KTH Numerical Analysis
and Computer Science

Visual Attention using Game Theory

OLA RAMSTRÖM

Licentiate Thesis
Stockholm, Sweden 2004

TRITA-NA-0423
ISSN 0348-2952
ISRN KTH/NA/R-4/23-SE
ISBN 91-7283-848-5
CVAP 292

KTH Numerisk analys och datalogi
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av Licentiate Thesis fredagen den 22 oktober 2004, kl 13.30 i E2, Lindstedsvägen 3, Kungl Tekniska högskolan, Stockholm.

Tryck: Universitetsservice US AB

Abstract

The question "what is on the table?" is normally simple for a human, but difficult for a machine. The problem is that the machine does not know what to search for, as no visual properties of the targets are known. Machine-vision algorithms, in general, need explicit knowledge of visual properties to perform object detection. Moreover, several visual properties must be considered to provide robustness. Such requirements make object detection computationally demanding and hence common algorithms scale poorly with respect to the number of objects and their visual properties. To address these problems a system has been developed that is inspired by findings from experimental psychology. The system is designed to search for objects on a specified place, e.g. things on a table or obstacles on a road. For such tasks many visual properties need to be processed. The presented system distributes the processing of visual properties and integrates only a relevant subset of the processed data. The relevant subset of data is found by forming object hypotheses from homogeneous regions in the scene. Hence the complexity of integrating a large set of visual properties is reduced.

This thesis first provides a survey of findings from experimental psychology, which give insight into the strategies used by the human visual system. From this survey it is clear that the processing of visual data is distributed across our visual cortex. Attentional mechanisms cooperate to fuse only a relevant subset of the data. One example of such mechanisms is object formation.

The presented system is also inspired by game theory, a field in which distributed computing and cooperation has been studied for quite some time. This thesis provides an overview of game theory and evaluates its applicability to visual attention.

The system is evaluated in the context of a tabletop scenario; detecting objects on a table in a natural environment. The evaluation demonstrates that a sparse set of data is indeed enough for object detection when the visual context is known and the scene not too cluttered.

Sammanfattning

Frågan “vad finns på bordet?” är vanligtvis enkel för en människa men svår för en dator. Problemet för datorn är att den inte vet vad den ska leta efter; den vet inte vilka visuella egenskaper de sökta objekten har. I allmänhet behöver datorseendealgoritmer explicit kunskap om visuella egenskaper för objektdetektion. Dessutom behövs många visuella egenskaper beaktas för att ge robusthet. Sådana krav gör algoritmer för objektdetektion beräkningsintensiva och begränsar deras skalbarhet med avseende på antal objekt och deras visuella egenskaper. Ett system som angriper dessa problem har utvecklats. Systemet är inspirerat av experimentell psykologi och är ämnat att söka efter objekt på ett specificerat ställe; t ex saker på ett bord eller hinder på en väg. Beräkningarna av visuella egenskaper är distribuerade och endast en relevant delmängd av data integreras. Den relevanta delmängden identifieras genom att forma objekthypoteser från homogena ytor i scenen. Därigenom minskas komplexiteten för integrationen av en stor mängd visuella egenskaper.

Denna avhandling ger en översikt av rön från experimentell psykologi, vilket ger insikter om strategier som det mänskliga seendet använder. Det är tydligt från översikten att bearbetningen av data distribueras över visuella cortex och att all data inte integreras. Flertalet mekanismer för visuell uppmärksamhet samarbetar för att integrera endast en relevant delmängd av all data. Ett exempel på en sådan mekanism är objektformering.

Det presenterade systemet är också inspirerat av spelteori, ett fält som har studerat distribuerade och samarbetande system sedan länge. Denna avhandling ger en överblick över fältet och utvärderar dess användbarhet för visuell uppmärksamhet.

Systemet utvärderas i ett bordsscenario; detektera objekt på ett bord i en naturlig miljö. Utvärderingen visar att en liten delmängd data verkligen räcker för att detektera objekt när den visuella miljön är känd och inte för rörig.

Acknowledgments

First of all, I would like to thank my supervisor Prof Henrik Christensen for giving me the opportunity to explore this fascinating research area. Thank you for your excellent advices and challenging enthusiasm. My co-supervisor Prof J O Eklund has also contributed with his profound experience in the field.

I would also like to thank the people at CVAP for creating such a great working atmosphere where the doors are open and questions are welcome. In particular my closest colleague Fredrik Furusjö has contributed with his knowledge and friendship. Thank you Ronnie for the fruitful discussions on game theory and Mårten for solving the hardware problems. Also a special thanks to Johan E, Ivan, and Jonas P who helped me relax from work with climbing.

The collaboration with the people in CogVis has been fun and informative. Prof John Tsotsos contributed more than he knows by pointing me in the right direction at a critical point in my work.

Finally, I would like to thank my family for their constant love and support and in particular my fiancee Eva who always believed in me.

This work has been sponsored by the CogVis consortium (IST-2000-29375), which is gratefully acknowledged.

Contents

Chapter 1

Introduction

1.1 Biological vision systems

In everyday life we have the impression of constantly perceiving everything in the visual field coherently and in great detail. One would normally expect to notice a salient event such as a gorilla walking across the floor while watching a basketball game. However, experimental psychologists have demonstrated that we often fail to notice unexpected salient events. Indeed, several experiments have demonstrated that only a small fraction of the visual properties of a scene is attended and consciously perceived. The vast amount of visual information that we do not attended to is unconsciously maintained as contextual information. Several attentional mechanisms use the contextual information to guide our conscious perception to aspects of a scene that is important for the task at hand.

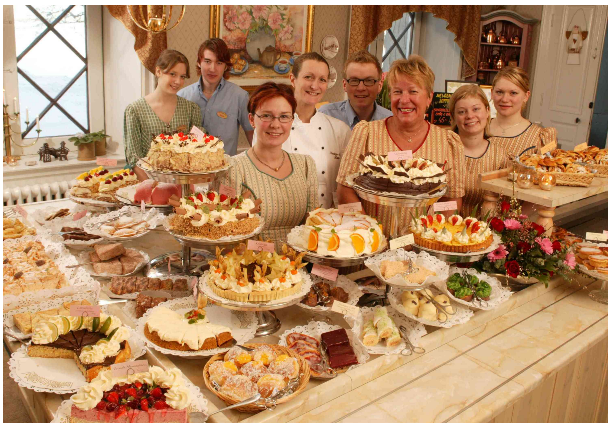

Figure 1.1: Table with cakes at Taxinge Slott

One example of such an attentional mechanism is object based visual attention; the grouping of visual information to perception of objects. The grouping can be performed at different hierarchical levels, e.g. the image in figure 1.1 can be perceived as personnel and cakes as a whole, while attending the cakes we can perceive a set of distinct cakes, or while attending one cake we can perceive its components. Object based visual attention is necessary in order to for example search for the cake that we expect to be tastiest in figure 1.1.

Object based perception and many other attentional mechanisms cooperate to maintain the visual context. The cooperation involves a vast amount of visual data and is therefore challenging to control. In the old days of experimental psychology a ghost named Homunculus was the solution to this cooperation. Homunculus was the "little man" inside who controlled all the mental processes. It is now clear that the case is quite the opposite; the control of mental processes is distributed to several areas in the brain. Also the computation of visual information is distributed and only a limited amount of data is integrated. Nevertheless, several attentional mechanisms successfully cooperate to maintain the visual context.

Depending on current task the attentional mechanisms selects only relevant visual features to our conscious perception; e.g. objects representing cakes when selecting what to eat with the coffee at Taxinge (in figure 1.1). Since we consciously perceive the relevant visual information for the current task we have the impression of constantly perceiving everything in the visual field coherently and in great detail.

1.2 Computer vision systems

Computer vision researchers have been studying visual attention for quite some time. However, natural scenes are still challenging. If targets are not globally salient, with respect to visual features across the view sphere, models of attention are limited in robustness. To gain robustness more data and features are needed. It is common in many models of visual attention to integrate all data at one point in the system, similar to the Homunculus, which is a potential bottleneck as the number of visual features is increased.

In this thesis we address this bottleneck by applying game theory to find a distributed solution. In short, the task is to search for items on a table. A color image is decomposed into a set of feature maps at individual computation nodes. At each node homogeneous regions are extracted at several redundant scales. Clusters of similar regions are formed by a negotiation scheme to form background regions. Since each background region is composed of several similar homogeneous regions, at different scales, robustness is increased. Having a set of background regions, we can extract local statistical models and a description of the layout of the scene, as visual context. Using this contextual information only a subset of the available data is needed to search for a target.

Interestingly, the strategy which is aimed at providing a solution for distributed computing, is similar to findings in experimental psychology on object based attention. The strategy does indeed reduce the need for integration of data. Moreover, the local target saliency is increased. However, the solution is currently restricted to scenes with limited clutter.

1.3 Outline

A model is proposed in this thesis, which is inspired from insights of experimental psychology and enables distributed computing without need for central control using game theory. A review on experimental psychology is given in chapter 2 and a review on game theory is given in chapter 3. Related work is presented in chapter 4.2 and in particular the proposed model and it contributions are discussed at the end of section 4.2.

The remaining chapters discuss details of the proposed model and its performance. In more detail, chapter 5 explains some image processing routines used by the proposed model, the formation of background objects is presented in chapter 6, and the search for targets at background objects is presented in chapter 7. A reference model is discussed in chapter 8 and the results of an evaluation are presented in 9. The properties of the model are finally summarized in chapter 10.

Chapter 2

Experimental psychology

2.1 Introduction

The proposed model is inspired from findings in experimental psychology. This chapter provides a review on a few aspects of visual attention. Other reviews on visual attention, with broader scope, can be found in for example (Palmer 1999), (Hoffman 1999), and (Harris and Jenkin 2001).

In section 2.2 the concept of attention is presented by a brief historical review. The methods for experimental psychology have emerged during the twentieth century, which is discussed in section 2.3. The first aspect of visual attention we will describe is space based attention, section 2.4. This provides an important introduction to object based attention, which is presented in section 2.5. In section 2.6 some experiments, which reveals how poor our conscious perception is, are presented. The experiments demonstrate that instead of having a rich conscious representation of the visual environment, certain aspects of the visual context are maintained. Section 2.7 describe some aspects of visual context. All of the above mentioned aspects of attention can be categorized as covert attention, visual attention without moving the eyes, head, or body In section 2.8 we will describe the opposite category, overt attention.

2.2 An historical description of attention

(James 1890) is an early treatise on psychology. In one chapter attention is addressed and it is one of the early contributions to psychological study of attention. Attention is defined as:

> “Every one knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition

> which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German."

Several important aspects of attention are discussed, such as our ability to direct our attention endogenously to certain events and regions, and that it also can be attracted exogenously by for example a sudden sound. Attention can furthermore be tuned by experience and expectations, which is exemplified by the following illustrative example:

> "A faint tap per se is not an interesting sound; it may well escape being discriminated from the general rumor of the world. But when it is a signal, as that of a lover on the window-pane, it will hardly go unperceived."

What is in common for these types of attention is that our ability to perceive, conceive, distinguish, remember, and react is improved.

2.3 Methods in experimental psychology

In (James 1890) and many other studies during the nineteenth-century much of the insights of attention came from introspection. Introspection is inspection of one's own thoughts and feelings. However, many relevant aspects of visual attention are processed unconsciously and is hence not susceptible to introspection. During the twentieth-century introspection has been questioned and is used with caution nowadays.

During the twentieth-century new methods to study attention experimentally emerged. (Posner 1978) provides an extensive review of results from different methods of experimental psychology with focus on performance methods. Performance methods measures the reaction time (RT) or accuracy in performing a task. A typical performance experiment is to measure RT when searching for a target among distractors of different kinds in a display. By varying the number of distractors (set size) the complexity of the search task is controlled. In a complex search task distractors are hard to distinguish from the target. Hence, the RT increase fast with the number of distractors and the slope of the RT × set size function is large, as illustrated in figure 2.1. If that slope is zero, the reactiontime is independent of number of distractors, the stimulus is is said to pop-out.

Most of the experiments in the remainder of this review will be based on performance.

2.4 Space based attention

Processing features

You might have experienced the difficulty in naming the color of a word printed in a conflicting color or similar tasks, such as the the one in figure 2.2:

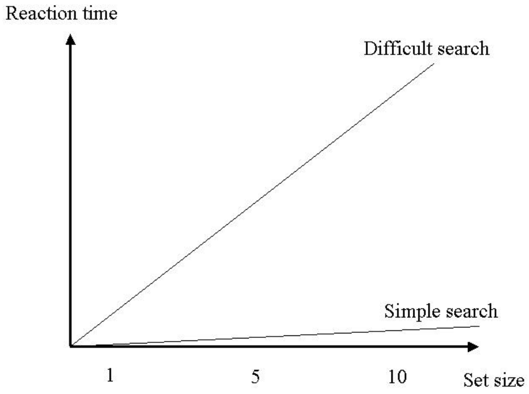

Figure 2.1: Performance experiment, reaction time × set size

Count the number of items on each row below:
aaa
bbbb
2
11
44444

Figure 2.2: An example of the Stroop effect

This is called the Stroop effect from (Stroop 1935) who investigated this effect. Interestingly, it was found that the reverse, reading the words printed with conflicting color, does not give any significant effect.

Stroop draw the conclusion that humans are used to read text of different colors and hence, the word stimuli has no association to color. But color stimuli have been associated to various responses, among these to identify the object.

(Garner 1974) gave an explanation to this effect by introducing the concept of separate and integral features.

The light that is received by the retina is processed to a set of features. Some features are separate, meaning that we can attend to one feature without being distracted by variations of another separate feature; e.g. we can attend to the color of a set of objects without being distracted by variations in their height. Other features are integral, e.g. we cannot attend to the width of a set of objects without being distracted by variations in height.

When we classify a set of objects we compare objects that differ in integral features using Euclidean measures but classification of objects differing in separable features is more complex.

For example, the boxes in figure 2.3 can easily be grouped either in columns by orientation or in rows by size, since they are separable features. However, in figure 2.4 the columns are defined by identical width and rows by identical height which are integrated features and hence grouping into columns and rows is harder.

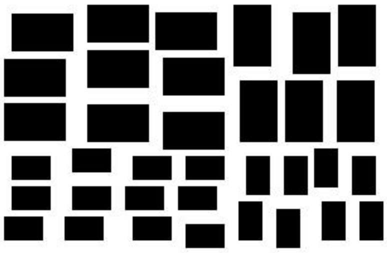

Figure 2.3: Difference in separable features

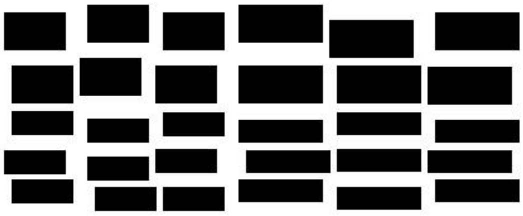

Figure 2.4: Difference in integral features

If we add extra integral features to the classification they will affect our ability to classify. If their variations are correlated with the original features, the classification performance is enhanced but if the extra integral features are uncorrelated the classification performance is decreased. However, extra separable features do not affect classification performance significantly.

Note, that we can intentionally perceive difference between objects that differs in separable features as if they were integral, but that includes cognitive processes and are therefore slow.

Moreover, according to the Stroop effect there is no clear cut between integral and separable features. Some features can be integral to some degree and the distinction is not symmetric; form (text is defined by form) is separable from color but color is integral, to some degree, with form.

Groups of integral features are denoted feature dimensions, e.g. color is a feature dimension with red, green, and blue as features. Other examples of feature dimensions are brightness, size, and orientation. It turns out that we can search for a target that is unique in one feature dimension regardless of how many other distracting items there is in a display. However, if the target is unique only with respect to a conjunction of two or more features the search time becomes proportional to the number of distractors; we have to move our attention over smaller regions of the display to increase our discriminability.

Before we discuss the role of feature dimensions in visual search we will take a quick look at strategies to move attention.

Spotlight

We can move our attention by moving the head or the eyes, overt attention. We can also move attention without moving any bodily parts, covert attention. Covert attention can be deployed much faster than overt, but overt attention provides higher acuity by bringing the target onto the foveal vision. Covert attention has been thought of as a spotlight that illuminates a part of the view sphere.

(Eriksen and Hoffman 1972) and (Eriksen and Hoffman 1973) investigated the size of the spotlight in a series of experiments. Subjects were instructed to identify letters as fast as possible in a display. Before a letter appeared on the display a visual cue indicated its location. The experiments demonstrated that the RT to identify letters at cued locations was impaired by adjacent letters and the identity of the adjacent letter did affect RT as well. The results indicated that the spotlight is of variable size but at least 1^o wide.

(Posner 1980) and (Posner, Snyder, and Davidson 1980) elaborated the concept of the spotlight metaphor and attention at cued locations. From a series of experiments it was concluded that prior information about the target shape does not improve RT, only information about location does. The enhanced performance at the cued location deteriorated with time, after 300 ms RT at the original fixation point was faster than at the cued location. If the cue on the other hand indicates a non-target location, an invalid cue, the RT was increased in simple search tasks.

These results can be summarized as: At preparation of a saccade the attention is directed as a spotlight to the target location and after the saccade the attention is diffused to the surrounding. An exogenous cue has the same effect, even when no saccade is issued.

Furthermore, it was concluded that for shape detection, and other tasks where the resolution of foveal vision is demanded, overt attention is applied. However, for simple tasks, such as detecting a bright or dark spot, the peripheral vision suffices and covert attention is preferred due to its faster deployment. Hence, overt and covert attention is not tightly coupled, their relation is more of a functional kind: when acuity is needed a saccade is issued.

(Eriksen and St. James 1986) extended the spotlight metaphor to a zoom lens model: The zoom lens can vary the size and effect of the spotlight. The theory predicts that the density of the processing resources decrease as the size of the zoom lens increase. The theory was verified by a series of experiments where the size of a cued area was varied. Indeed the RT is decreased as the cued area was increased and it could be concluded that this effect was due to a decrease in density of processing resources rather than increased complexity. The effect of the cue increased with the time it preceded the target, but the effect was asymptotic after 100 ms.

Feature integration

(Treisman and Gelade 1980) formulated the Feature Integration Theory (FIT). According to FIT separate feature dimensions are processed in parallel across the visual field, if a target is unique in one separate feature dimension it will appear to pop-out. For example a red target among green distractors can be detected with no delay with respect to the number of distractors. However, a target defined by a conjunction of separate feature dimensions takes linear time to detect with respect to the number of distractors; e.g. a vertical white bar among vertical black and horizontal white bars as in figure 2.5. Moreover, spatial relations between features does also need linear time with respect to the number of distractors; e.g. and L and T differs only in the relation of the horizontal and vertical line. Illusionary conjunctions may occur for items that have not been attended.

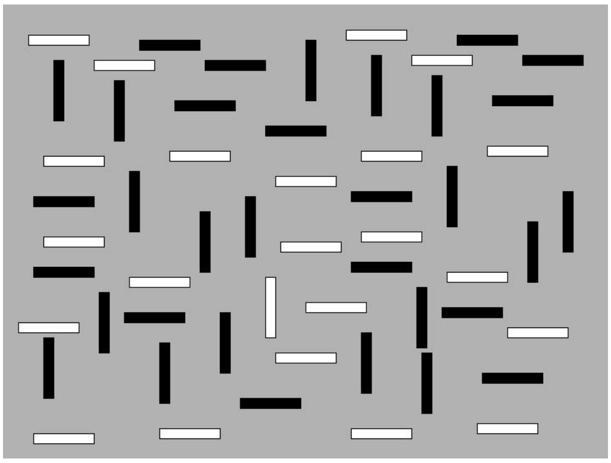

Figure 2.5: Conjugated search: “Find the vertical white bar”

The theory speaks in favor for a spotlight, when separate feature dimensions need to be integrated a spotlight need to search through the display, only within the spotlight can conjunctions and spatial relations of separate dimensions be discriminated.

Furthermore, the theory predicts that texture segregation can improve search performance, such as in figure 2.6. Our vision system preattentively segregates black from white bars, and in the black segment the white target appears to pop-out. large from small boxes, and in the segment of large boxes only one is tall. The FIT predicts that texture segregation is done preattentively when the texture is described by a unique feature dimension but not when it is described by a conjunction. Disjunctive textures (red O and green N versus blue O and green V) are more difficult to segregate than textures described by a unique feature but easier than textures described by a conjunction.

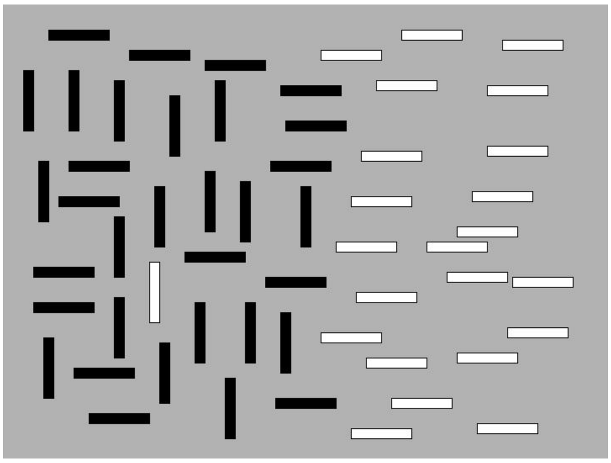

Figure 2.6: Conjugated search improved by texture segregation: "Find the vertical white bar"

(Julez 1981) continued the analysis of what features enables texture segregation. He found that we preattentively segregate textures defined by parameters from first order statistics (e.g. density and standard deviation) and textons. Examples of textons are color, terminators, and elongated blobs (lines with arbitrary width, length, and orientation). For example the 10/S symbols in figure 2.7 are easy to discriminate when attended but preattentive texture segregation is not possible since their textons are identical: same color, same number of terminators, and same number of short/long/vertical/horizontal lines.

