Visuo-cognitive Routines: reinterpreting the theory of visual routines as a framework for visual cognition

Wouter Kruijne and John K. Tsotsos

Technical Report CSE-2011-05

August 31 2011

Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario M3J 1P3 Canada

Visuo-cognitive Routines

reinterpreting the theory of visual routines as a framework for visual cognition

Authors:

Wouter Kruijne
Dept. of Artificial Intelligence
Faculty of Exact Sciences
Vrije Universiteit Amsterdam
&
John K. Tsotsos
Dept. of Computer Science & Engineering
and Centre for Vision Research
York University

August 31st, 2011

Abstract

In solving its tasks, the visual system must be capable of more than simple detection of features. Good performance requires cognitive reasoning about the information that is extracted, which is highly task-dependent. Therefore, a mechanism of visual cognition is needed in order to guide the interaction of information coming from the senses and higher order processes that regulate task performance. Such visual problem solving was addressed by Ullman (1984) with the theory of visual routines, where tasks are solved by sequences of elemental operations. Although succesful, this theory relies on assumptions on vision and attention that are challenged by our modern understanding of these domains. This study presents a functional analysis of the visual routines framework and identifies elements that need to be reconsidered in order to provide the same functionality, yet conform with our modern understanding of visual processes. The proposed reconsiderations help shape a new framework for visual cognition that integrates the visual pathway, peripheral vision processing, inhibition of return, visual working memory processes, the attentional mechanisms that interact with these components, and higher order cognitive production rules. The operations are expressed both as general *methods* and applied *scripts*. Example problems that have been tackled using the classical visual routines framework are used to illustrate how this new framework operates.

List of Figures

Contents

Chapter 1

Introduction

In vision research literature it is not uncommon to introduce the topic of vision by describing the authors amazement over the capacities of the human visual system. There is an impressive amount of visual information that is presented to our sensory systems, and somehow the brain is able to extract relevant information from this constant stream. Executing our daily tasks largely relies on our capabilities to visually detect, recognize, search or obtain descriptions of what is presented in the scenes we percieve. It may be debatable whether the visual system is the most important sensory system, but it is certainly the most extensively studied one, and the striate and extrastriate cortex – dedicated to the processing of visual stimuli – occupy about 40 percent of all cortical structures.

Despite the complexity of some visual tasks, the underlying principles of visual processing in the brain seem relatively simple. The neurons in the visual areas appear to act as simple feature detectors. The neurons in early areas, such as V1 and V2, detect the presence of simple features such as spatial frequency, orientation, direction and speed, whereas neurons in later areas can respond to more complex motion patterns and stimuli, including for example cells tuned to respond to faces (the FFA) or scenes (the PPA) (Orban, 2008). This increase in complexity originates from the connectivity between the neurons in the visual pathway. The hierarchical organization of the system allows for information from simple feature detectors to integrate into more complex cells, first introduced by Hubel and Wiesel (1959).

However, the simple detection mechanism implemented by the neurons in the visual pathway is not enough to solve all tasks the visual system must complete, due to the versatility of possible visual scenes. For example, the visual system can analyze contours and curves, but there are infinite possibilities to construct arbitrary curves and contours, which would make it impossible to represent every possible combination by a collection of feature detectors (Ullman, 1984). A similar problem is inherent in the task of visual search for a target in a scene. Due to the combinatorial explosion of

possible feature combinations of the target, and the possible composition of the scene, a brute-force feedforward detection mechanism is impossible for such tasks (Tsotsos, 1990). These tasks illustrate that other mechanisms besides simple detector neurons operate in the visual system as well. Probably the most extensively studied among these mechanism is visual attention, generally viewed as a selective mechanism that limits visual processing to information that is relevant to the task. However, visual attention alone appears not to be enough: as the example of contours and curves illustrate, the visual system needs mechanisms that allow for cognitive reasoning about the information extracted, in order to analyze complex configurations of features accurately. The set of such mechanisms, including visual attention, can be addressed as visual cognition.

