shifts of attentional selection are clearly distinct from eye movements. Recently more and more active vision systems attempt to employ attentional mechanisms to help eye movements for their goal locating (e.g. [65, 2]), but research into integrating both of the shifts of (covert) attentional selection and eye movements in one system is lacking. To be complete, a biologically plausible vision system should consider foveal sensing together with visual attention but importantly make a clear distinction between them.

In conclusion, modeling and implementing object-based visual attention must engender a framework satisfying the enlisted issues and requirements. To our knowledge, only Sun [64] proposed a hierarchical object-based attention framework. This framework aims at integrating object-based and space-based attention, and employing grouping-based competition for attentional selection to achieve object-based hierarchical selectivity of visual (covert) attention and attention-guided overt saccadic eye movements. The concept of grouping is defined as the underlying unit of attention selection and is used to link object-based and space-based attention together so as to obtain hierarchical selectivity within visual attention.

A grouping is defined as a hierarchically structured unit, and can be a point, a feature, an object, a group of objects or features, or a region.

[64] provides answers and implementation details to the above requirements 1, 4, and 6, and also includes considerations on how to integrate visual (covert) attention and attention-guided overt foveal eye movements. However, early segmentation into *proto-objects*, intertwined perceptual organization and attention, focus of attention and inhibition of return mechanisms in a spatiotemporal context are not described with implementation details.

Complementary to [64], we propose in the following section our approach to object-based attention modeling, based on the foundational discussions presented in the current chapter. Our approach will formulate explicit answers, with implementation details to requirements 2, 3, 5, 6, and 7. We will only consider covert attentional selection, in a bottom-up approach, omitting (overt) saccadic eye movements and top-down influences on attention.

Bibliography

[1] U. 01. ilab at the university of southern california. http://ilab.usc.edu.

[2] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(12):1415 – 1429, 2001.

[3] G. Baylis. Visual attention and objects: two-object cost with equal convexity. *Journal of Experimental Psychology: Human Perception and Performance*, 20:208–212, 1994.

[4] G. Baylis and J. Driver. Visual attention and objects: evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, 19:451–470, 1993.

[5] C. Bundesen. A computational theory of visual attention. *Phil. Trans. R. Soc. Lond. B*, 353:1271–1281, 1998.

[6] K. R. Cave. The featuregate model of visual selection. *Psychological Research*, 62:182–194, 1999.

[7] K. R. Cave and J. M. Wolfe. Modeling the role of parallel processing in visual search. *Cognitive Psychology*, 22(2):225–271, 1990.

[8] D. Chung, R. Hirata, T. N. Mundhenk, J. Ng, R. J. Peters, E. Pichon, A. Tsui, T. Ventrice, D. Walther, P. Williams, and L. Itti. A new robotics platform for neuromorphic vision: Beobots. *Lecture Notes in Computer Science*, 2525:558–566, 2002.

[9] G. Davis, J. Driver, F. Pavani, and A. Shepherd. Obligatory edge assignment in vision: the role of figure and part segmentation in symmetry selection. *Vision Research*, 40:1323–1332, 2000.

[10] B. Draper and A. Lionelle. Evaluation of selective attention under similarity transformations. *Computer Vision and Image Understanding*, 100(1-2):152–171, Oct-Nov 2005.

[11] J. Driver and G. C. Baylis. *The Attentive Brain*, chapter Attention and visual object segmentation, pages 299–325. Cambridge, MA: MIT Press, 1998.

[12] J. Driver, G. Davis, C. Russell, M. Turatto, and E. Freeman. Segmentation, attention and phenomenal visual objects. *Cognition*, 80:61–95, 2001.

[13] J. Duncan. Selective attention and the organization of visual information. *J. Exp. Psychol.*, 113:501–517, 1984.

[14] R. Egly, J. Driver, and R. Rafal. Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–177, 1994.

[15] C. W. Eriksen and J. D. S. James. Visual attention within and around the field of focal attention: a zoom lens model. *Perception and psychophysics*, 40(4):225–240, 1986.

[16] G. R. Fink, R. J. Dolan, P. W. Halligan, J. C. Marshall, and C. D. Frith. Spacebased and object-based visual attention: shared and specific neural domains. *Brain*, 120:2013–2028, 1997.