The FIT furthermore predicts that it is harder to search for targets that are described by absence of a feature. This was however already found by (Neisser 1963), he asked subjects to search for words with or without a target letter; e.g. find the word with the letter Z below:

EIVWIM WEMVXM MZVIMW EIEVXM IEMXWE

and find the word without the letter Z below:

EIVZIM WZMVXM MMVIZW EIEVEM IZMXWE

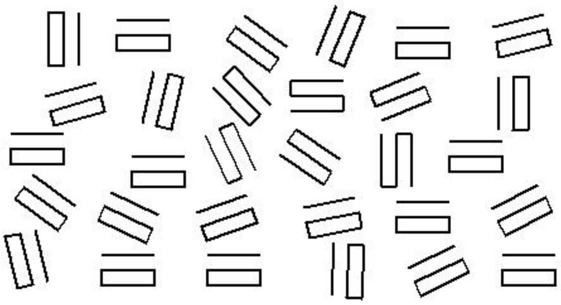

Figure 2.7: Segregation of the “S” region from the “10” region is not possible without attentional scrutiny

(Treisman and Gormican 1988) continued to investigate search asymmetries for a large set of features, for example it was found that:

- Search for a curved line among straight lines is easier than the opposite.
- Search for a tilted line among vertical is easier than the opposite.
- Search for an ellipse among circles is easier than the opposite.
- Search for an open circle (C) among closed circles (O) is easier than the opposite.
- Search for a pair of convergent lines among parallel line pairs is easier than the opposite.
- Search for a + or an L among vertical and horizontal lines is easier than the opposite.
- Search for a magenta, lime, or turquoise colored item among red, green, or blue items is easier than the opposite.

Their conclusion of these asymmetries is that search uses deviation from prototypes. According to the list above straight vertical/horizontal lines, closed circles, parallel pair of lines, junctions, intersections, red, green, and blue are examples of visual prototypes.

(Wolfe, Cave, and Franzel 1989) constructed experiments which contradicted some predictions by FIT. When the features of target and distractors are very conspicuous the search performance is better than FIT predicts, it is argued that a guiding mechanism aids the search. Consider the example above with conjugated search for a white vertical bar among distractors with conjugated features. If all positions with white and vertical are exited additively and used to guide attention then the target will be detected faster than what FIT predicts. The mechanism can equally well be explained with inhibition of non-targets. This mechanism is denoted Guided Search. An explanation to why this was not detected by (Treisman and Gelade 1980) can be that the features was not conspicuous enough and the guiding was swamped by noise. Guided Search furthermore predicts that conjugated search defined by more than two features improves search performance. With more features more information can be processed to guide the attention. And indeed, search for targets described by three features among distractors with conjugated features was found to be faster than search for targets describe by only two. However, guided search cannot improve search for targets discriminated only by spatial relation, such as an L among T.

(Enns and Rensink 1990) found even more troubles with FIT and Guided Search. Not only simple features are processed preattentively, also complex relations of lines and shades describing orientations of 3D objects are processed preattentively across the visual field. One cube appears to pop-out in figure 2.8. They investigated the reason for this effect by measuring the impact of different modifications of features: When shading was removed (figure 2.9 leftmost object) the pop-out effect was impaired but still present. If only the Y-junction (figure 2.9 middle object) or the shading (figure 2.9 rightmost object) of corresponding orientations was used no pop-out effect was present. This led to the conclusion that the spatial relations of features are processed preattentively, rather than a feature which can be implemented as a filter response. This is a direct contradiction to the FIT and Guided Search; which predicts that targets discriminated by spatial relations need linear search time with respect to number of distractors, such as L among T. Spatial relations for 3D orientation seem to be a prioritized task for preattentive processing whereas spatial relations of L and T objects are not.

(Treisman and Sato 1990) revised the FIT and acknowledged that different attentional strategies is used dependent on the environment. Search can be improved by grouping non-targets and inhibit such regions iteratively. Moreover, search for targets among distractors with conjugated features can be faster if one feature is more common than the other. Consider the example above with search for a white vertical bar, if there are less white bars than black, it is faster to search for a vertical bar among the set of white bars.

This is consistent with Guided Search ((Wolfe, Cave, and Franzel 1989)) but puts more emphasis on the inhibition interpretation.

It should be noted that preattentive feature processing is not believed to be a hard coded filter bank generating a saliency map at a specific location in visual cortex, as many attentional models implies (see section 4.1). Instead it is considered by many researchers to be a functional mechanism that can be tuned to specific tasks involving many cortical areas ((Di Lollo, Kawahara, Zuvic, and Visser 2001) and (Wolfe and DiMase 2003)).

(Bundesen 1998) have formulated a mathematical theory which explains the Guided Search. The theory predicts that filters preattentively compute the likelihood of an item $x \in S$ to belong to a visual category $i \in R$ as $\eta(x, i)$; where S is the set of items in a display and R is the set of visual features. Visual categories

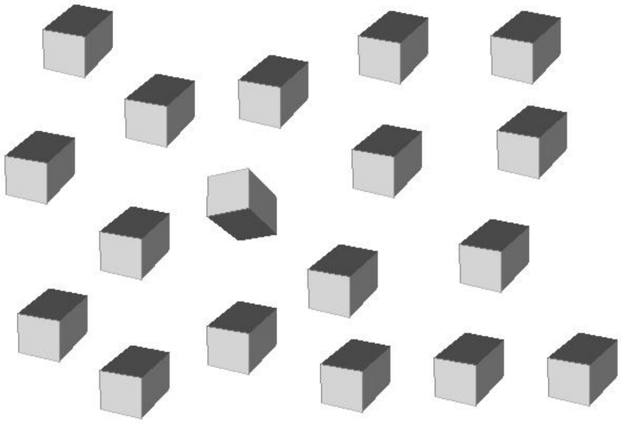

Figure 2.8: Pop-out due to 3D orientation

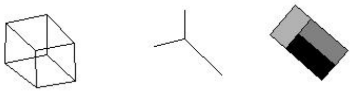

Figure 2.9: Cube without shading impress pop-out effect, only Y-junction or shading does not create any pop-out effect

are weighted by two factors: β_i and π_i where again $i \in R$. The processing power allocated to compute if element x belongs to the visual feature i is

$$v(x,i) = \eta(x,i)\beta_i \sum_{j \in R} (\eta(x,j)\pi_j) N(\eta, \pi, R, S) \tag{2.1}$$

where $N(\eta, \pi, R, S)$ is a normalization factor. The consequence is that π can act as a guiding mechanism and β as a selection mechanism; e.g. if we search for red digits and $\pi_{red} = 1$, $\pi_{digit} = 0$, $\beta_{digit} = 1$, all non-red digits are inhibited and attention is guided to red digits.

Furthermore, $\eta(x,i)$ can be extended to include texture segregation. All filter responses leak to adjacent locations with an exponentially decaying factor $c(x,y)$,

$x, y \in S$, as the two peaks with solid lines in figure 2.10.

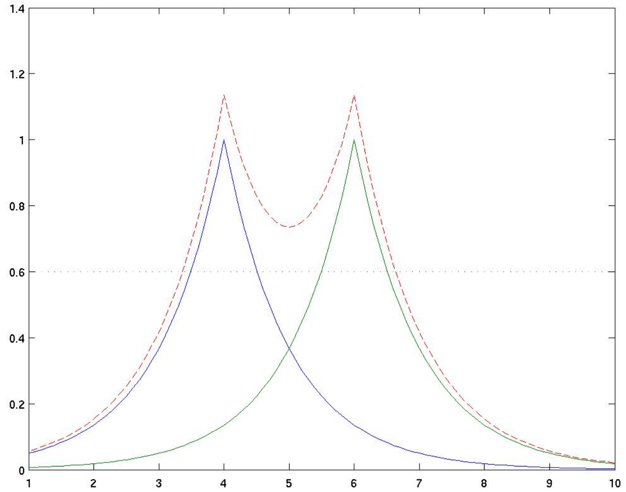

Figure 2.10: Two exponentially decaying peaks in solid lines and their sum in dashed line, illustrating the leakage factor $c(x, y)$

The preattentive filter responses are influenced by its near surrounding according to $c(x, y)$; adjacent responses are added as illustrated by the dashed line in figure 2.10. By applying a threshold we can segment a region F, segment above dotted line in figure 2.10. We can now extend $\eta(x, i)$ to be influenced by its neighbors and the segmented region F:

$$\eta_e(x, i) = a_F(x) \sum_{j \in S} c(x, y)\eta(x, j) \tag{2.2}$$

where $a_F(x) = 1$ if $x \in F$ and 0 otherwise.

Thus, a preattentive texture segregation process prunes the search to a limited region of interest. This extension also explains the effect of illusionary feature conjunctions, since it predicts feature responses to influence its neighbors.

2.5 Object based attention

Object files

(Duncan 1984) was one of the pioneer in the field of object based attention. In a series of experiments he asked subjects to make judgments about two features in a display. The results clearly indicated that subjects could more accurately detect the two features if they did belong to the same object rather than if they did belong to two different objects; e.g. the size and orientation of a box is easier to remember than the size of a box and the tilt direction of a line through it (see figure 2.11).

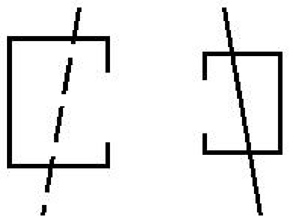

Figure 2.11: Duncan-stimuli: Tilt and texture is varied for the line, the size and orientation of the gap is varied for box. Two changes to one object are easier to detect than changes of one feature on two different objects.

(Lee and Chun 2001) extended the experiment by (Duncan 1984) to include displays where the boxes and lines were separated. It was demonstrated that the memory was unaffected by spatial separation. Thus, regardless of whether the lines and boxes in figure 2.11 do overlap or not, they are perceived as different objects when stored in the visual short-term memory (VSTM). A capacity of about four objects in the VSTM was found. Moreover, the VSTM capacity limit was only on the number of objects, no capacity limit on number of feature associated with each objects was found. (Olson and Jiang 2002) investigated this capacity of features and found that the accuracy of VSTM degrades with the number of features, but is clearly better when features can be aggregated to objects.

The memory for objects can hence be viewed as object files where features and form properties are stored. The VSTM has capacity for about four such object files.

Consequences of this was investigated by (Kahneman, Treisman, and Gibbs 1992). They also give an illustrative description of the nature of object files by an episode where a person approaches you on the street. When the stranger is close you recognize him as a familiar friend whom was unexpected in this context. Through out this episode the approaching person remains the same individual, although his retinal size, shape, and mental label has changed. The person had not lost his individuality even if he would have been occluded for some time. Perception appears to define objects by spatio-temporal constraints rather than their sensory properties or labeled identity.

An object file is addressed by its position at a given time, as we observe the object more information about its properties is stored in the file, and already stored information might be updated.

Visual objects are grouped hierarchically; a group of dancers can be a visual object, as can an individual dancer. The resolution of an object file is limited, therefore the perception of the individual dancers is coarse when a group of dancers are observed as a visual object. There is also a limit to the number of object file that can be maintained simultaneously. Therefore observing the individual dancers puts a limit to how many dancers can be observed at a given time.

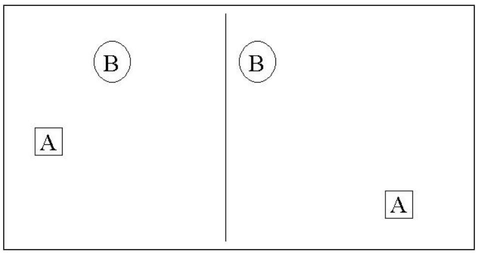

Figure 2.12: The letter A is easy to follow due to object-correspondence

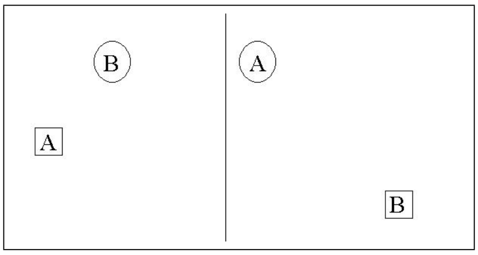

Figure 2.13: The letter A is harder to follow due to lack of object-correspondence

As the scene changes three distinct operations are needed to provide perceptual continuity through change:

1. Correspondence, determine if each object is new or if an old that has been moved.

2. Reviewing, retrieve the information history of the object.

3. Impletion, combine current and reviewed information to produce a percept of motion.

This is illustrated in figure 2.12 and 2.13 where the left square illustrate iteration 1 and right square illustrate iteration 2. In figure 2.12 it is easy to follow the letter A; the correspondence operation in iteration two determines that the square with the letter A is the same in the previous iteration, the reviewing that the object previously was located further up and to the left, the impletion combines this information to a right downward motion. In figure 2.13 it is harder to follow the letter A although it is spatially closer in iteration 2; the reviewing operation determines that the object has disapeared and the letter A is found in a new object.

A series of experiments clearly demonstrated that a new visual stimulus is identified faster if it matches a previous stimulus seen as a part of the same perceptual object. The same perceptual object could be a letter at a constant location, a letter inside a moving square, or one of two letters with constant relative spatial relation as they were both moved. Constant colors, shapes, and other feature properties could not contribute to the perception of an object in these experiments.

As the number of perceived objects increased, the benefit of prior presentation was reduced, and if the identity of the perceived object changed in subsequent presentations the benefit was wiped out.

(Hommel 2002) demonstrated by a clever experiment that object files are not limited to visual features, they also include actions associated with the object. Subjects were instructed to react with a right of left button depending on symbol identity. The RT was significantly longer if a symbol, which was instructed to be associated with a right button, was displayed on the left side of the display compared to when it was displayed on the right side. It was also demonstrated that this object-action association was spontaneous and stored in the VSTM.

Object versus space based attention

The experimental results by (Duncan 1984) were revised by (Lavie and Driver 1996) using different stimuli. The task was to identify differences at the end of two intersecting bars. The new experimental result agreed with (Duncan 1984); when the difference was at the same bar the reaction time was faster than when the difference was on different bars, but spatially closer ends. However, they found that object based attention could be eliminated. By using a visual cue, which narrowed the focus of attention spatially, the RT times was best describe by a spatial based attention mechanism. What they did not mention explicitly was that in this experiment the stimulus duration was reduced from 177 ms to 130 ms. (Law and Abrams 2002) has shown that this reduction in stimulus duration was the main effect for elimination object based attention. They used similar intersecting bars, varied endogenous and exogenous cues, and stimulus duration. Their results demonstrated that when a spatial location is cued (endogenously or exogenously) spatial proximity is better than same-object advantage but the same-object advantage was always clearly observed. Only when the stimulus duration was reduced from 186 ms to 129 ms the same-object advantage was eliminated.

(Egly, Driver, and Rafal 1994) elaborated the details of object based attention versus space based attention further. A stimulus consisting of two parallel vertical bars was cued at on end of one bar. After a brief interrupt one end of the two bars was filled. If the filled end coincided with the cued end RT was best. If the filled end was at the opposite side of the same bar RT increased. RT increased even more if the other bar, but spatially closer end, was filled. Thus, attention degrades with distance from cued location within one object but is better than shifting to another object.

(Deubel and Schneider ress) got similar results in a manipulation experiment. Subjects were instructed to grasp one end of a bar, G in figure 2.14. The identity of a symbol at D was queried after the trial. The experiments clearly indicated that discrimination accuracy at D in the left-most cross was better than in the right-most cross. Hence, when D did belong to the same object as G it was more attended, although the distance between D and G was identical in the two cases.

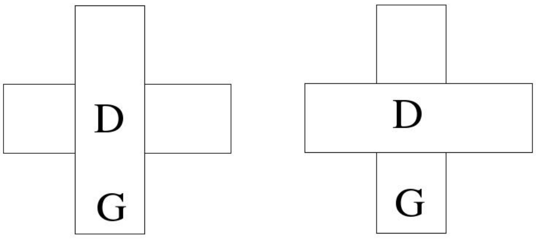

Figure 2.14: Overlaid bars with marked positions for grasp (G) and discrimination (D) location.

A natural question at this point is how the Feature Integration Theory (Treisman and Gelade 1980) and Guided Search (Wolfe, Cave, and Franzel 1989) are related to object based attention. (Wolfe and Bennet 1997) addressed this question by varying the spatial configuration of bars in a conjugated search task such as finding a red vertical line among red horizontal and green vertical lines. They found that the conjugated search performance follow the prediction of Guided Search only if they are not connected. When they are connected, e.g. in '+' configurations, the same conjugated search becomes much worse (compare figure 2.5 and 2.15). It appears as if separate features are grouped into objects and cannot be used to guide attention to likely regions as predicted by Guided search.

Hence, the features are grouped into object files and the conjugation of separate features can only be accessed by attending the individual files. Moreover, local form properties are stored in the object files but not the shape. Form is here defined as a local property (such as orientation, junction, intersection) as opposed to shape which means the global configuration of all features and form properties of the object (e.g. an L and a T both have a vertical and a horizontal line and a junction as form properties but different shape). Each object needs to be attended in turn to attain their shape. The experiments also indicated that we are able to learn new, more discriminative, form properties to improve search performance but not the shape.

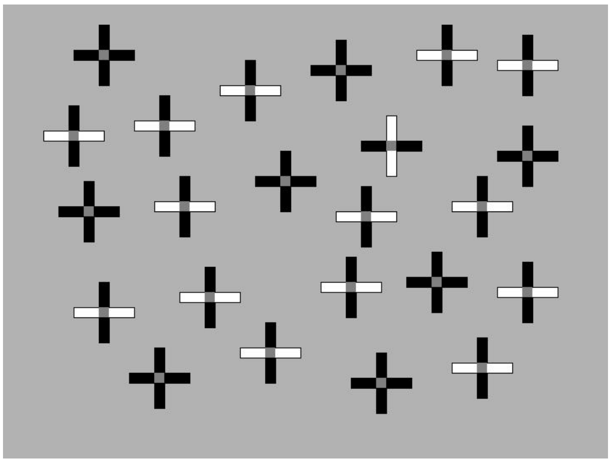

Figure 2.15: Conjugated search, such as find a vertical white bar, is harder when conjugated features are grouped to objects. The gray square at the center of each cross make us perceive it a one cross instead of two intersecting lines.

Formation of objects

(Enns and Rensink 1992) demonstrated the nature of object formation in a clever experiment. Subjects were instructed to track a rectangle with a notch in one corner. If the rectangle was originally adjacent to another object that did fit the notch it was more difficult to relocate it in a subsequent display where the other object was not adjacent (see figure 2.16). It was concluded that we preattentively compensates for occlusions, which further implies that preattentive segmentation is performed also at peripheral parts of our visual field. Interestingly, (Yeshurun and Carrasco 1998) has demonstrated the segmentation is not only performed preattentively, overt attention can even impair the segmentation performance of certain textures.

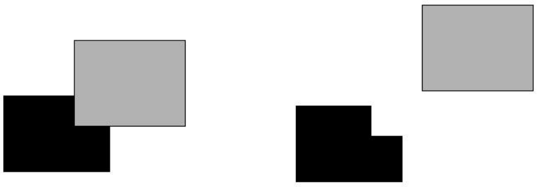

Figure 2.16: Illustration of object completion

(Marcus and van Essen 2002) demonstrated, by neural recording, that monkeys does preattentive segmentation across the whole view sphere. They could detect activation in V1 and V2, corresponding to segmentation of a region with deviating texture, although the monkeys were cued to attend another location. The time for segments to appear seemed to vary dependent on saliency, but was on the order of 100 ms, which is similar to the time found by (Law and Abrams 2002).

According to these findings the performance of object-based attention should depend on the saliency of targets. It has been demonstrated by (Duncan and Humphreys 1989) that target saliency can be dependent background homogeneity. When the target in a visual search task is salient the background homogeneity has little or no effect. However, as the target saliency is decreased, search performance becomes proportional to the similarity among distractors.

(Wolfe, Oliva, Horowitz, Butcher, and Bompas 2002) investigated further how the performance of visual search depends on target saliency and background clutter. Subjects were instructed to search for targets in displays where the background clutter was controlled. When the background clutter was moderate, the time to start the search was increased but search time per item was unaffected. Only when the clutter was similar enough to the target, the search time per item was increased. It was concluded that when the target is salient background clutter only slows the preattentive formation of objects prior to target identification, which is consistent with (Marcus and van Essen 2002). When the target saliency is low the background clutter starts to interfere with target identification as well, which is consistent with (Duncan and Humphreys 1989).

Similar results are obtained by varying the number of targets. As (Zelinsky 2001) increased the number of targets more errors was induced and the latency to the first saccade was increased. Moreover, the experiments indicated a capacity of about 5 salient items in VSTM. As the saliency was decreased the memory capacity was decreased.

Coherence Theory

The Coherence Theory (Rensink 2000a) and (Rensink 2001) gives a plausible explanation for object based attention.

Vision is divided into three stages, high-, mid-, and low-level vision. Where low-level vision performs object formation from observed visual data across the visual field. Such objects are named proto-objects and correspond to the object formation investigated by (Marcus and van Essen 2002) and (Enns and Rensink 1992). Proto-objects are volatile until the mid-level vision attends to them. Attention clusters up to six proto-objects to a coherent entity, denoted nexus. High-level vision infers identity and meaning from the attended nexus. After attention is released the nexus looses its coherence and dissolves back to its constituent set of volatile proto objects. Thus, the memory of it is lost.

The Coherence Theory gives a plausible explanation to the search asymmetries for targets described by the presence or absence of a feature, which was discussed in section 2.4. If a target is complex, as in conjugated search, the display has to be searched by attentional scrutiny. According to the Coherence Theory low-level vision will form proto-objects corresponding to items in the display, mid-level vision will group up to six of these items to a nexus, and high-level vision will decide if the target is present among the attended set of items. Consider the example where the nexus groups 5 proto-objects, each proto-object signals a '1' if the target feature is present and a '0' otherwise, and the nexus signals the sum of proto-object signals. When searching for presence of a target feature, the high-level vision can infer that the target is present when the signal from the nexus is at least 1 and absent when 0. When searching for absence of a feature the high-level vision can infer that the target is present when the signal is at most 4 and absent when 5. Hence, if the signals from the proto-objects are noisy the signal to noise ration (SNR) is larger when searching for the presence of a target feature. To improve SNR when searching for absence of a feature, fewer proto-objects has to be grouped by the nexus and the search time is thereby increased.

Furthermore, the Coherence Theory predicts that details about attended objects are forgotten after attention is released. If an object changes a visual property we will perceive it as motion, but if we are distracted at the moment of change we have to rely on visual memory to detect the change. According to the Coherence Theory we should be poor at detecting such changes. This is indeed the case and experiments reveal how little we actually consciously perceive at each moment. The phenomenon is named Change blindness and will be discussed in the next section.

2.6 Change blindness

The "real" mysteries of vision

When we study our visual system it is amazing that we can observe the world with such precision. We unconsciously change gaze direction several times per second and the eye has an irregular resolution with a blind spot; nevertheless we perceive all of the visual field as continuous, coherent, and with full resolution. This is what (O'Reagan 1992) calls the "real" mysteries of visual perception.