Visual cognition is an important aspect of visual perception, as it is needed in any task that requires further processing than hierarchical, convergent feature detection. In this case, "further processing" can be defined as any process that utilizes these sensory representations, such as integrating information from multiple fixations, or even interpreting the scene and reasoning about its meaning. This makes visual cognition a broad concept: it can be viewed as the intermediate level of computation between low-level sensory mechanisms – such as feature detection – on the one end, and higher level cognition processes – such as monitoring visual task performance – on the other end. This implies that a theory for visual cognition will have to seamlessly integrate into theories of both ends. Probably due to the versatility of mechanisms on both ends, visual cognition as a whole is studied relatively little and a complete theory of visual cognition is still absent (Cavanagh, 2011).

A notable exception is the theory of visual routines (Ullman, 1984). Devised over 25 years ago, this theory originated from the simple observation that the visual system seems to be capable of effortlessly solving tasks that can not be solved by feedforward detection of features, but instead required a set of mechanisms that operated on the representation constructed by these feature detectors. An appropriate sequence of such simple operations could be used to model the reasoning process that was used to solve such tasks. Purely based on the functional requirements of such operations, Ullman illustrated a framework where a finite set of such operations, if sequenced accordingly to construct visual routines, could be used to solve more complex visual tasks. Without mentioning the term, he had devised a model for visual cognition, albeit based on relatively simple models of sensory vision and visual attention. The visual routines framework became a popular approach for computer vision systems that used reasoning beyond simple detection to extract visual information. Despite its succes, the visual routines framework was constructed based on assumptions of the visual system that since have been proven to be wrong or at least incomplete. However, this does not invalidate the concept as a whole and some of the proposed mechanisms

might still be used to address the issues of visual cognition.

The formulation of the visual routines theory is heavily influenced by the theory of vision by Marr (1982) – which will be discussed in chapter 2. This is reflected in Ullman's view on vision, but also in the type of problem analysis that is used to formulate visual routines theory. Both Marr and Ullman use a functional approach and focus on extensive analysis of the problem as a starting point, rather than on experimental observations. Due to this type of analysis, the visual routines theory provides a framework that can not only be applied as a theory for human visual cognition, but also provides an approach to the design of computer vision systems. Since then, vision research has elucidated many facets of the visual pathway and it's properties, but in the design of a model for visual cognition, the empirical approach has been less succesful. Due to the nature of visual cognition, a complex interwoven collective of mechanisms that can be used to extract goal-dependent independent, the functional approach as used by Ullman may still prove to be more succesful.

In this thesis, an attempt will be made to update the theory of visual routines in order to construct a framework for visual cognition that is consistent with our modern understanding of vision and visual attention. This requires (1) a thorough analysis of the theory of visual routines, including the experimental findings and the understanding of vision and cognition, as well as the assumptions that helped construct it; (2) a review of our modern understanding of vision and visual attention and the prominent changes in the view on these systems and mechanisms; (3) an illustration of the discrepancies between the modern view and the classical view, as well as how this affects the visual routines theory; (4) a proposal how the visual routines theory would have to updated to provide a model for visual cognition that includes our modern view, yet is still capable of solving the same visual problems that have been attributed to visual cognition.

Therefore, the following chapters will address these points as follows: **Chapter 2, Classical visual routines and attention** reviews several classical theories of vision and attention that have influenced the formulation of the visual routines theory. Then, the visual routines theory will be detailed, followed by a review of various systems and models that are based on the framework. These implementations help fill in details and problems that were not addressed in the original description of the framework. **Chapter 3, Vision and attention** reviews findings that illustrate our modern understanding in vision and visual attention. First, new findings on the visual system will be discussed. Then, attention will be reviewed, using both a computational approach to address its function and a mechanical approach detailing the timescale and localization of the attentional mechansims. Based on these chapters, **Chapter 4, The theory for visual routines**, discusses in which respects the visual routines framework is incompatible with the new view and introduces adjustments to the framework

that could resolve these issues. This way, step by step, a new framework is constructed. **Chapter 5, Visual problem solving using the new framework** then illustrates how this modern framework could be used to organize operations to realize visual cognition in terms of visual problem solving. This is illustrated by three example problems that are solved using the new framework. Finally, **Chapter 6, Discussion** critically reviews the new framework and addresses its implications for research in vision and visual cognition.