[17] F. Hamker. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance*, 100(1-2):64–106, Oct-Nov 2005.

[18] F. H. Hamker. Distributed competition in directed attention. In *Proceedings in Artificial Intelligence, Vol. 9. Dynamische Perzeption*, pages 39–44, Berlin, 2000. G. Baratoff and H. Neumann.

[19] Z. J. He and K. Nakayama. Visual attention to surfaces in 3-d space. *Proceedings of the National Academy of Sciences USA*, 92:11155–11159, 1995.

[20] D. Heinke, G. W. Humphreys, and G. diVirgilo. Modeling visual search experiments: Selective attention for identification model (saim). *Neurocomputing*, 44:817–822, 2002.

[21] J. E. Hoffman. *Attention*, chapter Visual attention and eye movements, pages 119–154. Psychology Press, 1998.

[22] G. W. Humphreys and H. J. Müller. Search via recursive rejection (serr): A connectionist model of visual search. *Cognitive Psychology*, 25:43–110, 1993.

[23] L. Itti. Real-time high-performance attention focusing in outdoors color video streams. In T. N. P. E. (B. Rogowitz, editor, *In: Proc. SPIE Human Vision and Electronic Imaging VII (HVEI'02)*, pages 235–243, 2002.

[24] L. Itti. *Advances in Neural Information Processing Systems*, volume 15 of *Hardware Demo Track*, chapter The Beobot Platform for Embedded Real-Time Neuromorphic Vision. MIT Press, Cambridge, MA, 2003.

[25] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[26] L. Itti and C. Koch. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, 10(1):161–169, 2001.

[27] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[28] J. Jarmasz. Towards the integration of perceptual organization and visual attention: The inferential attentional allocation model. Ph.d. prospectus, Carleton University, Ottawa, Ontario, 2001.

[29] J. Jarmasz. *Objects, Pilots, and the Act of Attending: A Conative Account of Visual Attention.* Ph.d. thesis in cognitive science, Carleton University, Ottawa, Ontario, 2003.

[30] J. P. Jarmasz. Integrating perceptual organization and attention: A new model for object-based attention. Technical report, Cognitive Science Program ad Centre for Applied Cognitive Research, Carleton University, Ottawa, Canada, 2002.

[31] N. Kanishwer and J. Driver. Objects, attributes, and visual attention: which, what, and where. *Current Directions in Psychological Science*, 1:26–31, 1992.

[32] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[33] N. Lavie and J. Driver. On the spatial extent of attention in object-based selection. *Perception & Psychophysics*, 58:1238–1251, 1996.

[34] G. D. Logan. The code theory of visual attention: an integration of spacebased and object-based attention. *Psychological Review*, 103(4):603–649, 1996.

[35] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[36] A. Mack and I. Rock. *Inattentional Blindness*. Cambridge, MA: MIT Press, 1998.

[37] A. Mack, B. Tang, R. Tuma, S. Kahn, and I. Rock. Perceptual organization and attention. *Cognitive Psychology*, 24:475–501, 1992.

[38] F. Miau and L. Itti. A neural model combining attentional orienting to object recognition: preliminary explorations on the interplay between where and what. In *Proc. IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 789–792, 2001.

[39] F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. *In: Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, 4479:12–23, 2001.

[40] R. Milanese. *Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation*. PhD thesis, University of Geneva, Switzerland, 1993.

[41] R. Milanese, H. Wechsler, S. Gill, J. Bost, and T. Pun. Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Proc of CVPR*, pages 781–785, 1994.

[42] V. Navalpakkam, J. Rebesco, and L. Itti. Modeling the influence of knowledge of the target and distractors on visual search. *Journal of Vision*, 4(8):690a, 2004.

[43] V. Navalpakkam, J. Rebesco, and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.

[44] U. Neisser. *Cognitive Psychology*. New York: Appleton-Century-Crofts, 1967.

[45] U. Neisser and R. Becklen. Selective looking: attending to visually specified events. *Cognitive Psychology*, 7:480–494, 1975.