The answer is argued to be that we use the retina as an external memory and a tool for acquiring detailed information about what is important for the task at hand. Just as we can blindfolded perceive a bottle as a complete object although our fingers just touches a fraction of its surface, we can perceive the world as continuous and coherent although we consciously only perceive a few attended aspects at a time. Since we at any time can return to most objects in a scene, we have the impression of having full resolution all over the visual field. The blind spot, irregular resolution and saccadic effects does not affect our visual perception since the visual information is not re-presented internally in detail.

The lack of detailed internal representation is exposed by the change blindness paradigm. When a transient masks the change we often fail to notice large differences in scenes. Without the mask a change is detected easily as a movement, e.g. an item that appears on the display will act as a exogenous cue. The mask can be a saccade when comparing two adjacent pictures, a blank screen between the original and a modified display, or a a set of salient blobs appearing simultaneously with the change. In the last case the blobs do not need to hide the changing object; if there are enough many and salient blobs the perceived movement, due to the target change, is swamped by the blobs. Interestingly, this is similar to a mud-splash on a windshield when driving.

(Simons and Levin 1997) provides a review of the change blindness paradigm. Effects of change blindness are often measured using precision methods. After a change the subjects are questioned if they have perceived the change and the accuracy of change detection is measured. In the flicker paradigm the original and modified display is altered with a transient between them and the number of alternations to detect the change is measured.

(Rensink 2000b) has shown that the complexity of search for masked changes is similar to search for complex pattern, such as conjunction search.

Looking without seeing and seeing without looking

(O'Regan, Deubel, Clark, and Rensink 2000) constructed an experiment where subjects inspected a normal everyday scene while their eye movements were recorded. At each eye-blink an item was modified. The experiment indicated that although changes to central interest items were more easily detected they were often missed even when the subjects were looking directly at them. Conversely, a change to items that was not directly fixated was often detected. The phenomena was explained as subjects were "looking without seeing and seeing without looking". The same phenomena was found by (Zelinsky 2001) using natural images.

The failure of seeing directly fixated items is similar to the "blanking effect": (Deubel, Schneider, and Bridgeman 1996) found that a small displacement of a target during a saccade is not perceived, a second unconscious corrective saccade will compensate (see (McFadden and Wallman 2001) for more details on correction saccades). However, if the target is blanked for 50-300 ms immediately after a saccade the displacement is perceived. The data can be interpreted as if the target it self is preferred as reference point for the saccade, but in absence of such reference the expected target position is estimated quite accurately. (Deubel, Bridgeman, and Schneider 2002) found that occluding objects may also cause a blanking effect. Moreover, distractors that appear during a saccade can serve as a reference point and erase the blanking effect by guiding the compensating saccade to the displaced location. Furthermore, if the appearing distractor is object-like it can replace the memory of the target. This is one possible explanation to why subjects failed to see changes at fixated locations in (O'Regan, Deubel, Clark, and Rensink 2000) and (Zelinsky 2001): The post-changed object can replace the memory of the target during a saccade. Another explanation was provided by (Posner 1980), attention at fixation is dispersed after a saccade.

The Role of Expectations in Change Detection and Attentional Capture

What has been discussed so far is intentional change blindness; subjects intentionally search for changes in scenes where the change is masked. Incidental change blindness and incidental attention capture are related phenomenon; subjects often fail to notice salient changes or pronounced events if they are not expected. Examples of incidental change blindness can be found in many motion pictures; e.g. in the movie Ace Ventura: “When nature calls” (Warner Bros 1995) all pieces on a chess board disapears from one camera shot to another.

An excellent review on change blindness and attentional capture, dealing with both the intentional and incidental case, can be found in (Simons and Mitroff 2001). In an experiment on incidental change blindness a motion picture with two women having a conversation was created. Every time the camera cut to a new position a change in the scene was introduced. For example in one scene one woman has a scarf around her neck, and in the next it is removed. Although these changes were made to the persons in focus, most observers failed to detect any change at all. In another motion picture a man is sitting at a desk. The phone start ringing and he gets up to answer. As he walks to the phone the camera cuts to a new position, after the cut the man has changed identity. Although the man was the central object of this scene people failed to notice the change.

To study incidental change blindness in a real world environment an experiment was made as follows: A man stops a pedestrian and asks for direction. During their conversation two men carrying a door passes between them, one of the men carrying the door changes place with the person asking for direction. After the interrupt most pedestrians failed to notice the change.

Incidental attentional capture is related to incidental change blindness, subjects fail to notice salient events when they are occupied with a distracting task.

One approach of testing incidental attentional capture is to engage the subjects in a primary task, e.g. determine which line of a cross is the longest. After a set of trials another objects is displayed along with the primary object. The results show that the subjects often fail to detect the unexpected object. However, these tests are not very natural; meaningless object are briefly flashed in a sequence.

In an experiment using a more natural environment a movie picture was shot with people playing basketball. A black and a white team was moving around and passing a ball within respective team, the teams had one ball each. After a while a person, in a black gorilla suite, walked across the scene. When the subjects were asked to count the number of passes among the white team they often failed to detect the gorilla. Subjects counting passes among the black team had a higher detection rate. However, when watching the sequence without a task the gorilla was quite prominent. Other experiments have verified that events, which are prominent but irrelevant to the task at hand, are often not detected. The detection probability of irrelevant items is also dependent on similarity between relevant and irrelevant items. Apparently, the subjects actively inhibit conscious perception of items similar to the ignored distractors.

Inattentional amnesia

Change blindness is explained by the Inattentional Amnesia Theory (Wolfe 1999) which can be summarized as:

1. Under normal circumstances we consciously perceive visual stimuli at all locations in the visual field. This is demonstrated in (Wolfe, Cave, and Franzel 1989) where processing of the whole display participated in guiding attention in a conjugated search task.

2. At the current locus of attention, visual information can make enhanced contact with other metal processes. This permits, for instance, object recognition and transfer into memory. This is illustrated in figure 2.5, attentional scrutiny is needed to discriminate the target.

3. The present conscious visual representation is composed of the visual stimuli and the effects of attention This is illustrated in figure 2.17, attention adds the perception of a square.

4. This present conscious visual representation has no memory. It exists solely in present tense (or about 100ms). When the visual stimuli is removed or the attention is directed elsewhere, no trace of the effects of attention remain in the visual representation (however, if the subject is questioned he can remember the stimuli).

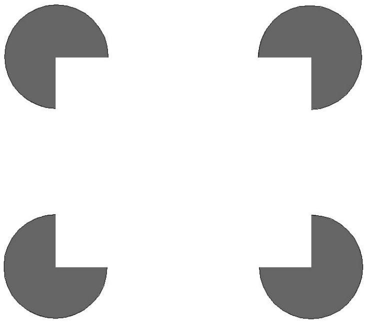

Figure 2.17: Kanizsa square - Attention adds information which gives us the illusionary perception of square.

To test bullet 4 above, a series of experiments was conducted involving visual search where subjects could, in theory, react faster by experience. However, the results indicate no improvements in reaction time although the search task was repeated hundreds of times with a similar environment.

In one experiment subjects were presented five letters in a circle and a probe letter in the center. The task was to determine if the probe letter could be found among the five surrounding letters. After hundreds of repetitions the subjects had a good conscious memory of the surrounding letters, but no improvement in RT. Thus, the visual representation was forgotten.

If vision has no memory and if attention is the gateway to other mental representations, it follows that unattended stimuli may be seen, but will be forgotten; hence inattentional amnesia.

(Mitroff, Simons, and Levin ress) demonstrates that some memory is preserved over changes. Subjects were instructed to search for masked changes in a display. When a change was detected the subjects had an accurate memory of post- and pre-change items, where the pre-change memory was better. More interestingly, also when the change was not detected the memory for both post- and pre-change items was well above chance.

Conclusion

It can be concluded that we consciously only perceive a small fraction of the visual properties in a scene and some global properties can be used to guide attention. (Simons and Levin 1997) argues that this behavior is necessary for us to handle our visual environment: Considering the perception of a busy street, people are occluded by each other, cars are passing, and you are walking along the street. A vision system that encodes with great detail would have a hard time integrating all information onto a coherent perception. Instead, a sparse coding of the visual context, with information added to items of central interest, can ease the processing of information to a coherent perception of the world.

In the next section we will discuss some aspects of visual context.

2.7 Visual context

Gist

(Biederman, Glass, and Stacy jr 1973) found that items are easier to detect if they are in an expected context, e.g. a boat is easier to detect in a sea landscape than in a city. In a set of experiments they cut pictures of normal everyday scenes into pieces The pieces were jumbled around and arranged to form a new picture where central interest items were kept intact but the background context was unrecognizable. The RT to identify specified central interest items was significantly longer in the jumbled pictures compared to the original. This implied that the global context of the picture was detected very fast and was used to guide the target search.

However, re-arranging pieces of a picture result in more effects than just confusing the background context, e.g. new artificial sharp edges are introduced.

(Potter 1976) investigated the speed of target detection and recognition. The investigation could support the assumption that the context of a scene is detected fast. It also gave interesting insights to the relation between detection and recognition. Her experiments revealed that we are able to detect a specified target among a sequence of 16 pictures presented at a rate of 113 ms per view. The detection accuracy is best if the target is specified as a picture but almost as good when specified only by title (e.g. a boat). In another experiment she investigated the corresponding recognition accuracy by specifying the target after the sequence of pictures rather than before. Since the subjects did not have a task when looking at the sequence they had to rely on their recognition when queried. The result demonstrated that the recognition accuracy is poor at such presentation rates. More than 300 ms per view was required to get recognition accuracy comparable detection at 113 ms. However, the picture memory was comparable to detection accuracy when a single picture was presented briefly and subsequently masked with a purely visual mask, i.e. conceptually meaningless picture such as salt and pepper noise.

It was concluded that we can identify a conceptually meaningful picture within 100 ms but about 300 ms of further processing is required to store the picture in long-term memory. A picture is subject to visual masking up to the point of identification, but is vulnerable to conceptual masking until the additional processing for long-term retention is completed. Since we normally change gaze direction several times per second it appears as if much more is seen and understood than is retained.

(Friedman 1979) denoted this rapid context detection as gist. In an experiment the subjects were instructed to identify deletions or changes to objects in previously viewed scenes. The result revealed that items that were unexpected with respect to the gist, were fixated for a longer period of time the first time they were seen and modifications were also easier to detect compared to expected items.

(Walker and Malik 2002) verified that subjects could capture the gist of a scene with one brief glance (37-69 ms). A model, which matches texton histogram against a memory, led to similar classification and confusions as subjects with limited processing time. It was concluded that a simple texture discrimination model explains early scene identification well.

Priming of pop-out

(Maljkovic and Nakayama 1994) and (Maljkovic and Nakayama 1996) found that the pop-out effect described by the Feature Integration Theory can be enhanced by previous presentations. This phenomenon was named priming of pop-out (PoP). Subjects were given a shape detection task (orientation of diamonds with one edge cut off). The RT was clearly tied to previous displays: If the target had the same color, spatial frequency, or position in the previous display RT was reduced. Up to 8 previous displays (or 30 sec) could influence RT, with exponentially decaying effect. Primed positions were coded as object relative rather than an absolute position in space. The PoP effect was due to target facilitation as well as distractor inhibition, although target facilitation was stronger.

(DeSchepper and Treisman 1996) had earlier found similar inhibitory effects. They investigated how much of an unattended visual stimulus is perceived, processed, and stored in memory. In one experiment a green and an overlapping red meaningless shape were displayed. The task was to identify if the green shape was equal to a white adjacent prototype. The red shape was a distractor. When a red shape became green in the next display the RT was clearly longer. This indicated that the distractor was stored in memory for inhibition; it was shown that this inhibitory memory could last for more than a month. The memory for both targets and distractors was improved for each repetition, but only the targets could be explicitly remembered when queried after a session. When a memorized distractor was presented as a target the inhibitory effect was eliminated for subsequent displays. Hence, the unattended distractor was implicitly memorized until its associated action was altered. This is consistent with the theory of object files.

Layout

According to the theory of object files the VSTM stores objects, which are indexed by their position. A relevant question is whether this spatial information is combined to perception of layout. This issue was investigated by (Fryklund 1975). It was demonstrated that the layout of targets was important for the accuracy of VSTM. The VSTM was accurate for targets arranged in a row, worse for targets arranged in an arbitrary but compound blob-shape, and poor when scattered. However, when the scattered pattern was symmetric the pattern it self was spontaneously memorized.

(Chun and Jiang 1998) investigated how the spontaneous memory for layout can be used to guide visual search. Subjects were instructed to determine the rotation of a T among rotated L in a sequence of presentations. In one experiment the presentations consisted of 12 different layouts, which were repeated in random order. The position of the T was constant with respect to each layout, but its position was random between the different layouts. This repetition of layouts significantly enhanced the performance compared to an experiment where no layout was repeated. Clearly the 12 different layouts were learned and used to guide attention. This effect was denoted contextual cueing. The contextual cueing was also observed to be rather stable; small variations in the layout did not affect the guiding and the target and distractor identities could even be changed with maintained guidance benefit.

Conclusion

The consequences of gist, PoP, and layout processing is summarized in (Chun 2000). The findings support the theory by (O'Reagan 1992) that the visual world serves as its own memory and we do not re-present the visual information internally. Gist serves to quickly give expectation on what to find in a scene. As we inspect a scene we perform many saccades where we detect and understand the view, but only a small set of items are explicitly remembered. However, we memorize much more information implicitly. This implicit memory is used to associate actions to items and to retrieve the layout of the scene if applicable. With this implicit information we relieve our conscious mind from much processing which is not directly related to the task.

Since the unconscious processing provides the necessary information to the conscious mind we perceive the visual world as continuous, coherent, and with full resolution. However, since we only have full resolution at the retina we need to change gaze repeatedly. In the next section we will discuss overt attention.

2.8 Overt attention

Relation between overt and covert attention

(Posner 1980) and (Posner, Snyder, and Davidson 1980) found that we can perform simple tasks by covert attention, which can be deployed faster than overt attention. According to the spotlight metaphor covert attention searches a display and activates overt attention only when foveal acuity is needed. This was challenged by (Findlay and Gilchrist 2001). They studied subjects in tasks which required overt attention, such as reading and complex object search. They found that saccades were made with such frequency that covert attention could not have scanned the display for new locations to fixate. The recorded scan-paths indicated that overt eye-movements were primary and covert attention only operated in association with these movements; gathering preliminary information from the next saccade location.

Clearly, the role of overt and covert attention is dependent on the task. (Johansson, Westling, Backstrom, and Flanagan 2001) have demonstrated that simple manipulation tasks can be performed without overt attention. However, the task is easier if overt attention is allowed.

Overt strategies

When overt attention is required the visual system tries to find some visual context properties to guide attention. Depending on the scene and task, different strategies are used.

In natural landscape scenes (Ouerhani, von Wartburg, Hugli, and Muri 2004) have demonstrated that search performance can be approximated by models using inhibition of return. Inhibition of return implies that previously attended locations are remembered, hence some sort of layout of the scene is perceived.

In experiments involving search for changes, researchers have reported that changes up to 10^o from gaze direction can guide attention to “jump ahead” to the changed object. However, (Henderson, Williams, Castelhano, and Falk 2004) has demonstrated that such guiding is only found in simple displays consisting of a few items on a homogeneous background, therefore changes to the gist or layout of the scene could be observed and utilized. In more complex scenes, where the change does not affect any global properties, the change need to be within 4^o of gaze direction to be detected. Hence, search strategy depends on complexity of the display; when applicable the visual context guides attention to probable locations.

(Horowitz and Wolfe 1998) studied overt attention in search tasks where it was difficult to extract any visual context. They recorded subjects scan-path while searching a display which was scrambled every 111 ms and found that the scan-path was best described by a random walk pattern. (Scinto, Pillalamarri, and Karsh 1986) used a large display of 10/S textures, which cannot be preattentively processed (see figure 2.7) and found a similar random scan-path pattern. However, (Scinto, Pillalamarri, and Karsh 1986) also demonstrated that saccade distances decrease and fixation duration increase during the search, which can be explained by the strategy: “If you can’t find what you are looking for, look more carefully”. (Hornof 2002) investigated subjects strategies when searching for target words in text. In the case of labeled text a hierarchical logical search strategy was applied: First the correct label was located, then the target word was searched. However, in the case of unlabeled text the search was near random walk. Hence, when no visual context can be derived, a random scan-path pattern is suitable. When a logical structure can be derived, such as with labeled text, the scan-path is be optimized with respect to this information.

(Johansson, Westling, Backstrom, and Flanagan 2001) have demonstrated that in manipulation tasks the scan-path is task specific. In a series of experiments they investigated the correlation between eye-hand motion when grasping and moving a bar to a target, directly or around an obstacle. It was demonstrated that the gaze preceded the hand movements. Landmarks where contact occurred (the bar and target) were fixated by all subjects. However, the contact points was always fixated prior to contact and gaze shifted away before the contact occurred. Nevertheless, the grasp and contact point was highly correlated with the previously fixated position. Obstacle was often but not always fixated, or rather an empty area next to the obstacle, where the target was supposed to pass, was fixated. The moving bar or fingers was never fixated.

(Land and Furneaux 1997) have studied overt attention in several different tasks, for example table tennis and driving. From these studies it is clear that the saccade pattern is optimized to the task. In table tennis players typically fixate points where the ball bounce, top of the net, and where the opponent strikes the ball a few milliseconds before the action. In car driving the tangent of the road is fixated in curve taking, hence the direction to drive and the only stable point. When an obstacle appears the attention quickly shifts between the tangent and the obstacle.

Conclusion

In some reaction tasks where foveal acuity is not needed, covert attention is preferred due to its faster deployment. However, in most everyday tasks such acuity is desired and we normally change fixation several times per second. The resulting scan-path varies depending on our task.

When we have a logical understanding of a task or can understand the visual context, this information is efficiently used to guide the saccades. If no such information is available, the visual search collapses to a random scan-path.

2.9 Summary

We have discussed several attentional mechanisms to extract relevant information from a scene: space based, object based, covert, and overt attention. We have also discussed some mechanisms to maintain this visual information: Object files, gist, PoP, and layout.

Thanks to these mechanisms only a fraction of the visual properties in a scene need to be consciously processed and maintained. This enable us to focus on what is important for the task at hand, without being distracted by the complexity of the visual environment.

Attentional strategies are clearly dependent on our task, prior experience, and the visual environment. The different attentional processing units seem to be distributed and cooperating in a complex way without central control. One way to model such distributed processing is game theory. Game theory will be briefly introduce in chapter 3.

Chapter 3

Game Theory

3.1 Introduction

We will in this thesis model some attentional strategies using a negotiation scheme from game theory. Therefore we will briefly introduce the subject here. More profound introduction to game theory can be found in (Osborne and Rubinstein 1999) and (Fudenberg and Tirole 1991), and (Kraus 2001). Most examples and theory are taken from these two books.

Game theory is a field that study the interaction of decision making agents. It is probably most known in economics (where game theoretic research has received several Nobel prizes) and the military.

A game is defined by a set of agents, the actions they can take, and their interests; but it does not specify what actions they do take. Agents are assumed to be rational and choose actions that are most beneficial to them, given beliefs and knowledge about the other agents. A solution of a game is a systematic description of possible outcomes.

In section 3.2 some early work on game theory are presented where agents have strictly conflicting interests. In section 3.3 we introduce a more general description of a game. In section 3.4 we generalizes further to games where the sequence of actions are included. Finally, in section 3.5 games where cooperation is necessary are presented which leads to the negotiation scheme used in this thesis.

3.2 Zero-sum games

(von Neumann and Morgenstern 1944) is often considered as the first comprehensive work on game theory.

They introduced a function Φ which maps the actions of agents to outcomes. In a two player game where agent 1 plays x and agent 2 plays y the outcome for agent $i \in \{1, 2\}$ is $u_i = \Phi_i(x, y)$, where $u_i \in \mathbf{R}$ is denoted utility. In comming examples utility will represent concrete interests such as money and abstract interests such as the value of watching a football game. It is not obvious that utility can represent an agents interest and the interested reader should refer to (von Neumann and Morgenstern 1944) for a detailed discussion on its applicability and properties. A rational agent will always strive chose an action which maximizes its expected utility. The main part of the book discusses games where the sum of all agents utility is zero, zero-sum games. In such games an agent cannot simply choose the action which maximizes its utility since the outcome depends on the other agents actions. A simple example of zero-sum game is Matching Pennies, where two agents chooses heads (H) or tails (T) of a coin. If both choos heads or tails agent 1 winns, otherwise agent 2 winns. The game is described in the table below, where the actions of agent 1 are represented by rows, actions of agent 2 by columns, and the resulting utilities are expresses as (u_1, u_2):

	H	T
H	(2,-2)	(-3,3)
T	(-1,1)	(1,-1)

We observe that agent 1 has a maximum and agent 2 has a minimum utility when both chooses heads (H,H) and agent 2 has a maximum and agent 1 a minimum utility in (H,T). If both agents only consider their own actions, agent 1 would always choose heads and agent 2 would choose tails resulting in the outcome (-3,3). Hence, agent 1 would attain his minimum utility by always striving to obtain his maximum utility without considering the opponents possible actions! It is profitable for agent 1 to solve the minmax-function:

$$u_1 = \max_x \min_y \Phi(x, y) \tag{3.1}$$

Thus, assume the opponent does the action yeilding the worst utility and choose the best response to that.

According to the minmax-function agent 1 would assume agent 2 to choose tails for which tails maximizes the expected outcome. Similarly, agent 2 would assume agent 1 to choose heads for which tails maximizes the expected outcome. The resulting outcome would hence be $\Phi(T, T) = (1, -1)$. Thus, in a game the agents actions are determined from the full set of all agents possible actions. The set of actions of the two agents is denoted an action profile, in Matching Pennies the action profile $\{T, T\}$ is an equilibrium state since it is prefered by both agents.