Chapter 2

Classical visual routines and attention

For a thorough understanding of the theory of visual routines as a theory on its own or as a model for visual cognition, one cannot discard the rich history of vision research that led to its conceptualization. Not surprisingly, many of the assumptions made in the model and many of the questions the model attempts to address, naturally follow from the view on vision, visual attention and visual cognition that were dominant at the time. This chapter will try to illustrate this by reviewing prominent theories on vision. After this review, the original theory of visual routines will be introduced, addressing some of the issues that the framework was designed to solve. However, as its section (2.2) will illustrate, the theory provides only a framework and does not address its full potential as a visual problem solver. Therefore, the final section of this chapter reviews several systems that use the various aspects of the theory as their inspiration. These systems will then be used, in part, to 'fill the gaps' in the framework, and thus will form the overall picture of of classical visual routines that will then be re-examined.

2.1 Classical vision and attention

When defining the classical view on vision and visual cognition that has led to the development of the theory of visual routines, it seems appropriate to describe the heavily influential theory of vision by Marr (1982). Marr pointed out that although the 1950's and the 1960's had seen great advances for the study of visual perception, including advances in the fields of neurophysiology, computer vision and exeperimental psychology, the 1970's had brought no revolutionary advances and the development of the field appeared stagnant. The reason for this, Marr proposed, was similar to the reason that the once influential Gestalt school of psychology – which searched for the laws and conditions under which objects and their seg-

ments were seemingly effortlessly percieved as a whole (see Koffka, 1999) – had been largely abandoned: although they managed to devise detailed 'laws' for visual phenomena, they provided little explanatory power due to the lack of a supporting theory. He pointed out that most neurophysiological developments since then had fallen victem to the same deficiency: pioneer studies on receptive fields and neuronal selectivity (e.g. Hubel & Wiesel, 1959) merely *described* the behavior of cells and failed to *explain* it. The main argument is that complex brain processes like vision can not be understood by mere descriptions of the properties of its components but require a theory which describes the goal of the process, so that findings can be linked to the underlying mechanisms that realize this goal.

To detail this approach to complex systems, Marr described three levels of analysis. The *computational level* describes the goal of the system or process; what the goal of the computation is, and why such a goal seems appropriate. The *algorithmic level* describes how this goal is computed; what representations are used and how they are transformed. The lowest *implementational level* finally describes how these mechanisms are physically realized. The problem with the dominant observational approach was that without a thorough understanding of their function within the context the appropriate level of analysis, the functionality of these observed phenomena within the system as a whole would remain hidden. For example, the observation that the visual system is equipped with edge detector cells is meaningless unless we can attribute how these cells are used to achieve the eventual computation goal of vision.

Thus, Marr proposed a computational goal for the vision problem, and derived a strategy to solve the problem. In his theory, the goal of visual perception is to create a complete, threedimensional, object-centered description of the percieved scene. The construction of this 3D-description consists of three stages. First, the retinal image is processed to construct a 'primal sketch', which extracts intensity changes and their geometrical organization. This is then used to construct a '$2\frac{1}{2}$D sketch', transforming the image into a viewer-centered field with rough depth descriptions and orientations. The final stage is reached by transforming this into an object-centered 3D representation, in which shapes and surfaces are hierarchically represented. Marr's proposed hierarchical representation was constructed of volumetric primitives; a skeleton constructed from cylinders of variable size (Binford, 1971; Marr & Nishihara, 1978).

Though construction of a 3D model is described as the computational goal, Marr acknowledges that this is not all vision does; the extraction of other information such as motion, color and reflectance is of significant importance. However, these goals are considered secondary, and can be achieved using the 3D model after it has been constructed, and only when needed. The process of construction and enrichment of the 3D model is also the only representation that is affected by task guidance and top-down pro-

cesses; Marr emphasizes how up to the construction of the $2\frac{1}{2}$D sketch, the process is little to not influenced by 'higher-order processes'. This implies that these sketches will have to be recomputed everytime the scene changes, which is why the construction of the $2\frac{1}{2}$D sketch will have to be fast. Therefore, Marr describes that it is done in a single feed-forward pass through the visual system, which would have to take about 160ms, a number that since indeed has been related to the processing speed of a single pass through the visual system (Thorpe, Fize, & Marlot, 1996). The 3D model on the other hand makes use of memory, is much more stable, and can be adjusted by integrating multiple views of the same objects. The process of enriching object knowledge by using multiple views has been defined as active vision (Bajcsy, Computer, & Electrical Engineering, 1985).