[46] H. C. Nothdurft. The rol of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33:1937–1958, 1993.

[47] B. A. Olshausen, C. H. Andersen, and D. C. V. Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neuroscience*, 13(11):4700–4719, 1993.

[48] S. E. Palmer. *Vision Science-Photons to Phenomenology.* Cambridge, MA: MIT Press, 1999.

[49] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[50] H. E. Pashler. Familiarity and visual change detection. *Percept. Psychophys.*, 44:369–378, 1988.

[51] R. H. Phaf, A. H. C. van der Heijden, and P. T. W. Hudson. Slam: A connectionist model for attention in visualseleciton tasks. *Cognitive Psychology*, 22:273–341, 1990.

[52] M. E. Posner. Orienting of attention. *Q. J. Exp. Psychol.*, 32:3–25, 1980.

[53] Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, 3:179–197, 1988.

[54] O. Ramström and H. I. Christensen. Visual attention using game theory. In *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 462–471, London, UK, 2002. Springer-Verlag.

[55] R. A. Rensink. Mindsight: visual sensing without seeing. *Invest. Ophthalmol. Vis. Sci.*, 39:631a, 1998.

[56] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.

[57] R. A. Rensink. Change detection. *Annual Review of Psychology*, 53:245–277, 2002.

[58] R. A. Rensink, J. K. O'Regan, and J. J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychol. Sci*, 8:368–373, 1997.

[59] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2:1019–1025, 1999.

[60] I. Rock, C. M. Linnett, P. Grant, and A. Mack. Perception without attention: Results of a new method. *Cognitive Psychology*, 24:502–534, 1992.

[61] B. J. Scholl. Objects and attention: state of the art. *Cognition*, 80(1-2):1–46, 2001.

[62] C. R. Sears and Z. W. Pylyshyn. Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*, 54:1–14, 2000.

[63] M. Singh and B. J.Scholl. Using attentional cueing to explore part structure. In *Poster presented at the 2000 Pre-Psychonomies Object Perception and Memory meeting*, New Orleans, LA, 2000.

[64] Y. Sun. *Hierarchical Object-Based Visual Attention for Machine Vision.* Phd. thesis, University of Edinburgh, 2003.

[65] B. Takacs and H. Wechsler. A dynamic and multiresolution model of visual attention and its application to facial landmark detection. *Computer Vision and Image Understanding*, 70(1):63–73, 1998.

[66] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognition Psychology*, 12:97–136, 1980.

[67] A. Treisman and S. Gormican. Feature analysis inearly vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[68] A. M. Treisman. *Attention, Selection, awareness and control*, chapter The perception of features and objects, pages 5–35. Clarendon Press, Oxford, 1993.

[69] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):p 507 – 547, 1995.

[70] J. Tsotsos, Y. Liu, J. Martinez-Trujillo, M. Pomplun, E. Simine, and K. Zhou. Attending to visual motion. *Computer Vision and Image Understanding*, 100(1-2):3–40, Oct-Nov 2005.

[71] S. Vecera, M. Behrmann, and J. McGoldrick. Selective attention to the parts of an object. *Psychonomic Bulletin & Review*, 7:301–308, 2000.

[72] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition a gentle way. *Lecture Notes in Computer Science (LNCS)*, 2525:472–479, 2002.

[73] D. Walther, U. Rutishauser, C. Koch, and P. Perona. On the usefulness of attention for object recognition. In L. Paletta, J. K. Tsotsos, E. Rome, and G. W. Humphreys, editors, *WAPCV2004: 2nd international workshop on attention and performance in computational vision*, Prague, Czech Republic, 2004.

[74] J. M. Wolfe. Guided search 4.0: A guided search that does not require memory for rejected distractors. *Journal of Vision, Abstracts of the 2001 VSS Meeting*, 1(3):349a, 2001.

[75] J. M. Wolfe. *Integrated Models of Cognitive Systems*, chapter Guided Search 4.0: Current Progress with a model of visual search, pages 99–119. New-York: Oxford, 2007.

[76] J. W. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–238, 1994.

[77] S. Yantis. *Attention*, chapter Control of visual attention, pages 223–256. Psychology Press Ltd., 1998.