If agents on the otherhand have a mixed strategy, that is choosing actions with a probability, the game would have different outcomes. Assume agent 1 plays heads with probability p and agent 2 with probability q. The expected outcome according to the minmax function for agent 1 is then

$$v_1(q) = pq2 + p(1-q)(-1) + (1-p)q(-3) + (1-p)(1-q)1 \tag{3.2}$$

Since the game is strictly competitive agent 2 maximizes the expected outcome by minimizing the outcome of agent 1:

$$\frac{dv_1}{dp} = 7q - 2 = 0 \tag{3.3}$$

The expected outcome of agent 1 is minimized by selecting $q = \frac{2}{7}$, similarly the expected outcome of agent 2 is minimized by $p = \frac{4}{7}$, which yields the expected outcome $(-\frac{1}{7}, \frac{1}{7})$. Thus, with mixed strategies the agents actions is determined from the probability space over set of all agents possible actions

(von Neumann and Morgenstern 1944) proved that there always exists a minmax-solution to Matching Pennies and all other two player zero-sum games.

The interested reader might have reacted on the rather sloppy notation, refer to (von Neumann and Morgenstern 1944) for detailed discussion regarding utility- and minmax-functions and other topics regarding zero-sum games.

3.3 Strategic games

Nash equilibrium

In strategic games all agent choose their strategy once and for all, the zero-sum game presented above is an example of a strictly competitive strategic game. In the more general form of strategic games agents do no neccesarily have strictly competing interests and the minmax-function is hence not normally applicable.

The “Battle of the sexes” is a game which illustrate a case where two agents do not have strictly competitive but different interests:

> A man an a woman want to meet this evening. The man prefers to watch a football game (F) and the woman a theater (T). However, both prefers the partners choice over beeing apart.

In this game both agents desire an outcome where both agrees, which is illustrated in the table below:

	F	T
F	(3,1)	(0,0)
T	(0,0)	(1,3)

The “Prisoner’s dilemma” illustrate an other strategic game:

> Two suspects on a burglar crime are put in separate cells. Each has to choose whether or not to confess and implicate the other. If neither suspect confesses, then both will serve one year on a charge of carrying a concealed weapon. If each confesses and implicates the other, both will go to prison for 10 years. However, if one suspect confesses and implicates the other, and the other does not, the one who has collaborated with the police will go free, while the other will go to prison for 20 years.

and can be expressed in table form as:

	Confess	Don’t confess
Confess	(-10,-10)	(0,-20)
Don’t confess	(-20,0)	(-1,-1)

A common solution to strategic games is the Nash equlibrium: A strategy profile from which no agent can profitably deviate; where a strategy profile is the strategies of the set of agents.

In the example of the "Prisoner's dilemma" no prisoner can profitably deviate if one chooses Confess. Thus, although their joint preference is Don't confess, they will both Confess due to the threat that the other might Confess. In the example of "Battle of the sexes" both Football and Theater is a Nash equlibrium. Hence, the theory does not predict what they will do, only the possible outcomes to this type of game.

The strictly competitive example "Matching Pennies" does however not have a Nash equilibrium.

Mixed strategy

A Mixed strategy game was exemplified in 3.2 when an agent chosed heads with probability p and tails with probability $(1-p)$, as opposed to a pure strategy when either heads or tail is chosen, $p = \{1, 0\}$. Each agent computed their optimal mixed strategy from expectations of the other agents mixed strategy.

In the more general strategic game we can apply the Nash equilibrium to mixed strategies as well by considering the probability space of the other agents actions. It can be proven that the set of pure Nash equlibriums is a subset of the the set of mixed-strategy Nash equilibrium and that every game with a finite number of possible strategies has a mixed-strategy Nash equilibrium.

Evolutionary equlibrium

Consider a popolation of e.g. animals, humanbeings, plants, etc, which has reached an equlibrium. In evolution a small subset of a population is mutated. A mutated individual will take non-equlibrium actions. In an evolutionary equlibrium each mutated individual must loose on deviating from the equlibrium so it dies out.

An evolutionary equilibrium is a Nash equilibrium with the addition that the equilibrium actions is also preferred against a deviating mutant action.

Thus, in the normal Nash equilibrium no sigle agent can profitably deviate; in the evolutionary equlibrium no small set of agents can profitably deviate, which is a harder requirement.

Bayesian games

If the agents do not know the other agents preferences, decisions need to be based on Bayesian inference on expectations of the other agents actions. Such games are called Bayesian games.

The Nash equilibrium can be applied to Bayesian games as well, if one incorporates the probability space. In a Bayesian Nash equilibrium every agent chooses the best action given what he can observe and his beliefs about the other agents' actions according to this observation.

Consider a market with one dominant company. The company has an old plant and considers to build a new. Simultainously, another company considers the possibility to enter the market. If it is expensive for the first company to build a new plant the outcome of their choices is represented by the table to the left below, if it is not expensive the outcome is represented by the right table below.

	Enter	Don't enter
Build	(0,-1)	(2,0)
Don't build	(2,1)	(3,0)

	Enter	Don't enter
Build	(3,-1)	(5,0)
Don't build	(2,1)	(3,0)

The second company does not know if it is expensive to build a new plant or not. His choice will depend on his expectation of the building prize.

The game is simple for the first company, it is profitable to build if and only if building cost is low. The second company gains on entering if and only if the first company does not build, which is not known. Instead the second company will enter if he expects that the first company does not build.

In the example above we have two Bayesian Nash equilibriums:

- The first company doesn't build if cost is high, the second company enters if he expects building cost to be high.

- The first company builds if cost is low, the second company doesn't enter, if he expects building cost to be low.

Thus, only a subset of all information shared among the agents in Bayesian games. This is an important aspect, in most relevant everyday situations we do not have all information but does rather build our strategies on expectations. Moreover, game theory has successfully been applied to some distribute computing applications (Kraus 2001) and in such applications it is normally of central interest to limit the amount of shared data.

3.4 Extensive games

Subgame perfect equilibrium

In distributed computing and several other applications of game theory the sequence of actions is of interest. This information is omitted in strategic games.

In extensive games the history of recent actions and the order agents make their future moves are considered. With this extension Nash equilibria can still be formulated but is normally of little interest; it might be an equilibria in the beginning of a game but need not be in intermediate states of an extensive game.

Consider for example the Prisoner's dilemma example with two relapsed criminals, which and expect to repeatedly confront the same dilemma and compute their utility as average discounted utility. The average discounted utility for agents i

after iteration $t = T$ with the history of actions $\{a^0, a^1, ..., a^T\}$ is defined as:

$$u_i^T = \frac{1-\delta}{1-\delta^{T+1}} \sum_{t=0}^{T} \delta^t u_i(a^t)$$

where δ is a discount factor ($0 < \delta < 1$) and $\frac{1-\delta}{1-\delta^{T+1}}$ is a normalization factor so that a sequence of $u_i = 1$ is 1 regardless of T.

Using average discounted utility a strategy to Not confess until the other does Confess and then always Confess might be profitable. For the benefit of easier calculations we rewrite the “Prisoner’s dilemma”-table as:

	Confess	Don’t confess
Confess	(1,1)	(-1,2)
Don’t confess	(2,-1)	(0,0)

which has the same properties as before. The game has two classes of subgames: A: both choses to Confess, B: both choses to Not confess. While the two suspects conforms to class A both will obtain the utility one. A deviating action will generate a higher utility for one iteration and both will conform to class B for the rest of the game, yielding zero utility. Hence, if agent i deviates after S iterations his average discounted utility after $T > S$ iterations is:

$$u_i^T = \frac{1-\delta}{1-\delta^{T+1}}(1 + \delta + ... + \delta^{S-1} + 2\delta^S + 0 + ...)$$

which is less than one when $\delta > \frac{1}{2}$ regardless of what choices he makes after $t > S{+}1$.

Thus, in the repeated “Prisoner’s dilemma”, Confess is still a Nash equlibrium in the first iteration, but in intermediate states is is not. In extensive games the Subgame perfect equilibria is often used: A strategy profile which is a Nash equilibrium of every subgame.

Hence, in intermediate states of the game the suspects prefer a subgame perfect equilibrium where both agents chooses “Don’t confess” until the other agent deviates.

Strict dominance

If a game is played for a fixed number of iterations backward induction can be applied to exclude strictly dominated solutions. An action profile is strictly dominated if there is another profile that is prefered regardless of the other agents action. In the rewritten “Prisoners dilemma” above Don’t Confess is strictly dominated by Confess for both agents.

Thus, Don’t confess is only a subgame perfect equilibrium if the game is played for ever.

Repeated games

Repeated games studies phenomena such as cooperation, revenge, and threats. For example in the infinitely repeated "Prisoner's dilemma" cooperation (Don't confess) could be enforced by the threat to Confess the rest of the game.

However, in a similar but different game presented below

	C	D
C	(2,3)	(1,5)
D	(0,1)	(0,1)

two agents have a common interest in cooperating (C) but agent 2 can profit by defecting (D). Agent 1 can threat to punish agent 2 by playing D for the rest of the game, as in the infinitely repeated "Prisoner's dilemma". However, it is not a credible threat here since agent 1 will gain on forgiving agent 2 and try to establish a new cooperation.

The Perfect folk theorem states that for a large enough discound factor δ there exists a punishment which enforces all agents to cooperate. In the example above agent 1 should play D for S iterations so

$$5 + \frac{1-\delta^S}{1-\delta} > 2\frac{1-\delta^{S+1}}{1-\delta}$$

Thus, just enough iterations so agent 2 will loose on defecting (D).

Bargaining games

In bargaining games one agent starts by proposing an action profile, the second agent can choose either to accept or reject. Accept results in that the bargain ends and the proposition is implemented, reject gives the posibility to respond with a counter offer in the next iteration.

Normally, iterations of bargaining are costly. Hence, an offer at iteration t is prefered to the same offer at iteration $t + 1$.

Split a pie

Consider two agents bargaining on how to split a pie. The utility for agent i of receiving x of the pie at iteration t is $\delta_i^t x$; where $0 \leq x \leq 1$ and $0 < \delta_i < 1$.

Let x'_i denote the equilibrium offer by agent i, then agent i will offer $1 - x'_i = \delta_{-i}x'_{-i}$; where $-i$ denote the other agent. Thus, an offer which has the same utility as the other agent can get in the next iteration by his equilibrium offer. In the first iteration the two agents will offer:

$$1 - x'_1 = \delta_2 x'_2 \tag{3.4}$$

$$1 - x'_2 = \delta_1 x'_1 \tag{3.5}$$

which yields the solution

$$x'_1 = \frac{1-\delta_2}{1-\delta_1\delta_2} \tag{3.6}$$

$$x'_2 = \frac{1-\delta_1}{1-\delta_1\delta_2} \tag{3.7}$$

It can be proven by backward induction that x'_i are subgame perfect equilibrium. Agent i, starting the negotiation, will offer $1 - x'_i$ and the other agent will accept immediatly.

In this game the outcome is dependent on their patience δ_i, the larger patience the larger the share. Moreover, the agent starting the negotiation will have a first move advantage depending on both agents patience $(\delta_1 + \delta_2)$, the more patient both agents are, $(\delta_1 + \delta_2) \rightarrow 1$, the smaller the first move advantage is.

Resource allocation

(Kraus 2001) exemplifies a bargain game where agents have the possibility to opt out, that is break the negotiation resulting in a conflict outcome for all agents. The conflict outcome is normally the worst out come for all agents, and there is always at least one offer which is better than opt out. However, the possibility opt out changes the set of possible outcomes of a negotiation.

In the chapter "negotiation about resource allocation" a joint scientific mission to Mars by ESA and NASA (the European and American space agency) exemplifies a negotiation with the possibility to opt out:

> NASA and ESA have embarked on a joint scientific mission to Mars involving separate mobile labs launched from a single shuttle in orbit around the planet. Each country has contracts with a number of companies for the conduct of experiments. These experiments were preprogrammed prior to launch. Arrangements were made prior to launch for the sharing of some equipment to avoid duplication and excess weight on the mission. Instructions to begin each experiment must be sent from Earth.
>
> NASA's antenna was damaged during landing, and it is expected that communications between NASA and its lab on Mars will be down for repairs for one day (1440 minutes) of the planned five day duration of the mission. NASA can use a weaker and less reliable backup line, but this involves diverting this line from other costly space experiments, and thus the expense of using this line is very high to the NASA. NASA would like to share use of ESA's line during the one-day period so that it can conduct its planned research program. Only one research group can use the line at a time, and that line will be in use for the entire duration of the particular experiment.

> A negotiation ensues between the two labs over division of use of the ESA line, during which time ESA have sole access to the line, and NASA cannot conduct any of its experiments (except by use of the very expensive backup). By prearrangement, ESA is using some of the NASA equipment for their experiments, and are gaining \$5000 per minute. While ESA cannot conduct any of their experiments without some NASA equipment, NASA could conduct some of its experiments without ESA equipment. NASA is losing \$3000 per minute during the period in which they must rely on their backup communications line. An agreement between NASA and ESA to share the communications line will result in a \$1000 gain per period (minute) for each group.
>
> If an agreement on sharing the line is not reached, NASA can threaten to opt out of the arrangement. In this case, NASA will be able to conduct a small portion of its experiments by using all of its equipment and no ESA equipment, and by using the backup communications line. The overall NASA gain will be \$550,000, but it will lose \$1000 per any minute of the negotiation. If NASA opts out, ESA will not be able to continue their experiments (without the NASA equipment) and their gain will be restricted to whatever they had gained at the point NASA opted out. If ESA opt out, they will need to pay NASA \$100,000 for use of NASA equipment up to that point. A dollar is the smallest unit of currency in this example.

We denote the utility as $U^x(a,t)$ where $x \in \{\mathrm{ESA}, \mathrm{NASA}\}$, $t = 0, 1, 2, ...$ is the negotiation period, and a is the action profile. The action profile can be an agreement on how to share the communication line during the first day $s = (s_N, s_E)$, hence $s_N + s_E = 1440$, both also have the possibility to opt out, $a = (Opt_x)$. With this notation the game can be formally expressed as:

- $U^E(s,t) = 1000s_E + 5000t$, $U^E(Opt_N, t) = 5000t$, $U^E(Opt_E, t) = 5000t - 100000$
- $U^N(s,t) = 1000s_N - 3000t$, $U^N(Opt_N, t) = 550000 - 1000t$, $U^N(Opt_E, t) = -1000t$

An agreement will be reached in the second period ($t = 1$) with $s = (887, 553)$. It should be noted that there are agreements in the future that both agents prefer over reaching the agreement $s = (887, 553)$ in the second period. This is because ESA gain more over time than NASA loses over time. For example, the agreement $s = (879, 561)$ in the fourth time period (period 3) is better for both agents than $s = (887, 553)$ in the second time period. The problem is that there is no way that NASA can be sure that ESA will offer them $s = (879, 561)$ when the fourth time period arrives. In that time ESA need to offer only $s = (885, 555)$ in order to prevent NASA from opting out, and they don't have any motivation to offer more.

Incomplete information

If the agents do not have full information about each others utility function, they have to rely on their expectations on the other agents. In section 3.3 we discussed a game with a company deciding on whether to enter a market or not, based on its expectations on building prices. Two types of games was considered, one for high building prices and one for low.

The situation is similar in extensive games with the addition that agents can improve their estimates of types from the history of recent actions. If the game in section 3.3 is played repeatedly the first can observe the established actions and retreive the building price.

If both agents utilities depend on both agents actions, such as in the NASA-ESA mission, they need to guess each others type from the history. Neither agent will reveal its type, and their beleifs will change over time.

In such situations can be examined using the notion of sequential equilibrium ((Kraus 2001) chapter 4). An action profile in sequential equilibrium is a bayesian Nash equilibrium given opponents strategies, recent history, and current beliefs.

3.5 Cooperative games

Competitive equilibrium

In cooperative games agents can only take actions in coalition with others.

One example of cooperative game is a market economy. A market is a place where agents can buy and sell goods. With a set of goods each agent can produce a value. The produced value is defined as the utility. Furthermore, the utility is a concave function with respect to the amount of goods; if we have no bread we will profit much on one loaf of bread, but if we have a barn filled with bread we care less if we can get yet another loaf.

If goods can be valued with money, the utility of the market can be shared arbitrary among its members. In (Osborne and Rubinstein 1999) it is shown that a market with money exchange will reach a competitive equilibrium. A competitive equilibrium is a state where all agents agrees on a price for each type of good and the agents will buy and sell until all have the same amount of all types of goods.

Let us consider a market with N agent and k number of available goods. Agent i has a stock of goods $n_i \in \mathbf{R}^k$ and the utility $f(n_i) \in \mathbf{R}$, see illustration in figure 3.1. Since the utility function is concave, agents with a large stocks of a good will gain on selling on the market and agents that are short on a good will gain on buying.

After trading agent i will have a new stock of goods z_i, where $\sum_{i \in N} z_i = \sum_{i \in N} n_i$. Each agent will strive to get the z_i that solves

$$max_{z_i}(f(z_i) - p \cdot (z_i - n_i)^T) \tag{3.8}$$

where $p \in \mathbf{R}^k$ a prize vector.

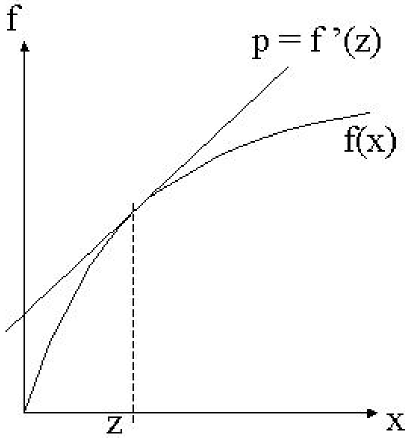

Figure 3.1: A market where the utility of x is $f(x)$. When the average allocation on the market is z the agents will agree on a prize $p = f'(z)$

We denote the average stock $\tilde{n} = \sum_{i \in N} n_i / N$. In (Osborne and Rubinstein 1999) section 13.4 it is shown that the solution where $z_i = \tilde{n}$ for all $i \in N$ and the price vector $p = f'(\tilde{n})$ is a competitive equilibrium.

Core

In the market economy example above all agents gain on cooperating, no subset of the market can be more fruitfull than the full market.

Solutions involving all agents of a cooperative game where no subset of the agents can profit on deviating from the full set of agents are denoted core. Note the similarities between core and Nash equilibrium.

It has been proven that every market with a currency, or some other means to transfer the value of the outcome of a game arbitrary among the agents, has a non-empty core. Furthermore, a competitive equilibrium is in the core of the market.

Market is of course only one example of cooperative games, in a more general game there is no guaratee for the existence of a core. Subsets of agents might form coalitions which are more profitable for its members. In such situations they will object to the solution which involves all agents. However, if there are many coalitions which might profitably deviate, an objection might unleash a process involving further objections by other coalitions. At the end some of the member of an objecting coalition might be worse off. Therefore, it is important for a coalition to be stable before objecting.

(von Neumann and Morgenstern 1944) studied requirements for a coalition to be stable, stable sets. A stable set is a set of imputations, divisions of the outcome of a game among the agents, where all other imputations outside the set will be objected to by some coalition in the set and no coalition objects to an imputation in the set.

Consider the "Three-player majority game":

> Three miners find a large gold ingot in a cave. At least two persons are needed to carry the gold. How will the men divide the gold?

A coalition of all three is not stable, since any miner can persuade any other miner to exclude the third miner. If they share the gold evenly the third miner cannot object with any coalition which is preferred by both miners in the original coalition. Indeed any coalition of two miners who share the gold evenly is a stable coalition.

When a coalition of all agents is stable it is proven that it is the only stable set. More generally, no subset of a stable coalition is a stable set.

Extensive cooperative games

Cooperative games can be modeled as extensive games as well. (Shehory and Kraus 1998) decribes some examples of extensive cooperative games whith distributed problem solvers (DPS). DPS act in order to increase the benefit for the group as a whole. The examples studies solution concepts when the task is to build a construction of a given configuration from a set of building blocks, as illustrated in figure 3.2. The building blocks are of various size and the agents need to cooperate to build the construction. When no central coordinator divides them into groups to solve subtasks, they have to rely on a distributed negotiation scheme. One such scheme forms disjoint coalitions for a current set of subtasks by iterative negotiation. Each agent calculates the value of each coalition it can form and announce the the most profitable. Each agent subsequently selects the most profitable among these. Since all agents are DPS they know that all other agents will do the same. The choosen coalition is implemented to solve a specific subtask, and the remaining agents updates their list of possible coalitions and tasks to solve. The procedure is repeated untill all subtasks have been assigned a coalition of agents.

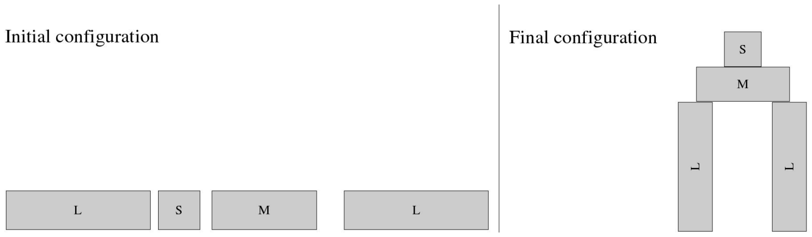

Figure 3.2: Cooperation is needed for a task involving building a construction of a given configuration from a set of building blocks

In contrast to the previous example where the prisoners resolved their dilemma by threat, companies decided strategy based on beliefs, space agencies reached an agreement due to the threat to opt out, this problem is resolved by common sense. Moreover, the agents have incomplete information of the other agents. Instead of maximizing their own utility based on estimated beliefs all agents acts according to the benefit of the group and they know that all other agents do the same.

Chapter 4

Context of the thesis

4.1 Related work

Winner-take-all

Many of the related models are inspired by the Feature integration theory, (Treisman and Gelade 1980): First a pre-attentive processing stage computes conspicuity for several feature dimensions. The conspicuity maps are subsequently integrated to a saliency map, which is used to perform a coarse to fine search for regions of interest.

(Koch and Ullman 1984) is one of the early works on implementing a computational model of the feature integration theory. Several features are computed in parallel and stored in feature maps. The feature maps are integrated to a saliency map. All pixels in the saliency maps are processed by a Winner-take-all (WTA) network, which has only logarithmic search complexity with respect to the number of pixels in the saliency map. The outcome of the WTA network is a region that is attended and subsequently inhibited. Due to the inhibition a scan-path is generated.