Marr's work did not only introduce a novel description of the visual processing stream, but it also discussed a novel approach to the analysis of complex information processing systems in order to support this description. Therefore, a majority of his work focusses on the justification both this method of analysis and the representations used in the visual processing pathway. As a consequence, his consideration of the features used to construct these representations is limited, in that the sketches seem to be constructed primarily out of edges and boundaries. The vastness of features for which early neuronal selectivity already had been discovered (e.g. color, motion and disparity) is mostly discarded and only used for later enrichment of the sketches when necessary for the task. Also, there is no explicit role for attention defined in Marr's theory of vision: processing up to the $2\frac{1}{2}$D sketch is determined by transformations of the entire input image, and the eventual 3D-model is constructed incrementally by combining these sketches. The absense of attention is remarkable, as the concept was already well established within vision research. In the recent years before Marr's book was published, various theories and models of attention had already been devised, including models of Early Selection (Broadbent, 1958), Late Selection (Deutsch & Deutsch, 1963), the Attentional Spotlight (Shulman, Remington, & Mclean, 1979), and Feature Integration theory (Treisman & Gelade, 1980). Although these theories do not present a clear consensus on how attentional mechanisms operate, they are all based on experimental evidence that indicates that the human visual system is incapable of immediate detailed processing of all the information present in the retinal image. Therefore, processing needs to be limited to the information at the focus of attention. Attention therefore appeared to be selective and to restrict processing to certain information only. The control of attention is therefore an important aspect of image analysis.

The view of selective attention does not imply that immediate processing of the entire input image as proposed by Marr does not occur, but it will simply not allow for full potential information analysis. The feature integration theory of attention (Treisman & Gelade, 1980) illustrates this notion

by proposing a dissociation for attentive versus unattentive processing, and it seems to be the most influential theory of attention with respect to the theory of visual routines. It describes how the visual system can process a wide number of features in parallel, but by default these features 'float free'. Selective attention is then defined as a process that 'glues' the features at a certain location together to form a coherent representation. This could account for observed behavior in visual search: looking for a red bar among green ones with the same orientation can be done rapidly because the singleton 'pops out', but looking for a conjunction of features (e.g. a red, vertical bar among combinations of red and green vertical and horizontal bars) requires serial processing of each item, because selective attention needs to fixate on each item to bind the features. Also, when attention can not be employed, the free-floating features might bind in an arbitrary fashion, an experience named 'an illusory conjunction'. These illusory conjunctions will be influenced by experience when possible, which is why in real-life situations the unattended objects in a scene will not arbitrarily glue together and we don't experience incoherencies in our periphery (such as a green sun and yellow grass). To *assure* a correct perception of the world however, feature-integration theory states selective attention is required.

As the next sections (2.2 and 2.3) will illustrate, the notion of selective attention resonates within the theory of visual routines, both in the original formulation and the work that followed. One model in particular has been very influential for many models that followed the visual routines paradigm, which will be discussed here; the first saliency map model (Koch & Ullman, 1985) [1] which was designed to implement the feature integration theory. In this model, early vision is implmented by the extraction of a set of topographical feature maps, which indicate the presence of that feature at an image location by activation in a node at the corresponding location in the map. Selectively attending a certain location is formulated as 'lifting out' the features at that location and combining them in a higher order more abstract representation, which is what feature integration theory proposed. The model proses that the location that captures attention and is selected first will be determined by the cospicuity of the image at that location. The conspicuity of each location is encoded in a global *saliency map*, which is constructed by summation of the activation in the feature maps at the corresponding locations. The point with maximum saliency is computed by a Winner-Takes-All (WTA) algorithm, and will be selected. Attentional shifts are then determined by decreasing the feedforward activation from all maps at the selected location, which will cause the next most conspicuous

[1] As the term 'first' indicates, the concept of saliency has become an important field of study in visual attention, and many models have been designed based on this principle. The next chapter on attention (section 3.3 in particular) will discuss the evolution of saliency-based models in more detail, as well as the relation between these models and attention.

location to win the WTA competition and to be selected by attention. This mechanism should then account for the findings of feature integration theory, because a singleton will be the first attended location with the most saliency, whereas the search for a conjunctive target will need to cover more steps.