(Tsotsos 1990) analyzed the computational complexity of visual search and found that an attentional strategy is necessary to cope with the vast amount of information that meets the eye. An example of a dense template matching algorithm, which does not assume any specific scale, is proven to be NP-complete. By comparing algorithm complexity to an upper estimate of computational resources in the visual cortex of humans it is concluded that a pyramidal structure, similar to the WTA network, is necessary. An attentional model with an improved WTA network was presented in (Tsotsos, Sean, Wai, Lai, Davis, and Nuflo 1995), denoted selective tuning. The selective tuning is a more flexible architecture than the WTA process in (Koch and Ullman 1984), which is motivated from neurobiological findings. In (Tsotsos, Pomplum, Liu, Martinez-Trujillo, and Simine 2002) this model is developed further to approximate the neural processing of translation-, rotation-, and growing-motion of humans.

Koch continued the work on his attentional model, which resulted in a saliency-based search mechanism presented in (Itti, Koch, and Niebur 1998) and further developed in (Itti and Koch 2000). A reference implementation, named "iLab Neuromorphic Vision C++ Toolkit", is publically available at URL: "http://ilab.usc.edu/toolkit/". Although some criticism has been raised by (Draper and Lionelle 2003) the model is well-known. Moreover, the authors and (Ouerhani, von Wartburg, Hugli, and Muri 2004) have given some psychophysical motivation to this model.

Top-down information

In (Navalpakkam and Itti 2002) a top-down component is added to the saliency-based search mechanism in (Itti and Koch 2000). The top-down component, which encodes a visual search task, is integrated with the bottom-up saliency map to produce task relevant interest regions. By adding the top-down component the system can perform task oriented visual search. However, models for task oriented visual search has been developed much earlier.

(Milanese 1993) did build an attentional system that was inspired by the structure of human visual system during his PhD. A bottom-up process computes a set of feature maps which are integrated via an iterative energy minimizing algorithm to a saliency map. The saliency map guides a top-down process and a scan-path is generated by inhibition of return. The top-down search could also be interrupted by an alerting system by abrupt movements.

One of the more well-known attentional models is Guided Search 2.0, (Wolfe 1994), which is a development from the findings in (Wolfe, Cave, and Franzel 1989). A set of non-linear, categorical, feature maps are computed. A saliency map is computed by differences to adjacent items at each feature map. A top-down process searches for targets guided by the saliency map. In (Wolfe and Gancarz 1996) it was further developed to include the dynamic overt component of attention, which uses the higher acuity of vision at the central visual field.

(Rao, Zelinsky, Hayhoe, and Ballard 2002) developed a model for overt attention using top-down information without guidance from bottom-up saliency. The top-down information is expressed as a template of the filter response from a set of Gaussian derivatives, of order 0-3, on three scales. Saliency, with respect to the template, is used to guide the attention.

(Oliva, Torralba, Castelhano, and Henderson 2003) and (Torralba 2003) has developed a model which use prior knowledge of contextual information to guide attention. The contextual information is computed as mean and variance of the feature distribution in receptive fields at different scales. After a training phase this information can be used to guide the attention to expected contexts for different categories. For example in a city scene, cars are likely to be located on the street. After having learned the contextual properties for roads, this information can be used to guide the attention to roads when searching for cars.

Object based attention

There seems to be less work on object based attention. Object based attention normally involves perceptual grouping and segmentation. Most models studies only specific, biologically motivated, issues in perceptual grouping.

(Grossberg and Raizada 2000) has developed a neural network-model for perceptual grouping which addresses the phenomenon of illusionary contours, such as the Kanizsa square 2.17.

(Thielscher, Schuboe, and Neumann 2002) has developed a pre-attentive segmentation model for oriented textures. It builds on two key assumptions:

- Human texture processing is based on the detection of salient discontinuities signaling texture boundaries.
- Texture border detection is achieved by cells in higher visual areas having large receptive field to integrate the preprocessed input from lower areas.

Regions of similar texture are iteratively attenuated and its boundaries enhanced via feedback responses. As the texture area become more and more dense the attenuation will become stronger and discontinuous elements with in a uniform area will "pop-out". This has strong connections to findings from biological experiments.

(Sun and Fisher 2003) has implemented and analyzed a more general model for object based attention. The perceptual grouping is made by hand and used to calculate visual saliency of objects and hierarchically selecting attentional shifts. Large salient areas are attended at the most coarse scale. Within each area other salient subareas can be found hierarchically.

This model has many similarities to the proposed. However, in (Sun and Fisher 2003) segments are found by hand and in the automatic and stable segmentation of proposed model is one of its main contributions. The proposed model, on the other hand, has less focus on how to generate a scan-path.

Game theoretic perspective

To enable efficient distributed processing there is a need to find algorithms which do not integrate all data. Game theory has been studying concepts for distributed processing without central control.

The attentional models in section "Winner-take-all" and "Top-down information" integrates all information from all feature dimensions and selects one single winning node. In a game theoretic perspective this is similar to a strictly competitive game with full information (see section 3.2). The models in section "Object based attention" are not complete systems. (Grossberg and Raizada 2000) and (Thielscher, Schuboe, and Neumann 2002) have modeled two different aspects of object based attention, but does not generalize to natural scenes. In (Sun and Fisher 2003) segments are found by hand.

4.2 Contribution

A model has been developed which searches for target objects of an expected size. The model is designed to be implemented on a distributed system and uses concepts from extensive cooperative games with incomplete information to minimize the need for inter-process communication.

The strategy is to use knowledge of the environment; e.g. in a living-room we might expect large items with homogeneous surfaces such as a table and a cupboard. We denote them background regions. The background regions provide layout information of the scene, which can be used to guide the attention. Furthermore, the feature statistics of each background region is used as contextual information. The model avoids integration of all data at one node; instead the processing is distributed over a set of processing nodes. Only a sparse set of data is shared among the nodes without need for central control.

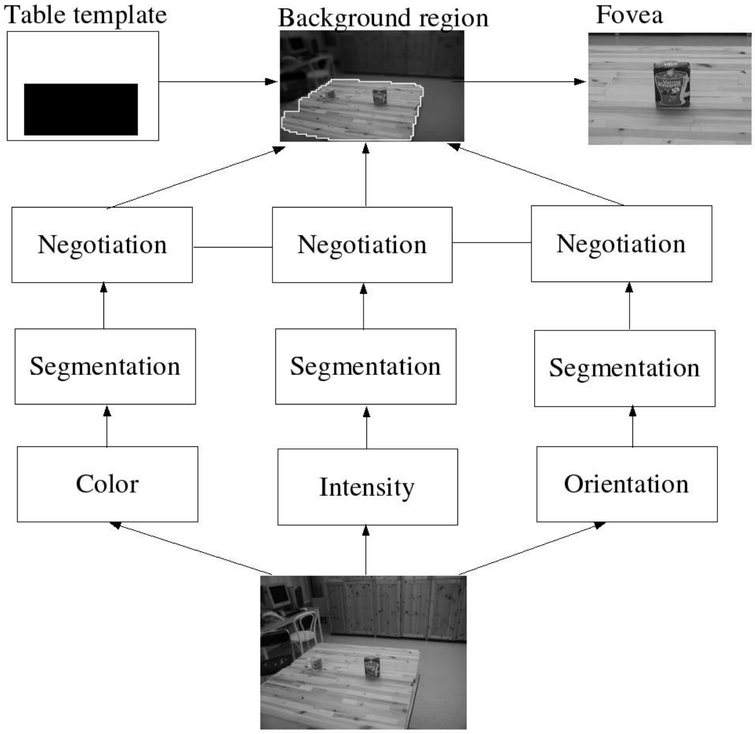

Figure 4.1: A raw image is processed by a set of distributed nodes resulting in a set of background regions, which often corresponds to large objects.

Figure 4.1 illustrates the system: A raw image from a camera is decomposed into a set of feature maps at separate nodes, namely color, intensity, and orientation (see chapter 5). A segmentation algorithm searches for large homogeneous regions locally at each node (see chapter 5). The resulting segments are sensitive to variations in the intrinsic parameters and the camera pose. A negotiation scheme forms coalitions of segments, which are more stable than the individual segments, by only sharing real valued coalition values across the nodes an the final winning segmentation mask (see chapter 6). The coalitions are denoted background regions and provide layout information; a spatial template for e.g. the table selects interesting background regions. Saliency can thereafter be computed locally at each node with respect to the feature statistics of each interesting background region (see chapter 7). As reference a center-surround saliency algorithm has been developed and is presented in 8. The performance of the model is inspected in chapter 9 and discussed in chapter 10.

The contribution of this thesis is usage of game theory for distributed control and the framework for generation of stable homogeneous regions, background regions. The contributions has be developed in several steps which has resulted in three publications:

Visual attention using game theory

In (Ramström and Christensen 2002) a model for computation of visual saliency, with respect to a template, is presented. The model is based on game theory. An input image is processed to a feature map and the pixels of the feature map are treated as actors on a market. The goal of the actors is to become similar to a template of a target object by trading on the market. The outcome of such market models saliency with respect to the target.

Object detection using background context

In (Ramström and Christensen 2004c) a system which computes large coherent regions using coalition formation is presented. The coherent regions are used to guide attention, using game theoretical concepts. Within each coherent region saliency, with respect to a template of a target object and the local background statistics, is computed.

In (Ramström and Christensen 2004b) the computation of large coherent regions in (Ramström and Christensen 2004c) is improved and used to compute saliency with respect to object-size. This enables the system to search for objects specified by size, rather than features.

Finally, in (Ramström and Christensen 2004a) the system was improved to the form that is presented in this thesis.

Chapter 5

Image processing

5.1 Introduction

The system processes an observed raw image into a set of feature maps, which are subsequently segmented into a set of homogeneous background regions. In this chapter we will discuss some of the image processing techniques that are applied.

Section 5.2 describes the decomposition of a color image into a set of feature maps and section 5.3 and 5.4 describes two different segmentation algorithms.

To evaluate what information is gained by extracting homogeneous regions the object detection performance will be compared to a center-surround saliency method based on (Itti and Koch 2000). Center-surround saliency is computed as feature difference between a center region and its surrounding on several scales. Such computations can efficiently be processed by computing a scale pyramid, where each consecutive layer is a down sampled copy of the layer below. Hence, a pixel in the top layer correspond to a region in the layer below, and a larger region in the layer below that, etc. The drawback of a scale pyramid is that we need to compute and store each layer. Instead of calculating a scale pyramid integral feature maps can be used, these are computed faster and consume less memory. Section 5.5 describes integral feature maps.

5.2 Low-level vision

The observed raw image has 300 × 224 pixels in resolution using the YUV color space and the decomposed feature maps have 75 × 56 pixels in resolution with different dimensionality. The format of the raw images enable accurate processing of image features and the four times down sampling of the feature maps reduce noise and thus improves extraction of homogeneous regions.

From experimental psychology (Enns and Rensink 1992)and biology (Marcus and van Essen 2002) it is clear that the visual cortex performs segmentation preattentively using several separate visual features, Julez denoted such features textons

(Julez 1981). Treisman (Treisman and Gelade 1980) found that segments cannot be formed by conjunctions of separate features. We will in this model restrict us to three separate feature dimensions: Color, intensity, and orientation. These are suitable to the environment we will use for evaluation. Note that it is not claimed that these are better than any other feature dimensions nor that three separate dimensions is an optimal number of dimensions. However, these feature dimensions allow us to compare the results to (Itti and Koch 2000), where similar although not identical features are used.

Feature map identity is denoted $d \in \{color, intensity, orientation\}$ and the feature maps are denoted f^d. The three different feature maps are described in section 5.2, 5.2, and 5.2. In order to make output from the different feature maps comparable when searching for homogeneous regions all feature maps are normalized to have zero mean and unit variance.

Intensity

Intensity is represented as a one dimensional vector $f^{intensity} \in \mathbf{R^1}$; the Y channel normalized to have zero mean and unit variance.

Figure 5.1 illustrates the normalized intensity features resulting from left most raw image.

Color

Color is represented as a two dimensional vector $f^{color} \in \mathbf{R^2}$, resembling opponent color.

The color channels $U, V \in [-127, 128]$ are independently normalized to have zero mean and unit variance.

Hence, $f^{color}(x,y) = [U_n(x,y), V_n(x,y)]$; where the n subscript indicates normalized values.

Figure 5.2, middle and right image, illustrates the two color features resulting from left most raw image.

Orientation

Orientation is represented by normalized second order derivatives of Gaussian at arbitrary scale and sign. The feature vectors have dimensionality three, $f^{orientation} \in \mathbf{R^3}$.

At each position (x,y) we compute the response from 24 second order derivatives of Gaussian, of different orientations o and scale s, $DG2^o_s$; where $s \in \{1,2,3,4\}$ denotes the scale (or variance of the Gaussian) and $o \in \{0^o, 30^o, 60^o, 90^o, 120^o, 150^o\}$ the orientation.

The responses are represented by a set of vectors:

$$DG2_s(x,y) = \{(DG2^0_s, DG2^{30}_s, DG2^{60}_s, DG2^{90}_s, DG2^{120}_s, DG2^{150}_s)(x,y)\}; s \in \{1,2,3,4\} \tag{5.1}$$

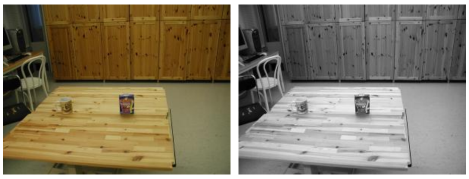

Figure 5.1: Raw image and feature map at the intensity node

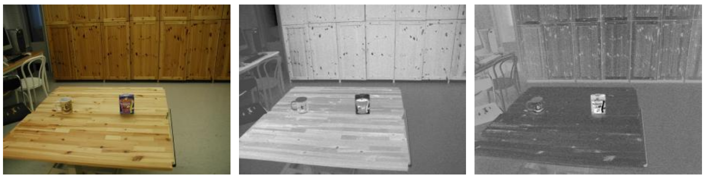

Figure 5.2: Raw image, U-, and V-feature map at the color node

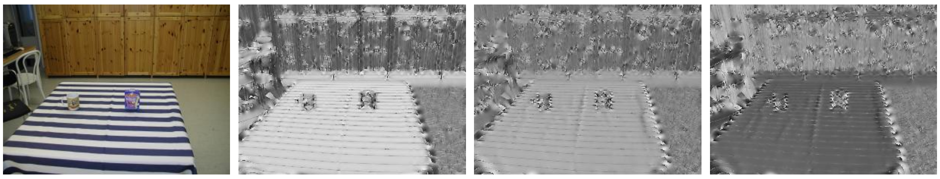

Figure 5.3: Raw image and feature map (0,30,60) at the orientation node

The dimensionality is reduced and the response is normalized by a two stage process which is inspired and motivated by (Varma and Zisserman 2002). In the first stage the dimensionality is reduced by selecting the strongest responses over scales:

$$DG2_{max} = \{DG2_{s'}(x,y)|||DG2_{s'}(x,y)\| \geq \|DG2_s(x,y)\|; \forall s', s \in \{1,2,3,4\} \quad (5.2)$$

where $||x||$ is the $L2$ norm of x. In the second stage DG_{max} is normalized to

$$DG2_{norm} = DG2_{max}\frac{log(1+10*\|DG2_{max}\|)}{||DG2_{max}||} \quad (5.3)$$

In (Varma and Zisserman 2002) the representation is further optimized to enable texture segregation, here we desire a representation which enable texture segmentation. An area with homogeneous texture with respect to orientation of ridges, e.g. a striped area, will normally have dark and light parallel stripes interleaved. With the he $DG2_{norm}$ representation a light and dark parallel stripe will have identical feature vectors except for the sign. This alternation of sign is not suitable for segmentation based on local similarities. We resolve this issue by multiplying the feature vectors with the dominant sign:

$$DG2_{sign} = \sum_{o} \frac{DG2^{o}_{norm}}{|DG2^{o}_{norm}|} DG2_{norm} \tag{5.4}$$

where $|x|$ is the $L1$-norm of x.

Finally, we reduce the dimensionality again by the definition:

$$f^{orientation} = (DG2^{0}_{sign} - DG2^{90}_{sign}, DG2^{30}_{sign} - DG2^{120}_{sign}, DG2^{60}_{sign} - DG2^{150}_{sign}) * q \tag{5.5}$$

where $q \in \mathbf{R^1}$ is a constant that normalize $f^{orientation}$ to unit variance.

The intuition behind this process can be illustrated in figure 5.3: Although the size of stripes of the tablecloth vary due to perspective and alternate between dark and light colors, the final orientation response is sufficiently smooth to be perceived as homogeneous by the presented model in the set of evaluation scenes 9. The $f^{orientation}$ operator has a peak at ridges and minimum at the border between two ridges. If this variation would interfere a more complex process should be applied, such as (Lindeberg 1998).

5.3 Mean-shift segmentation

Many segmentation algorithms can be considered to segment an image into homogeneous regions. Refer to (Forsyth and Ponce 2002) (chapter "Segmentation using clustering methods") for a good overview. The mean-shift segmentation (Comaniciu and Meer 1999) is a fairly well established segmentation algorithm based on local similarities.

The mean-shift segmentation algorithm has two intrinsic parameters for the clustering process: spatial scale s and feature range r. The (r, s) parameters are used in two sequential processing steps, filtering and segmentation, which are described below.

A filtered feature map z is computed iteratively from the input feature map f by the following algorithm (the d-superscript is omitted in this section for ease of notation, f refers to an arbitrary node):

At iteration $i = 0$ and at each pixel (x, y), we initialize a mean feature vector $m^f_0 = f(x, y)$ and a mean position vector $m^p_0 = (x, y)$. At each successive iteration i the vectors m^f_i and m^p_i are updated by the following mean-shift procedure:

- Select all pixels $w \in W$ for which the Euclidean distances $||f(w) - m^f_i|| < r$ and $||w - m^p_i|| < s$.

- Shift m^p_{i+1} to the mean position of W: $m^p_{i+1} = \frac{\sum_{w \in W} w}{|W|}$; where $|W|$ is the size of set W.

After convergence, $||m^p_c - m^p_{c-1}|| < \epsilon$ for a small value ϵ, we assign $z(m^p_0) = m^f_c$.

The mean-shift filtering algorithm is proven to converge at modes of the feature density function of input feature map (see details in (Comaniciu and Meer 1999)).

In the segmentation step we connect all filtered pixels for which $||z(x_1, y_1) - z(x_2, y_2)|| < r$ and $||(x_1, y_1) - (x_2, y_2)|| < s$ and give each such segment a unique label $\lambda = 0, 1, ...$ Finally, we remove all segments smaller than 400 pixels, which is about 10% of the feature map area and select the N largest remaining segments at each node. Hence, we extract only the large homogeneous regions at each node.

Figure 5.4 to 5.9 illustrates the resulting segments at the color node from the left most image. In more detail 5.4 is the segmentation result using $(r, s) = (0.15, 3)$ and 5.5 using $(0.22, 4)$. We observe that $(0.22, 4)$ in the same scene result in segments which represent the table and cupboard well. If we use the same parameters for two other views in figure 5.6 and 5.7 we observe that the table is well segmented but no segment corresponding to the cupboard is obtained in figure 5.6. If we change the tablecloth we observe that we do not get any segments in figure 5.8 and one noisy in figure 5.9. We can conclude that it is not straightforward to select (r, s) and feature dimension which are stable to small variations of the scene.

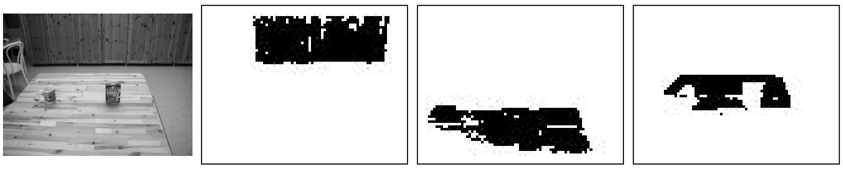

Figure 5.4: The left most image result in three mean-shift segments with r=0.15, s=3

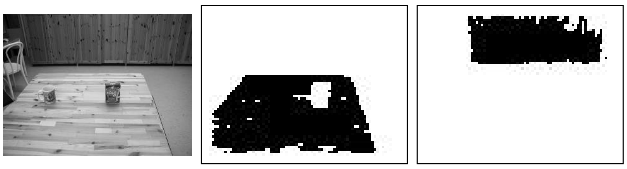

Figure 5.5: The left most image result in two mean-shift segments with r=0.22, s=4

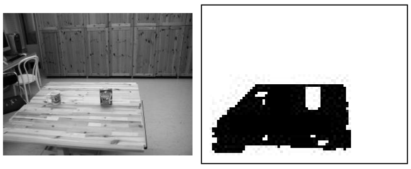

Figure 5.6: The left most image result in one mean-shift segments with r=0.22, s=4

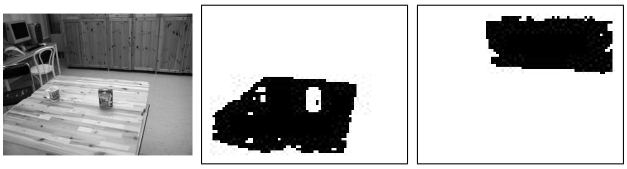

Figure 5.7: The left most image result in two mean-shift segments with r=0.22, s=4

Figure 5.8: The left most image result in no mean-shift segments with r=0.22, s=4

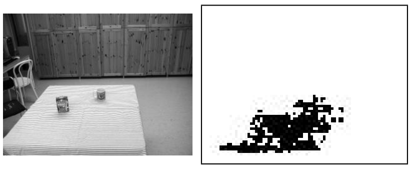

Figure 5.9: The left most image result in one mean-shift segments with r=0.22, s=4

5.4 Gaussian Mixture Models

The mean-shift algorithm is based on local similarities; Gaussian Mixture Models is based on global feature properties. The features-values are considered to belong to one of a fixed set of sources with Gaussian variation. However, the number of sources and their statistics are not known. More importantly we do not know if the image can be describe by this model! Nevertheless, Gaussian Mixture Models can often represent the features of an image well. Refer to (Forsyth and Ponce 2002) (chapter "Segmentation and fitting using probabilistic methods") for a more detailed discussion.

A Gaussian Mixture Model (GMM) of each feature map can be computed by the Expectation-Maximation algorithm:

A random set of Gaussian kernels $\{(m,v)_j\}_{j=0}^{k}$ are selected, where m is mean vector and v is covariance matrix. In the expectation step the expectation for each pixel i to belong to the Gaussian kernel $(m,v)_j$ is computed as

$$E_{ij} = \frac{p(x = x_i|(m,v) = (m,v)_j)}{\sum_{n=1}^{k} p(x = x_i|(m,v) = (m,v)_n)} \tag{5.6}$$

In the Maximization step the estimates of $\{(m, v)_j\}_{j=0}^{k}$ is improved using E_{ij}:

$$m_j = \frac{1}{n} \sum_{i=0}^{n} E_{ij} x_i \tag{5.7}$$

$$v_j = \frac{1}{n-1} \sum_{i=0}^{n} E_{ij} (x_i - m_i)(x_i - m_i)^T \tag{5.8}$$

The expectation and maximization step is processed iteratively. Moreover, at each iteration we identify Gaussian models that are too similar and replace one model by a new random kernel. In more detail if two Gaussian models m_i and m_j, $i < j$, have normalized histogram intersection:

$$2 \frac{m_i \cap m_j}{|m_i| + |m_j|} > 0.9 \tag{5.9}$$

then m_j is replaced by a new random Gaussian kernel.

Figure 5.10 illustrates the resulting mixture of three Gaussian applied to figure 5.4

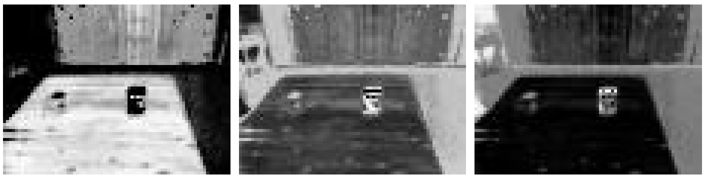

Figure 5.10: Gaussian Mixture Models, left to right image correspond to probability to fit first, second, and third Gaussian model

5.5 Integral feature map

Integral feature maps enables efficient computing of integrals over arbitrary rectangular areas; the sum over an arbitrary rectangular area is computed from the values at the four corners of the rectangle in the integral image. Instead of computing a scale pyramid where an integral value over a region is accessed by selecting appropriate layer and position, a rectangle of appropriate size and position is applied to the integral image. The drawback is that we are restricted to rectangular areas.

Each pixel of an integral feature map has the value of the sum inside a square from origo to the pixel. Hence, the integral pixel

$$I^d(i, j) = \sum_{x=0}^{i} \sum_{y=0}^{j} f^d(x, y) \tag{5.10}$$

This representation enables us to calculate the area within an arbitrary rectangle $(x1 < x < x2, y1 < y < y2)$ using only the feature vectors from the four corners of the rectangle $(x1, y1)$, $(x2, y1)$, $(x1, y2)$, $(x2, y2)$:

$$\sum_{x=x1}^{x2} \sum_{y=y1}^{y2} f^d(x, y) = I^d(x1, y1) + I^d(x2, y2) - I^d(x1, y2) - I^d(x2, y1) \tag{5.11}$$

5.6 Summary

We have presented some examples of image processing techniques which are used in this thesis to process an observed raw image into a set of feature maps and subsequently to segment a feature map into homogeneous regions.

Three feature maps are extracted: $f^{intensity}$, f^{color}, and $f^{orientation}$. Homogeneous regions are extracted in each feature map by two different segmentation algorithms: Mean-shift segmentation and Gaussian Mixture Models.

Moreover, integral feature maps have been described. Using integral feature maps, an integral over an arbitrary rectangular area can be computed by a linear combination of only four data.

Coming chapters will use these techniques extract homogeneous regions and to search for targets using distributed control.

Chapter 6

Background regions

6.1 Introduction

By distributing the processing to a set of nodes the computational power is increased. However, if the nodes have to share much data the communication between them becomes a bottleneck. In order to decrease the communication demand it is of interest to use attentional methods. Inspired from experimental psychology we perform preattentive segmentation across the field of view. Homogeneous regions are extracted and local context information at such regions is used to guide the attention. The local context is a segmentation mask and the feature statistics, which is approximated by a Gaussian model. If a homogeneous region is much larger than an expected target it can be used to guide attention in target detection, such regions are denoted background regions. In chapter 7 we will see how background regions can be used to decrease the communication load.

In chapter 9 we will see that when the task is to find objects on a table, it is of interest to find background regions, which might represent large homogeneous objects in the scene, in particular the table. If the distance to the table is known, the mean-shift segmentation algorithm can be applied with predefined intrinsic parameters to extract a segment corresponding to the table. However, if the distance is known with some uncertainty, we need a method for extracting regions which are stable with respect small variations in camera pose. This can be achieved by clustering redundant mean-shift segments, corresponding to different intrinsic parameters.

We will in this chapter describe a method to extract background regions and their segmentation masks. First, coalitions of similar mean-shift segments are formed, section 6.2. The coalition formation is a complex task with respect to the number of available segments. This complexity is reduced by a negotiation scheme, section 6.3. The resulting clusters are denoted background regions and are used to find a segmentation masks of homogeneous regions, section 6.4. The segmentation masks are post-processed to form regions which are more object-like, section 6.5. The process in summarized in section 6.6.

6.2 Clustering of redundant mean-shift segments

The mean-shift segmentation algorithm depend on two intrinsic parameters (r, s). Different values of these parameters might result in different segmentation result. Since they relate to distances in the spatial and feature domain, such variations are related to variations in pose and illumination. To increase the stability we compute the mean-shift segmentation algorithm varying $r \in M_r$ and $s \in M_s$. We will in this work restrict $M_r = \{2, 3, 4\}$ and $M_s = \{0.15, 0.22, 0.33\}$. Furthermore, we remember from section 5.3 that for each selection of $(r, s) \in M_r \times M_s$ we select only the $N = 4$ larges segments and discard all segments smaller than 10% of the feature map area.

The mean-shift segmentation algorithm is processed locally at each node for each $(r, s) \in M_r \times M_s$. Let P^d represent the resulting set of mean-shift segments at node d; hence the size of $|P^d| \leq N|M_s \times M_r|$. Each mean-shift segment $p_i \in P^d$ is associated with a segmentation mask S_i^d and a mean feature value inside the segmentation mask m_i^d. The similarity between two mean-shift segments p_i and p_j is defined as the normalized intersection of their segmentation masks and mean features:

$$Sim(i, j, d) = \frac{S_i^d \cap S_j^d}{|S_i^d| + |S_j^d|} \cdot \frac{m_i^d \cap m_j^d}{|m_i^d| + |m_j^d|} \tag{6.1}$$

The second factor is fairly standard in histogram matching, the first factor borrows the same normalization technique and gives a penalty when they differ spatially in size, location, and shape.

To find the optimal background regions we need to evaluate all possible cluster combinations of mean-shift segments, which has exponential complexity $O(2^{|P^d|})$. This complexity is reduced using a modified version of the coalition formation process proposed by (Shehory and Kraus 1998), which only have square complexity with respect to the number of mean-shift segments $O(|P^d|^2)$.

6.3 Negotiation

Coalitions of mean-shift segments are formed by an iterative negotiation process. At iteration $t = 0$ each mean-shift segment $p_i \in P^d$ broadcast its description, (S_i^d, m_i^d), and selects a set of possible coalition members $C_i^d(0) \subseteq P^d$, including all other mean-shift segments with a similarity larger than th:

$$C_i^d(0) = \{\forall p_j \in P^d | Sim(i, j, d) > th\} \tag{6.2}$$

We define the value of a coalition $C_i^d(t)$ at iteration t as:

$$V_i^d(t) = \sum_{j \in C_i^d(t)} Sim(i, j, d) \tag{6.3}$$

Mean-shift segments are valued relative to their contribution, hence the value of p_j in coalition $C_i^d(t)$ is:

$$V_i^d(j,t) = Sim(i,j,d) \tag{6.4}$$

Thus, each segment forms a coalition including similar segments at the same node. From this set of coalitions, background regions are iteratively extracted by a distributed negotiation scheme. In each iteration of the negotiation the strongest coalition across all nodes is chosen and used to form a background region and to inhibit segments in succeeding negotiation iterations.

In more detail, stable coalitions are formed when each mean-shift segment p_i at each node d iteratively perform the following:

1. Compute and announce $V_i^d(t) = \sum_{j \in C_i^d(t)} V_i^d(j,t)$ to all other segments at all nodes.

2. Choose the highest among all announced coalition values, $V_{max}(t)$.

3. If no other coalition at any node has a stronger coalition value, $V_i^d(t) = V_{max}(t)$, then compute the weighted segmentation mask $W^d = q \sum_{j \in C_i^d} V_i^d(j,t) S_j^d$; where $q \in \mathbf{R^1}$ is a constant which normalize W^d to have maximal value one. Remove all $p_i \in C_i^d$ from further negotiation.

4. Update $V_i^d(j,t+1) = (1 - 2\frac{S_i^d \cap S_{max}^d}{|S_i^d| + |S_{max}^d|})V_i^d(j,t)$;

5. Start over from 1.

At each iteration a weighted segmentation mask, W^d, is computed and $|C_i^d|$ is decreased for at least one mean-shift segment. The process will be repeated until all C_i^d are empty or for a fixed number of iterations.

Note that the set of nodes only compares values of maximal coalitions, all other computation of image data is processed locally at each node. The weighted segmentation masks resulting from five different scenes are illustrated in figure 6.1.

6.4 Segmentation mask

A segmentation mask can be extracted by thresholding each weighted segmentation mask W^d. However, we do not have any analytic way to extract such threshold value. Instead we calculate a Gaussian-mixture model (GMM) of the complement region of W^d in the associated feature map f^d, using 5 Gaussian models.

The resulting probability maps are suitable competition to W^d. We compute the E_{ij} maps, which is the probability of pixel i to belong to the Gaussian model j, for each $j = 1,2,3,4,5$. Furthermore, we denote the Gaussian model at W^d with $j = 0$, hence E_{i0} is the probability map for pixel i to belong to the Gaussian model at W^d. Finally, the probability for pixel i to belong to background region C_{max}^d is the joint probability $W_i^d E_{i0}$.

The segmentation mask, S^d, associated with C^d_{max} is defined as the pixels i where $W^d_i E_{i0} > E_{ij}$ for all $j = 1, 2, 3, 4, 5$.

We observe in figure 6.2 that in this example the segmentation masks S^d are far more stable than the corresponding mean-shift segments shown in figure 5.4 to 5.9.

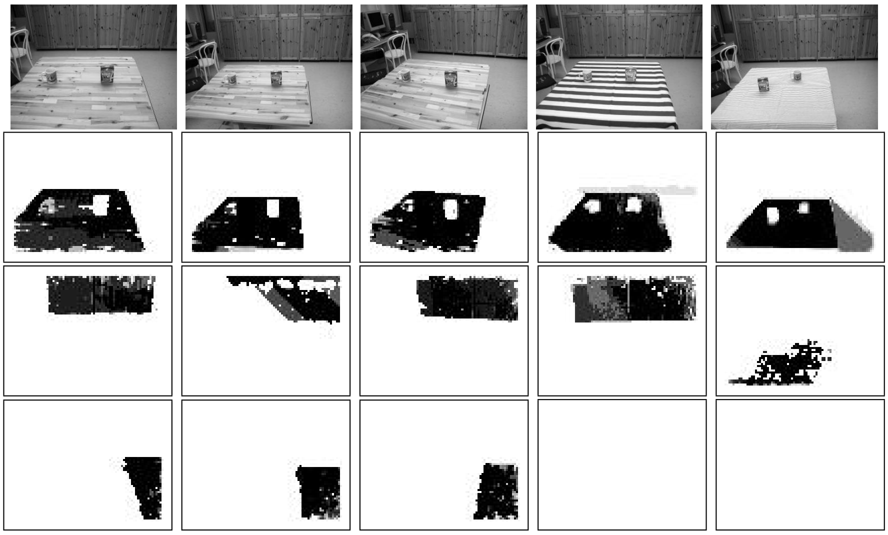

Figure 6.1: Top row: Raw images with variations in camera pose and tablecloth. Row two to four: corresponding weighted segmentation mask for first, second, and third background region.

6.5 Region completion

(Enns and Rensink 1992) demonstrated in a series of experiments that the visual cortex preattentively completes homogeneous regions to object hypothesis. Inspired by this result we perform region completion of the segmentation masks S^d. The object completion is not as advanced as the completion process demonstrated by (Enns and Rensink 1992), however it enables detection of objects within homogeneous regions. One obvious completion process is to fill in holes. Furthermore, objects e.g. at the border of a table often pop-out from the table leaving a notch on the border of the segmentation mask. Regions corresponding to artificial large indoor-objects, in the set of evaluation scenes, are often square or have some vertical or horizontal straight lines. Following this discussion, we define the region completion as filling in gaps where a vertical or horizontal straight line can be attached to the original segment. Note that a notch in the corner will not be completed by this process as predicted by (Enns and Rensink 1992) .

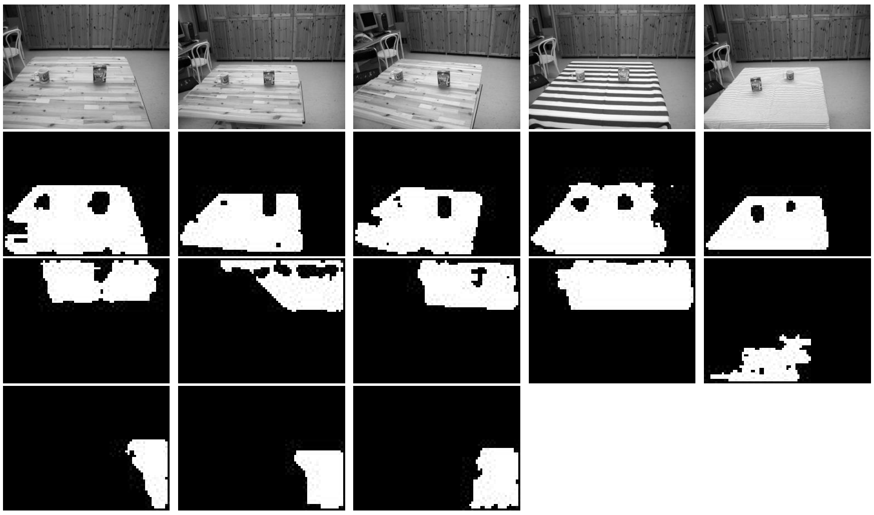

Figure 6.2: Top row: Raw images with variations in camera pose and tablecloth. Row two to four: corresponding segmentation mask for first, second, and third background region.

Moreover, we do not want to detect other overlapping background regions as salient. Therefore we restrict object completion to regions not occupied by other original segmentation masks.

Figure 6.3 illustrates completed segmentation masks from the same five different scenes. We observe that most regions are compound. Regions, which partially overlap with other regions, are exceptions.

This process is solely based on intuition from the (Enns and Rensink 1992) experiments and can obviously be improved. However, the completion process enables accurate objects detection and is sufficient at this point in the development process.

6.6 Summary

We have in this chapter presented a negotiation scheme for coalition formation of similar mean-shift segments. Such coalitions are used to find a segmentation mask of background regions.

By using redundant segmentation results, with respect to spatial and feature scale, robustness is improved. The complexity of the clustering process is reduced by a negotiation scheme. Each cluster will have a weighted segmentation mask. A Gaussian mixture model process is subsequently applied to extract a binary segmentation map from the each weighted segmentation mask. Finally, an object com-

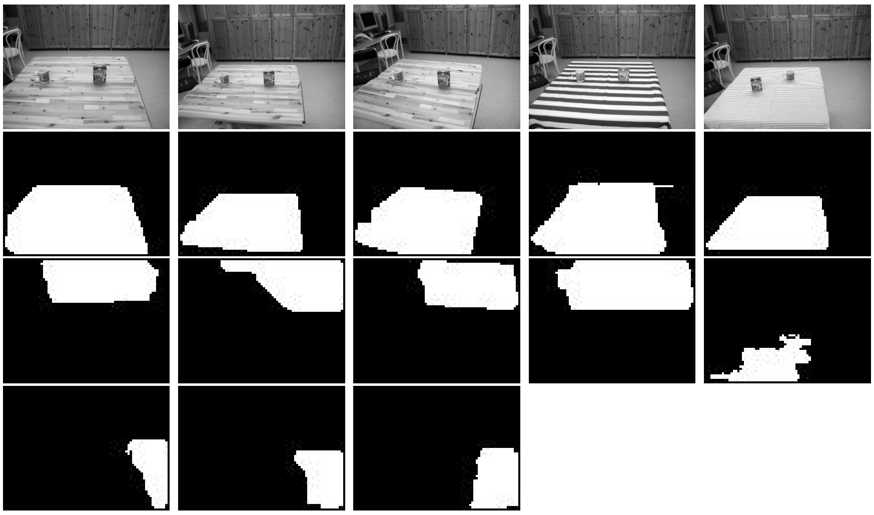

Figure 6.3: Top row: Raw images with variations in camera pose and tablecloth. Row two to four: corresponding completed segmentation mask for first, second, and third background region.

pletion process makes the regions more compound, which is useful when searching for targets within the region.

The negotiation scheme is fully distributed, i.e. no central control node is required. Only real valued coalition-values are integrated during the negotiation phase and a coarse segmentation mask is distributed at the end of each negotiation.

Chapter 7

Target objects

7.1 Introduction

At a background region we have a statistical model of the feature distribution and a segmentation mask. Both the feature and spatial information can be used to guide the search for targets.

First, all background regions that are spatially similar to a template are attended, section 7.2. The feature statistics of each such background regions is extracted and used to search for outliers which are likely to be a target object, section 7.3. Using the statistics of the background object and the target hypothesis a figure-ground segmentation can be processed, section 7.4. The processing steps are summarized in section 7.5.

7.2 Background region of interest

The task is to find objects on a table which are expected to occupy a large fraction of the lower half of the scene. We use a template to represent this knowledge. A background region which overlaps more than 25% with the template it is considered a background region of interest. Figure 7.1 illustrates the template used here.

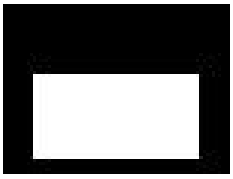

Figure 7.1: Template for background region

7.3 Foreground region of interest

To find foreground regions of interest (ROI) we calculate the feature statistics of each background region of interest. The weighted summation mask W^d of each such background region is used to calculate (m^d, Σ^d) at each node $d \in \{color, intensity, orientation\}$:

$$m^d = \frac{1}{|W^d|} \sum_{x=0}^{X} \sum_{y=0}^{Y} f^d(x,y) W^d \tag{7.1}$$

$$\Sigma^d = \frac{1}{|W^d|} \sum_{x=0}^{X} \sum_{y=0}^{Y} (f^d(x,y) - m^d)(f^d(x,y) - m^d)^T W^d \tag{7.2}$$

where W^d is the sum of all pixels in W^d.

We calculate the set of pixels p_S^d which can be excluded from background region S^d with confidence γ with respect to a Normal distribution (m^d, Σ^d). If this set is larger that q, the confidence value is increased and a new set p_S^d is computed. This process is repeated until the size $|p_S^d| < q$.

In the current implementation we have chosen the set $\gamma \in 0.5, 0.6, 0.7, 0.8, 0.9, 1$ and $q = 0.25|S|$ without further investigation.

The sparse set of conspicuous pixels p_S^d is distributed to all other nodes. At each node an integrated saliency map is constructed from the sum of all p_S^d. The integrated saliency map in convolved with a Gaussian kernel with a standard deviation equal to the expected target size. Each peak of the saliency map, which is larger than one, is extracted as ROI. Hence, we only consider regions, of expected target size, that at least one node has found conspicuous.

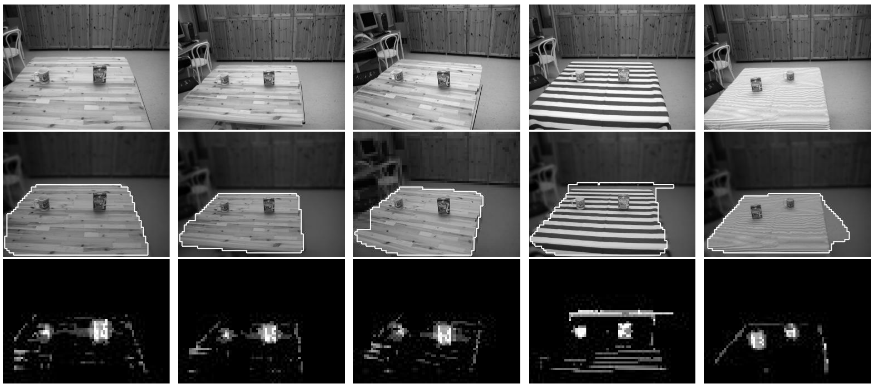

Figure 7.2: Top row: Raw images with variations in camera pose and tablecloth. Row two: background regions of interest. Row three: corresponding saliency maps

7.4 Target segmentation

Each peak, which is selected in the saliency map, is attended with a foveated camera with sixteen times higher resolution. A region of 19×14 pixels in the saliency map is attended with the foveated camera and will be observed with 300×224 pixels in resolution. The corresponding 19×14 region of the saliency map is cropped and up-sampled to 300×224 pixels (see figure 7.3). The up-sampled saliency image is denoted σ and is used to compute the foreground feature statistics.

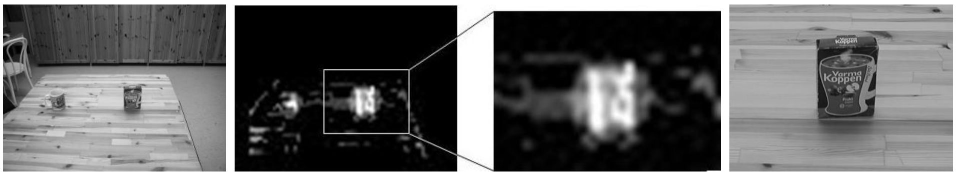

Figure 7.3: Seed for foveated object segmentation

The complement of σ, within the background region, is denoted β and is used to calculate the feature statistics of the background.

The foveated feature maps are denoted f_f^d; where $d \in \{color, intensity\}$. Note that *orientation* is omitted, the reason for this is that in the foveated image stripes which can be detected by the peripheral vision are too large to give accurate statistics by the foveated vision. This result is interestingly to what (Yeshurun and Carrasco 1998) found; foveated vision can impair texture segmentation.

The foreground and background statistics are computed according to equation 7.3 to 7.6:

$$m_{fg}^d = \sum_x \sum_y \sigma(x,y) f_f^d(x,y) \tag{7.3}$$

$$\Sigma_{fg}^d = \sum_x \sum_y \sigma(x,y)(f_f^d(x,y) - m_{fg}^d)(f_f^d(x,y) - m_{fg}^d)^T \tag{7.4}$$

$$m_{bg}^d = \sum_x \sum_y \beta(x,y) f_f^d(x,y) \tag{7.5}$$

$$\Sigma_{bg}^d = \sum_x \sum_y \beta(x,y)(f_f^d(x,y) - m_{bg}^d)(f_f^d(x,y) - m_{bg}^d)^T \tag{7.6}$$

The foreground segmentation mask is computed by:

$$S_f(x,y) = 1 \text{ iff } (f_f^d - m_{fg})\Sigma_{fg}(f_f^d - m_{fg})^T > (f_f^d - m_{bg})\Sigma_{bg}(f_f^d - m_{bg})^T \tag{7.7}$$

Figure 7.4 illustrates interest points and three first foveated foreground segmentations at the background regions of interest for five different scenes.

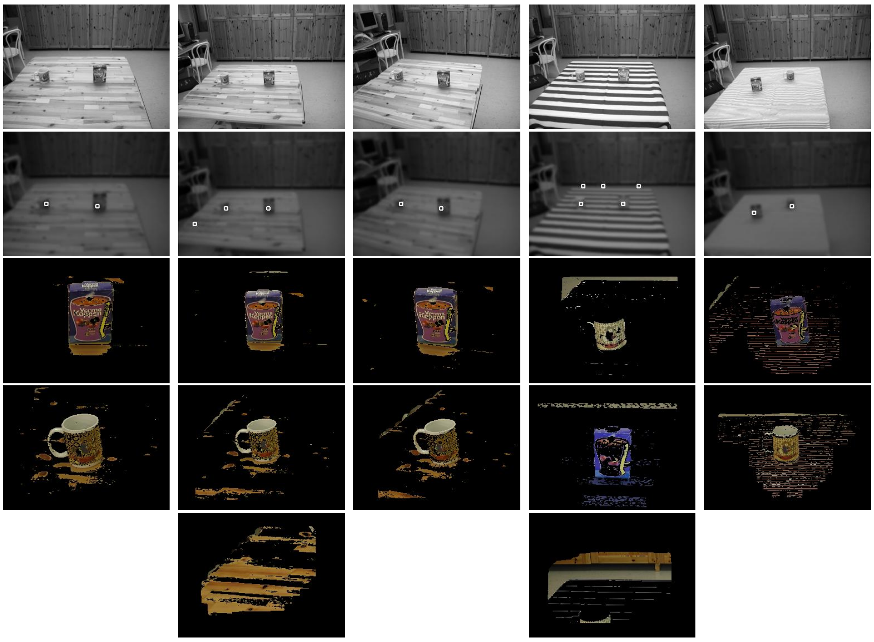

Figure 7.4: Top row: Raw images with variations in camera pose and tablecloth. Row two: Interest points at background regions of interest. Row three to five: Foveated foreground segmentation of first three interest points.

7.5 Summary

Targets are searched by attending to a background regions which are likely to contain the target. Such background regions of interest are extracted by comparing their spatial layout to a template, as illustrated in figure 7.1.

Each node compute the feature statistics of the background regions of interest and extract a sparse set of pixels which are not likely to belong to the statistics. Such conspicuous pixels are likely to occupy an occluding object and are distributed to the other nodes. From the set of integrated conspicuous pixels a saliency map is created and peaks in the saliency map are extracted as foreground regions of interest, ROI.

Each ROI is attended with a foviated camera. Using knowledge of the background and foreground region of interest in the peripheral view, a figure-ground segmentation can be computed in the foviated view.

By restricting the attention to a background region and its feature statistics, only a sparse set of data need to be integrated across the nodes to detect target hypotheses and perform figure-ground segmentation.

Chapter 8

Center-surround saliency

8.1 introduction

A well-established attention model is the center-surround saliency model developed by (Itti and Koch 2000) and was previously mentioned in section 4.1. The source code is available at `http://ilab.usc.edu/toolkit/`. However, to suite our choices of feature map definitions, an own implementation has been developed based on (Itti and Koch 2000). Our implementation uses fewer scales and integral feature maps (described in chapter 5), and is hence not equally good as the original. However, it has similar properties as the original and is used here as a comparison of typical behavior. It will be denoted CS-search.

It should be pointed out that CS-search does not have the same focus on distributed processing as the proposed model; at the final step complete pixel maps are integrated.

The CS-search model is presented in section 8.2 and summarized in section 8.3.

8.2 CS-search

Using the integral images we define center-surround saliency as the Euclidean distance between the mean feature vector inside a center rectangle (rc) and the mean feature vector inside a surrounding larger rectangle (rs). Let (rc, rs) denote a center rectangle with width rc and a surrounding rectangle with width rs, as illustrated in figure 8.1.

Six different center-surround saliency maps $CS^d_{(rc,rs)}$ are computed at each node with:

$$(rc, rs) \in \{(20, 50); (20, 60); (30, 60); (30, 70); (40, 70); (40, 80)\}$$

and

$$d \in \{color, intensity, orientation\}$$

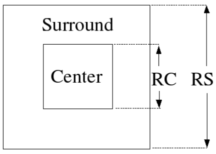

Figure 8.1: Center-surround rectangles (rc,rs)

The 18 center-surround saliency maps (six maps at three nodes) are integrated using a weighted summation

$$CS_{tot} = \sum_{\forall(rc,rs)} \sum_{\forall d} w^{d}_{(rc,rs)} CS^{d}_{(rc,rs)}$$

The weights, $w^{d}_{(rc,rs)}$, are defined as the square of the difference between the maximal peak value and the mean peak value at each center-surround saliency map. Peaks are defined as local maxima with respect to a four-connected neighborhood. The weights act as lateral normalization: Center-surround maps with many equally strong peaks are attenuated whereas maps a few salient peaks are amplified. See (Itti and Koch 2000) for more details.

If we search for cups, or simply want to inspect the scene for objects of target size, we do not want to stop after the first cup or item is found. Rather, we want to generate a scan path for attentive scrutinizing. We simulate interest points for such a scan path by the ten strongest CS-search peaks within 50% of the strongest peak.

In figure 8.2, right most image, we see that the package is detected as a salient point but not the cup, given 10 points within 50% of the strongest. It is not surprising that the clutter to the left in the image attract more attention than the cup on the table.

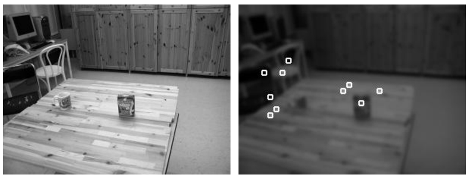

Figure 8.2: Raw image, center-surround saliency map, and saliency peaks (50%)

We address these drawbacks of CS-search with our proposed model, as illustrated in figure 8.3: From a raw image (left image) we extract a background region (middle image) which is likely to correspond to the table and search for target hypotheses using its context (right image).

We observe in figure 8.3 that using the context of the tabletop, only the package and the cup are considered salient.

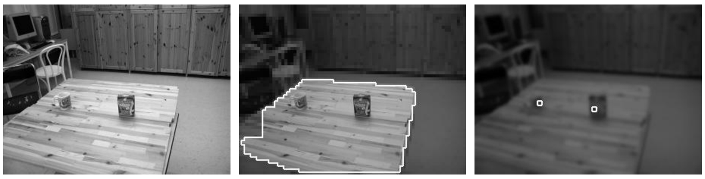

Figure 8.3: Raw image, the CS-search peaks and salient sub-regions of background regions

8.3 Summary

A saliency model based on center-surround saliency has been discussed. Center-surround saliency is defined as normalized differences between a center area and a surrounding area, both of varying sizes.

It does not have the same focus on distributed processing and does not attempt to extract any background context. However, since it is a well-known attention model it serves well as a comparison to the proposed model in the next chapter.

Chapter 9

Result

9.1 Introduction

There is no single correct ground truth in segmentation and attention; both vary with scale and task, among many other factors. To avoid such issues the system has been evaluated in terms its ability to detect targets on a table. The detection performance will depend on its ability to extract background regions.

For the evaluation five different scene set-ups has been used and in each scene the camera pose and the targets on the table has been varied in eight different configurations. Figure 9.1 lists the set of 40 resulting evaluation raw images. The scenes were chosen so the proposed model performs well but not perfect. Quite naturally, the proposed model does not perform well in cluttered scenes where no homogeneous regions are found. The selection of raw images provides a sense of what the proposed algorithm considers a non-cluttered scene.

We observe that in seven of the eight configurations of each scene there are two objects on the table, in one configuration the table is empty. Thus, there is a total of $2 \times 7 \times 5 = 70$ objects to be detected on 40 different appearances of the table.

In section 9.2 two experiments are presented which illustrate the power of foveated vision and the need for an accurate peripheral attention mechanism. In section 9.3 the proposed model is evaluated as peripheral attention mechanism.

9.2 Recognition

SIFT recognition

A recognition experiment has been done in cooperation with Fredrik Fuesjö (Ramström and Fuesjö 2004). SIFT features described in (Lowe 1999) was used for recognition on both the peripheral and foveated view. The peripheral view correspond to the set of raw images in figure 9.1 and foveated view correspond to the

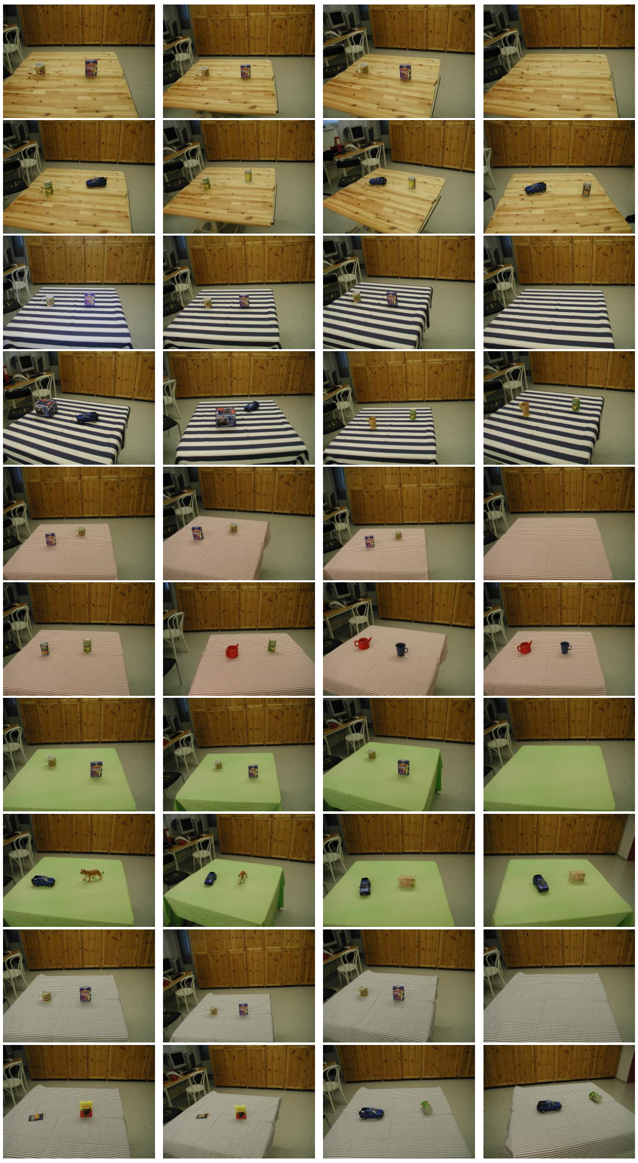

Figure 9.1: Raw images used in the experiments

set of views foveated by the attentional mechanism. Note, that no figure-ground segmentation in the foveated view was processed in this experiment.

The cup and box present in the first three views of each scene (i.e. image 1,2,3,8,9,10,16,17,18,24,25,26,32,33, and 34) was selected as targets. The motivation here is that their texture enables accurate detection by the SIFT features. All redundant saccades and other objects are candidates for false positive responses.

The SIFT recognition algorithm results in a set of matches between the observed image and a memory. By applying a threshold on the number necessary of matches it is possible to determine if the image correspond to the memory or not. The results are presented as ROC curves in figure 9.2; where the x-axis correspond to the probability to get false positive responses and the y-axis to the probability to get true recognition responses from by SIFT features for different thresholds. As the threshold is decreased the probability to get recognition and false positives are increased; chance performance is a 45^o curve similar to the top-left curve and perfect performance is a step function as the lower-right curve. We observe that both the box and cup is close to chance performance in the peripheral view and close to perfect in foveated view.

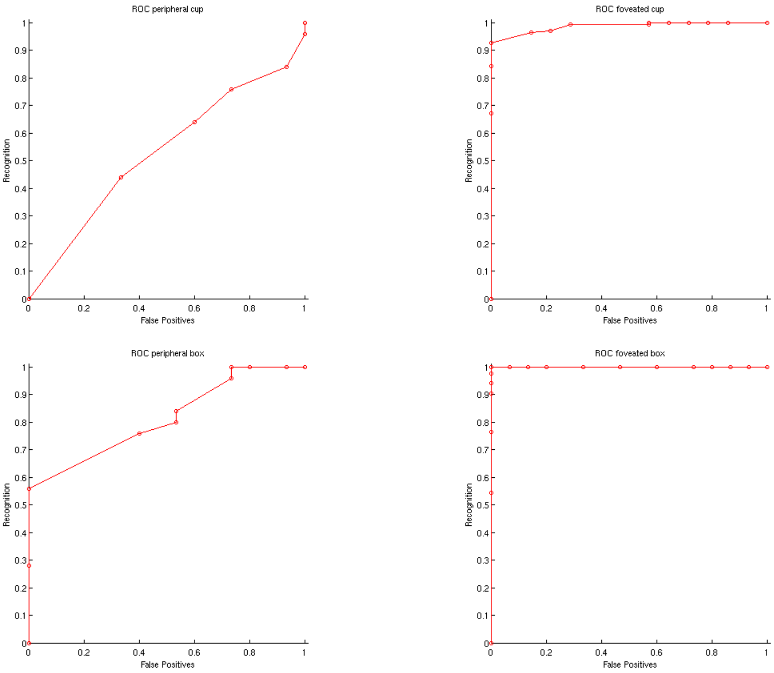

Figure 9.2: ROC curves. Top row: cup target for peripheral and foveated view. Bottom row: box target for peripheral and foveated view

It can be concluded that recognition in the peripheral view is just above chance but nearly perfect by foveated vision.

Figure-ground segmentation

When a target is detected on a background regions we have enough information to do figure-ground segmentation on the foveated view as discussed in section 7.4. It is assumed that such figure-ground segmentation can aid e.g. recognition. However, it is outside the scope of this thesis to actually perform such recognition, instead we present the figure-ground segmentation results for the detected cup and box in figure 9.3 and 9.4. The cup and box are present in the first three raw images of each scene. We observe that the cup is similar the wooden surface and the box similar to the blue stripy tablecloths.

From figure 9.3 and 9.4 it can be concluded that most of the background is suppressed, the red tablecloth induced the most background noise. It is left for future investigations how this result can be used for recognition and other vision-processes.

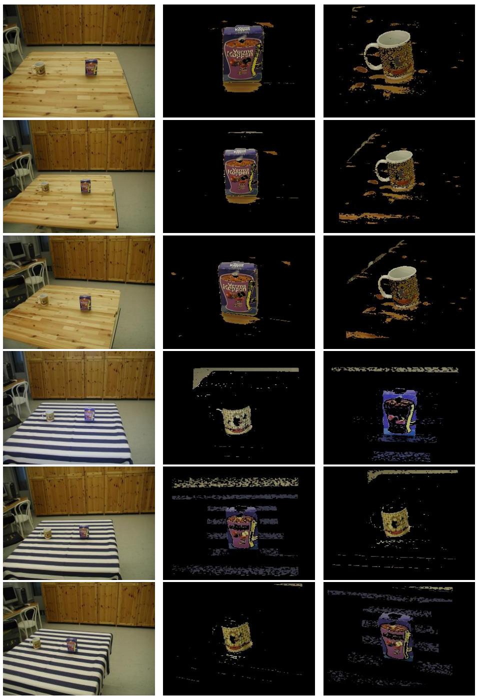

Figure 9.3: Raw images and figure-ground segmentation of detected targets

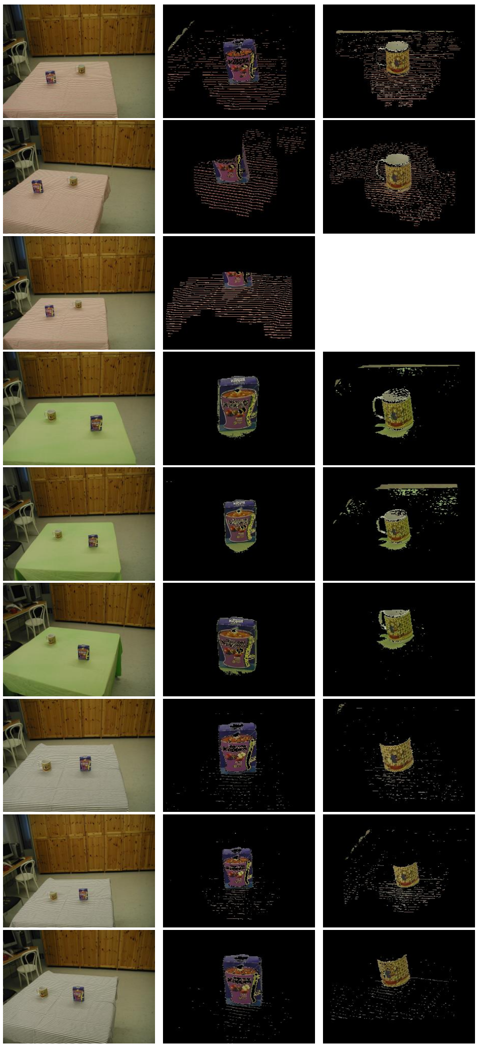

Figure 9.4: Continued list of raw images and figure-ground segmentation of detected targets

9.3 Target detection

CS-search

In CS-search we do not attempt to extract any background context other than a local surrounding, see chapter 8 for details. Since the information about large homogeneous regions is omitted, a salient table may attract many interest points around its border and between targets. Moreover, since saliency is computed on a global scale a target, which is salient with respect to the table, may not be salient enough compared to the rest of the scene.

Using each object on the table in figure 9.1 as targets, the CS-search did find 62 out of the 70 possible targets. A total of 350 interest points was generated, which correspond to about nine points per image and target detection at 18% of the interest point.

It should be emphasized that the CS-search only extracts the ten most salient points within 50% of the saliency of the strongest peak. A short evaluation indicated that the chosen thresholds are fairly good; to increase the detection rate many more irrelevant interest points was generated and hence decreasing the 18% detection rate.

Background region extraction

In the proposed model background regions are extracted which provides information of the layout of the scene. A table template, illustrated in figure 7.1, is used to take advantage of this information: the table is said to be extracted when at least one extracted background region spatially intersects 25% of the table template .

Moreover, the feature distribution of the background regions can be used as contextual information in target detection. The target objects are salient enough to be detected if the table is well extracted as a background region. However, if the background region is larger than the table the feature variance is increased and the objects can escape detection. On the other hand, if the background does not extract the whole table, the objects might not be included in the region and therefore not detected.

As a reference, each feature dimension and each parameter selection which was used redundantly in the proposed model, has been evaluated separately and the result is presented in the table 9.5. Each row represent one selection of (s, r)-mean-shift parameters and each column represent one selection of feature dimension ("combined" is a feature vector $\in \mathbf{R^6}$ with color, intensity, and orientation concatenated). The results are presented as
(#tables extracted, #targets detected); where the maximum is (40,70).

We can conclude that no single selection of mean-shift parameter and feature dimension is appropriate for all scenes and scene variations in this evaluation. The orientation node with $(s, r) = (4, 0.225)$ was best by extracting the table in 32 instances and detecting the target in 42 instances.

(s,r)	color	intensity	orientation	combined
(2,0.1)	(5,5)	(23,38)	(8,6)	(3,3)
(2,0.1)	(17,20)	(26,37)	(18,22)	(6,6)
(2,0.1)	(17,22)	(27,39)	(19,25)	(8,11)
(3,0.15)	(13,15)	(28,42)	(14,17)	(9,11)
(3,0.15)	(18,27)	(25,40)	(21,26)	(16,13)
(3,0.15)	(17,27)	(19,25)	(30,34)	(17,19)
(4,0.225)	(17,20)	(24,36)	(21,22)	(12,12)
(4,0.225)	(15,24)	(19,26)	(28,33)	(17,23)
(4,0.225)	(15,23)	(16,14)	(32,42)	(17,25)

Figure 9.5: Performance of the proposed model for a set of (s, r)-mean-shift parameters

th	#tables extracted	#targets detected	#interest points
0.1	70	66	137
0.2	70	66	137
0.3	70	66	138
0.4	70	66	137
0.5	70	66	137
0.6	70	67	131
0.7	70	65	125
0.8	70	66	123
0.9	69	64	132

Figure 9.6: Performance of the proposed model for a set of different similarity thresholds, th, in the coalition formation

Moreover, from inspection it was observed that the table in each raw image was always accurately segmented by at least one selection of feature dimension and mean-shift parameter. In more detail, the raw images on row 1-2 in figure 9.1 was well segmented by the color dimension, row 3-4 by the orientation node, and row 5-10 by the intensity dimension.

It can be concluded that with appropriate choice of feature dimension and mean-shift parameters the table can always be well segmented in each of the 40 raw images, but no single such choice is appropriate for all raw images.

The performance of the proposed model is presented in table 9.6 for a set of different similarity thresholds, th, in the coalition formation (see chapter 6). We observe that the proposed model always extract the table and 65 to 67 targets, except for $th = 0.9$. The number of interest points range from 123 to 138 and hence the detection rate from 48% to 54%. In the range $th = 0.3$ to $th = 0.8$ the number of interest points decrease monotonically from 138 to 123, which is reasonable;

Th	#targets detected	#interest points
0.1	57	82
0.2	66	123
0.4	66	139
0.6	66	144
0.8	66	146
1.0	66	146

Figure 9.7: Performance of the proposed model for a set of different thresholds of number of integrated data

the higher the threshold the higher requirements on the coalition similarity, which result in more homogeneous regions. At $th = 0.9$ the requirements are too high and the table is not extracted in one scene, since the scenes are configured to be easy to extract this is not acceptable. Moreover, $th \leq 0.5$ seems to have reached an asymptotic quality on the coalitions, with 137-138 interest points. The detection rate have a maximum at $th = 0.8$ with target detection at 54% of the interest points.

It can be concluded that although the extraction of the table is not always good enough to detect the targets, the coalition scheme clearly improves the table extraction. The number of detected targets is better than what can be achieved with any of the other evaluated methods and the detection rate is superior compared to CS-search.

Integration of conspicuous data

Within background regions the feature statistics is assumed to have Gaussian distribution. Conspicuity is defined as the confidence C with which a pixel-feature can be excluded from the Gaussian model. The confidence C is increased until less than Th % of the pixels are selected. Hence, the higher Th the lower C is allowed. When Th is equal to one, all pixels are integrated. It is desired to integrate as few pixels as possible and hence select an Th as low as possible. The performance, in terms of number of detected targets and number of interest points, for different Th-values are presented in the table in figure 9.7:

We observe that the number of target detections does not increase when Th is at least 0.2 in the set of evaluation scenes. Only the number of irrelevant interest points is increased in that range.

It can be concluded that within background regions only a sparse set of conspicuity data need to be integrated across the nodes in the set of evaluation scenes.

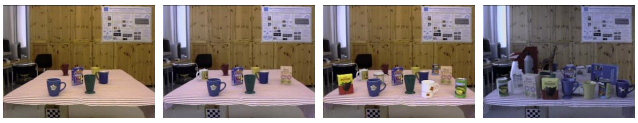

Figure 9.8: Cluttered table. Left to right image: 6 items, 8 items, 13 items, 30 items

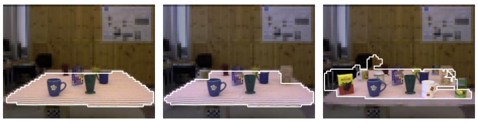

Figure 9.9: Background regions from cluttered table above

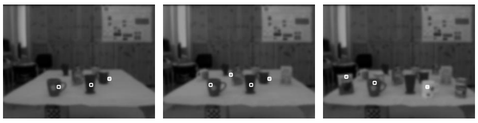

Figure 9.10: Target detections at background region above

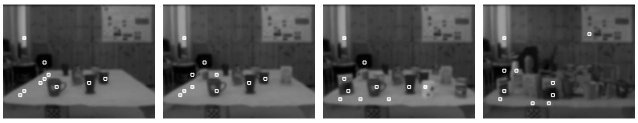

Figure 9.11: Center-surround search

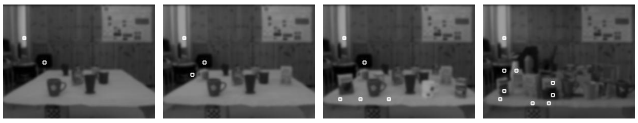

Figure 9.12: Center-surround search outside background regions

Clutter

In this section we will evaluate how the proposed model behaves in cluttered scenes. The same table as in the evaluation scenes was used, however the number of items on the table was larger: 6, 8, 13, 30 items in four consecutive observation as illustrated in figure 9.8. In figure figure 9.9 we observe that the table was represented by a background region in the first three images. However, in the second image the region does not include two objects on the left and right side and in the third image the region is larger than the table. In figure 9.10 we observe that only three to four items are detected when a background region is extracted. As a reference, the center-surround search also detects three to four items but with many more irrelevant interest points (figure 9.11). If we apply center-surround search at non-homogeneous regions, that is regions where no background regions has been detected, we get the same performance as the original center-surround search but with fewer irrelevant interest points (figure 9.12).

Interestingly, if the center-surround search is applied at non-homogeneous regions in the set of evaluation scenes, all 70 targets are detected at a total of 317 interest points. Hence, a method similar to the center-surround search is a suitable complement to the proposed model.

Finally, we will take on the challenge of the image of the cookie table at Taxinge slott in Sweden. We observe in figure 9.13 to 9.16 that only three interest points are found, representing a flower, cake and staff, and finally a cake. By increasing the threshold for number of integrated pixels from 20% to 100% the number of interest points are increased from three to five. However, in the raw images in figure 9.8 and the set of evaluation scenes this threshold does not make any difference.

It can be concluded that such clutter as in figure 9.13 is not properly handled by the proposed method.

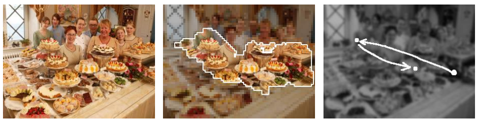

Figure 9.13: Cookie table at Taxinge slott. Left to right image: Raw image, background region, and interest points.

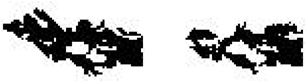

Figure 9.14: The two mean-shift segments at the intensity node, which contributed to the background region.

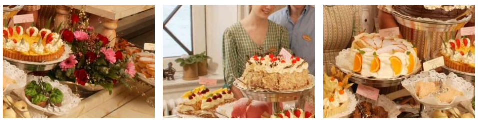

Figure 9.15: Left to right image: First to third foveal view at interest points.

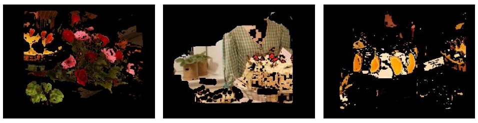

Figure 9.16: Left to right image: First to third figure-ground segmentation.

9.4 Summary

An attentional mechanism on a peripheral view, which guides a foveated view, does indeed improve recognition. Since saccades take time it is of interest to prune the number of irrelevant saccades.

A comparison to a CS-search mechanism demonstrated that the gained layout information clearly decreased the number of irrelevant interest points. Moreover, it was demonstrated that no single selection of feature dimension and mean-shift segmentation parameters could perform as well as the proposed model; the clustering among redundant segments clearly improved the quality of contextual information at the background regions. Finally, it was demonstrated that only a sparse set of conspicuous data needed to be integrated across the processing nodes at background regions.

In conclusion, the proposed model uses information of the layout to reduce the number of irrelevant interest points and the context of homogeneous regions to increase target saliency. Hence, the target detection rate is increased while only integrating a sparse set of data across the processing nodes.

Chapter 10

Discussion and future work

10.1 Discussion

It is clear from psychological experiments that visual attention plays an important role in perception. Many aspects of visual attention have been studied and quite well understood. However, when it comes to implementation of these aspects we are very limited. One aspect that has been quite successfully implemented by many researchers is the pop-out effect described in the Feature Integration Theory and other succeeding theories. Many of these implementations are based on a winner-take-all strategy and inhibition of return and correspond to human search performance well in certain situations. Most such implementations integrate a set of feature maps of different characteristics at one point in the system. However, such integration is computational and biological challenging.

Implementations of other aspects, such as object based attention, are far more limited. Object formation of certain artificial patterns and overt attention based on hand-segmented objects has been successfully implemented. However, the author has failed to find any successful implementations of visual attention based on object formation in natural scenes.

In order to address the complexity problem in integrating all feature maps at one point in the system, the proposed model attempts to find a strategy in order to only integrate a sparse set of data. One such strategy is to search for homogeneous regions, which provides a local statistical model. Using the homogeneous regions and their statistics it is possible to detect target hypothesis while only integrating a sparse set of data. Hence, object based attention is modeled as a consequence of addressing the complexity problem in data integration. Experiments show that in natural living room scenes where targets are expected to be found on a table in front of the camera, the proposed system is superior to a center-surround attentional system. The target saliency is enhanced and irrelevant interest points are suppressed.

Since the full data set is not integrated a global optimal solution cannot be found by the proposed model. However, the drawbacks of these approximations are biologically plausible. Below we list some examples of potential drawbacks of the proposed system (SYS) and related biological effects (BIO):

- (SYS) Only a sparse set of data is integrated across the computation nodes in the proposed system. (BIO) (Garner 1974), (Treisman and Gelade 1980), and (Wolfe, Cave, and Franzel 1989) have demonstrated that the visual cortex only integrate a limited amount of data across feature dimensions.

- (SYS) A set of redundant volatile mean-shift segments need to be computed by the proposed system and a negotiation scheme is needed to integrate a subset of these to extract target hypotheses. (BIO) According to the Coherence Theory, (Rensink 2000a), a set of volatile proto-objects is computed across the view sphere. Only when attention is applied to a subset of the proto-objects the subset becomes stable and makes contact to higher-level mental processes.

- (SYS) Where no background regions are found by the proposed system another attentional strategy, such as the center-surround mechanism (see chapter 8), is needed; the center-surround saliency mechanism is limited to globally salient items. (BIO) (Duncan and Humphreys 1989) has demonstrated that target saliency is important at non-homogeneous regions but not at homogeneous regions.

- (SYS) The mean-shift segmentation process takes time. (BIO) (Law and Abrams 2002) found that object based attention needs almost 200 ms.

More research is needed to verify the proposed model against psychophysical findings before any claim in its biological plausibility can be made. However, the results are appealing in the perspective that the proposed model implements an hypothesis for object based attention, not by mimicking psychophysical finding but as a consequence of computational complexity.

10.2 Future work

The background regions provide layout information and local context to enable efficient target search, hence similar to the benefits of object files. However, to model the consequences of object file maintenance in the proposed system, a dynamic scenario is needed. E.g. a mobile robot moving along a road should be able to detect the road as background region and hence obstacles on the road as targets. The road and the obstacles can thereafter be maintained as object files in a memory structure resembling the VSTM.

The experiments demonstrate that quite accurate segmentation masks of target hypotheses can be extracted. It is of interest to investigate how these masks can be used in object recognition. In a dynamic scenario the motion will be an extra dimension, which might enhance the result further. E.g. obstacles on a road are seldom defined by color or any other filter response-feature, but rather by shape. The segmentation masks of target hypotheses might aid a mobile robot to classify obstacles based on shape.

References

Biederman, I., A. Glass, and E. Stacy jr (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, Genral 97*(1), 22–27.

Bundesen, C. (1998). A computational theory of visual attention. *Philosophical Transactions of the Royal Society of London, Series B 353*, 1271–1281.

Chun (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across views. *Visual Cognition 7*, 65–81.

Chun, M. and Y. Jiang (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology 36*, 28–71.

Comaniciu, D. and P. Meer (1999, September). Mean shift analysis and applications. *Proc. Seventh Int'l Conf. Computer Vision*, 1197–1203.

DeSchepper, B. and A. Treisman (1996). Visual memory for novel shapes: Implicit coding without attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*, 27–47.

Deubel, H., B. Bridgeman, and W. Schneider (2002). Transsaccadic memory of position and form. *Progress in Brain Research 140*, 165–180.

Deubel, H. and W. Schneider (In Press). Attentional selection in sequential manual movements, movements around an obstacle and in grasping. In G. Humphreys and M. Riddoch (Eds.), *Attention in Action: Advances from Cognitive Neuroscience*. Hove: Psychology Press.

Deubel, H., W. X. Schneider, and B. Bridgeman (1996). Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vision Research 36*, 985–996.

Di Lollo, V., J. Kawahara, S. Zuvic, and T. Visser (2001). The preattentive emperor has no clothes: a dynamic redressing. *Journal of Experimental Psychology: General 3*(130), 479–492.

Draper, B. and A. Lionelle (2003). Evaluation of selective attention under similarity transforms. *Workshop on Performance and Attention in Computer Vision, Graz, Austria*, 31–38.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology, General 113*(4), 501–517.

Duncan, J. and G. Humphreys (1989). Visual search and stimulus similarity. *Psychological Review 96*(3), 433–458.

Egly, R., R. Driver, and R. Rafal (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General 123*, 161–177.

Enns, J. and R. Rensink (1990). Sensitivity to three-dimensional orientation in visual search. *Psychology Science 1*(5), 323–326.

Enns, J. and R. Rensink (1992). An object completion process in early vision. *Investigative Ophthalmology & Visual Science 33*(1263).

Eriksen, C. and J. Hoffman (1972). Temporal and spatial characteristics of selective encoding from visual displays. *Perception & Psychophysics 12*(1), 201–204.

Eriksen, C. and J. Hoffman (1973). The extent of processing of noise elements during selecive encoding from visual displays. *Perception & Psychophysics 14*(1), 155–160.

Eriksen, C. and J. St. James (1986). Visual attention within and around the field of focal attention: a zoom lens model. *Perception & Psychophysics 40*(4), 225–240.

Findlay, J. and I. Gilchrist (2001). Active vision perspective. In M. Jenkin and L. Harris (Eds.), *Vision & Attention*, Chapter 5, pp. 83–103. Springer Verlag.

Forsyth, D. and P. Ponce (2002). *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology, Genral 108*(1), 316–355.

Fryklund, I. (1975). Effects of cued-set spatial arrangement and target-background similarity in the partial-report paradigm. *Perception & Psychophysics 17*(4), 375–386.

Fudenberg, D. and J. Tirole (1991). *Game Theory*. MIT Press.

Garner, W. (1974). *The processing of information and structure*. Hillsdale NJ: Erlbaum.

Grossberg, S. and R. Raizada (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research 40*, 1413–1432.

Harris, L. and M. Jenkin (2001). Vision and attention. In M. Jenkin and L. Harris (Eds.), *Vision & Attention*, Chapter 1, pp. 1–18. Springer Verlag.

Henderson, J., C. Williams, M. Castelhano, and R. Falk (2004). Eye movements and picture processing during recognition. *Perception & Psychophysics 65*(5), 725 – 734.

Hoffman, J. (1999). Stages of processing in visual search and attention. In B. Challis and B. Velichkovsky (Eds.), *Stratification in Cognition and Consciousness*. John Benjamins Publishing Co.

Hommel, H. (2002). Responding to object files: Automatic integration of spatial information revealed by stimulus-response compatibility effects. *The Quarterly Journal of Experimental Psychology 55A*(2), 567–580.

Hornof, A. (2002, May). Cognitive modeling, visual search, and eye tracking. *ONR Attention, Perception and Data Visualization Workshop George Mason University*.

Horowitz, T. and J. Wolfe (1998). Visual search has no memory. *Nature 357*, 575–577.

Itti, L. and K. Koch (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*, 1489–1506.

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

James, W. (1890). *The Principles of Psychology, vol. I*, Chapter XI, Attention. New York: Henry Holt and Company.

Johansson, R., G. Westling, A. Backstrom, and J. Flanagan (2001, September). Eye-hand coordination in object manipulation. *The Journal of Neuroscience 21*(17), 6917–6932.

Julez, B. (1981). Textons, the elements of texture perception and their interactions. *Nature 290*, 91–97.

Kahneman, D., A. Treisman, and B. Gibbs (1992). The reviewing of object files. *Cognitive Psychology 24*(2), 175–219.

Koch, C. and S. Ullman (1984, January). Selecting one among the many: A simple network implementing shifts in selective visual attention. Memo at Massachusetts Institute of Technology Artificial Intelligence laboratory and Center for Biological Processing Whitaker College.

Kraus, S. (2001). *Strategic negotiation in multiagent environments*. MIT Press.

Land, M. F. and S. Furneaux (1997). The knowledge base of the oculomotor system. *Phil Trans R Soc Lond B*(352), 1231–1239.

Lavie, N. and J. Driver (1996). On the spatial extent of attention in object-based visual selection. *Perception & Psychophysics 58*(8), 1238–1251.

Law, M. and R. Abrams (2002). Object-based selection within and beyond the focus of spatial selection. *Perception & Psychophysics 64*(7), 1017–1027.

Lee, D. and N. Chun (2001). What are the units of visual short-term memory, objects or spatial locations? *Perception & Psychophysics 63*(2), 253–257.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision 30*(2), 77–116.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pp. 1150–1157.

Maljkovic, V. and K. Nakayama (1994). Priming of pop-out: I. role of features. *Memory & Cognition 22*, 657–672.

Maljkovic, V. and K. Nakayama (1996). Priming of pop-out: Ii. the role of position. *Perception & Psychophysics 58*(7), 977–991.

Marcus, D. and D. van Essen (2002, November). Scene segmentation and attention in primate cortical areas v1 and v2. *J Neurophysiol 88*(5), 2648–58.

McFadden, S. and J. Wallman (2001). Shifts of attention and saccades are very similar. are they causally linked? In M. Jenkin and L. Harris (Eds.), *Vision & Attention*, Chapter 2, pp. 19–40. Springe Verlag.

Milanese, R. (1993). *Detecting Salient Regions in an Image:From Biological Evidence to Computer Implementation.* Ph. D. thesis, University of Geneva.

Mitroff, S., D. Simons, and D. Levin (In Press). Nothing compares 2 views: Change blindness results from failures to compare retained information. *Perception & Psychophysics*.

Navalpakkam, V. and L. Itti (2002). A goal oriented attention guidance model. In H. B. et. al. (Ed.), *Biologically Motivated Computer Vision*, pp. 453–462. Springer.

Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology 76*, 376–385.

Oliva, A., A. Torralba, M. Castelhano, and J. Henderson (2003). Top-down control of visual attention in object detection. *International Conference on Image Processing (ICIP) 1*, 253–256.

Olson, I. and Y. Jiang (2002). Is visual short-term memory object based? rejection of the "strong object" hypothesis. *Perception & Psychophysics 64*, 1055–1067.

O'Reagan, J. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology 46*, 461–488.

O'Regan, J., H. Deubel, J. Clark, and R. Rensink (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition 7*(1), 191–211.

Osborne, M. and A. Rubinstein (1999). *A course in game theory.* Cambridge: The MIT Press.

Ouerhani, N., R. von Wartburg, H. Hugli, and R. Muri (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis 3*(1), 13–24.

Palmer, S. (1999). *Vision Science, Photons to Phenomenology*, Chapter Visual Selection: Eye Movements and attention. Cambridge: The MIT Press.

Posner, M. (1978). *Chronometric explorations of mind*. Hillsdale NJ: Erlbaum.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology 32*, 3–25.

Posner, M., C. Snyder, and B. Davidson (1980). Attention and the detection of signals. *Journal of Experimental Psychology General 109*, 160–174.

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory 2*(5), 509–522.

Ramström, O. and H. Christensen (2002). Visual attention using game theory. *Biologically Motivated Computer Vision*, 462–471.

Ramström, O. and H. Christensen (2004a). Distributed computing of attention. *To appear in the Springer edition on the WAPCV 2004 workshop*.

Ramström, O. and H. Christensen (2004b). Object based visual attention: Searching for objects defined by size. *WAPCV proceedings*, 9–16.

Ramström, O. and H. Christensen (2004c). Object detection using background context. *ICPR proceedings*.

Ramström, O. and F. Fuesjö (2004, June). Recognition with and without attention. CogVis (IST-2000-29375) Deliverable, Enclosure 5 to DR.1.5.

Rao, R., G. Zelinsky, M. Hayhoe, and D. Ballard (2002). Eye movements in iconic visual search. *Vision Research 42*, 1447–1463.

Rensink, R. (2000a). The dynamic representation of scenes. *Visual Cognition 7*(1), 17–42.

Rensink, R. (2000b). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition 7*(1), 345–376.

Rensink, R. (2001). Change blindness: Implications for the nature of visual attention. In M. Jenkin and L. Harris (Eds.), *Vision & Attention*, Chapter 9, pp. 169–188. Springe Verlag.

Scinto, L., R. Pillalamarri, and R. Karsh (1986). Cognitive strategies for visual search. *Acta Psychologica 62*, 263–292.

Shehory, O. and S. Kraus (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence 101*(1–2), 165–200.

Simons, D. and D. Levin (1997). Change blindness. *Trends in Cognitive Science 1*, 261–267.

Simons, D. and S. Mitroff (2001). The role of expectations in change detection and attentional capture. In M. Jenkin and L. Harris (Eds.), *Vision & Attention*, Chapter 10, pp. 189–208. Springe Verlag.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology 18*, 643–662.

Sun, Y. and R. Fisher (2003). Object-based visual attention for computer vision. *Artificial Intelligence 146*, 77–123.

Thielscher, A., A. Schuboe, and H. Neumann (2002). A neural model of human texture processing: texture segmentation vs. visual search. In H. B. et. al. (Ed.), *Biologically Motivated Computer Vision*, pp. 99–108. Springer.

Torralba, A. (2003). Contextual priming for object detection. *Proc. of the 2001 conference, Adv. in Neural Information Processing Systems 2*(14), 1303–1310.

Treisman, A. and G. Gelade (1980). A feature-integration theory of attention. *Cognitive Psychology 12*, 97–136.

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

Treisman, A. and S. Sato (1990). Conjunction search revisited. *Journal of Experimental Psychology:Human Perception and Performance 16*(3), 459–478.

Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Science 13*, 423–469.

Tsotsos, J., M. Pomplum, Y. Liu, J. Martinez-Trujillo, and E. Simine (2002). Attending to motion: localizing and classifying motion patterns in image sequences. In H. B. et. al. (Ed.), *Biologically Motivated Computer Vision*, pp. 439–452. Springer.

Tsotsos, J., M. Sean, W. Wai, Y. Lai, N. Davis, and F. Nuflo (1995). Modelling visual attention via selective tuning. *Artificial Intelligence 78*, 507–545.

Varma, M. and A. Zisserman (2002, may). Classifying images of materials: Achieving viewpoint and illumination independence. In *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, Volume 3, pp. 255–271. Springer-Verlag.

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior.* Princeton, N.J. : Princeton University Press, 1980, c1972.

Walker, L. and J. Malik (2002). When is scene recogniton just texture recognition? *Journal of Vision 2*(7), 255a.

Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review I 2*, 202–238.

Wolfe, J. (1999). Inattentional amnesia. In Coltheart (Ed.), *Fleeting Memories*, pp. 71–94. Cambridge, MA: MIT Press.

Wolfe, J. and S. Bennet (1997). Preattentive object files: Shapeless bundles of basic features. *Vision Research 37*(1), 25–43.

Wolfe, J., K. Cave, and S. Franzel (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance 15*(3), 419–433.

Wolfe, J. and J. DiMase (2003). Do intersections server as basic features in visual search? *Perception 32*, 645–656.

Wolfe, J. and G. Gancarz (1996). Guided search 3.0. *Basic and Clinical Applications of Vision Science, Dordrecht, Netherlands: Kluwer Academic*, 189–192.

Wolfe, J., A. Oliva, T. Horowitz, S. Butcher, and A. Bompas (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision Research 42*, 2985–3004.

Yeshurun, Y. and M. Carrasco (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature 396*, 72–75.

Zelinsky, G. (2001). Eye movements during change detection: Implications for search constraints, memory limitations, and scanning strategies. *Perception & Psychophysics 63*, 209–225.