These two models do not only illustrate the view on visual attention at that time, but also a critical hole in the theory of vision up until then. The last section of Koch and Ullman's paper states that the saliency model would be able to account "for such visual routines as tracking of contours, counting objects or marking a specific location", but it seems that such visual tasks would require more complex involvement of task guidance than the automatic and straightforward approach of pointwise lowering of feature values described here. Rather, a mechanism is needed that interprets the feedforward activation using more than one single selected location in order to determine the next point of fixation. This is what the classical theory of visual routines was intended to provide (Ullman, 1984)

2.2 The visual routines framework

Although the previous section illustrates visual routines theory as a model of visual cognition, this term was never used in the original paper. Instead, what drove Ullman to formulate the theory was the observation that certain tasks can be solved by our visual system, but not by the simple feedforward approach that so many theories such as Marr's proposed. The main example used to illustrate these tasks is the extraction of visuospatial relations; tasks such as determining whether a marker is inside or outside a contour, or whether two markers are on the same curve require not only the location of targets as in visual search, but also reasoning about these locations. This, Ullman proposed, is done by a visual processor which operates upon the feedforward representation by means of 'elemental operations'.

Therefore, Ullman proposed a visual framework, where visual processing is divided into two stages. In the first stage, the bottom-up *base representation* is constructed which is effectively Marr's $2\frac{1}{2}$D-sketch. After construction of this representation, sequences of elemental operations are applied by the visual processor. These sequences of elemental operations, that describe a program to solve a visual task, are called *visual routines*. Each elemental operation uses the information from the base representation to construct *incremental representations*, which contain new information which can be used by the next elemental operation. Thus, by application of a visual routine, the sequence of elemental operations can construct a series of incremental representations, where the last representation provides the solution to the visual problem.

Based on the problem type of visuospatial relations, Ullman detailed a

set of elemental operations that, if sequenced correctly, should be able to solve such tasks. These elemental operations are:

- *Shift of processing focus*: In order to derive information at different parts of an image, the processing focus should be able to shift. This operation is loosely based on the 'spotlight' model of selective attention.

- *Indexing* the locations where the processing focus should shift to: Ullman describes how certain pop-out objects in the visual field allow for immediate indexing. There may be other mechanisms at work here, but these are not addressed in detail.

- *Marking* meets the need to mark certain locations, for example those that already have been attended. This is needed in tasks such as counting objects on a screen, or tracing a circle and marking the end point.

- *Boundary tracing* denotes the capability to trace curves. This mechanism needs to also be capable to trace incomplete boundaries (such as dashed curves), and be capable to handle intersections. How this is achieved is not discussed.

- *Coloring*: spreading activation all over an object within its boundaries. This operation could be used to identify whether a location lies inside or outside a contour. Like boundary tracing, this mechanism should be able to spread activation, even with incomplete boundaries.

These operations are entirely based on the example problems of visuospatial tasks, but this list is illustrative and not exhaustive. Instead, they form a subset of a fixed library of elemental operations that can be used by the visual processor.

Although it was not explicitly mentioned as one of its goals, the visual routines framework provides a very powerful theory for visual cognition. The description of vision as a two-stage process illustrates the difference between the sensory aspect of vision and the task-dependent aspect that requires cognitive reasoning, which is expressed by the visual routines. The routines that are used are determined by the current task of the system, and thereby the visual processor provides the intermediate layer between the sensory system and higher-order cognitive processes. The strength of visual routines as a model for visual cognition is its modular nature: A routine provides a program that can be used to solve various problems of the same type, and all visual tasks can be solved by appropriate sequencing of elemental operations that can be drawn from a finite library. Moreover, as the 'indexing', and 'shifting' operators illustrate, the theory also encorporates control over visual attention. Therefore, the framework also provides

a theory of how attention is controlled and can be utilised in a manner dependent on the visual task at hand, which is absent in an automated model of visual attention such as the saliency map model.

As a model for visual cognition, the visual routines framework details the wide range of aspects there are to solving a visual task, which need to be detailed for a complete understanding of the visual system. However, the only component of this framework that is detailed in the paper itself are the proposed elemental operations, which only form a subset of the complete library. Still, the framework in itself provides a description of how a wide variety of visual tasks is solved, and a large body of work followed this description, which seemed to fill in the details of the various components of the framework. Below, an overview is given of the work on visual routines after Ullman, that has had considerable success in filling in the framework provided.

2.3 Implementations based on the framework

The Visual Routines framework makes many claims on how the visual system processes an input image. An important notion is that there appears to be a *finite* number of basic operations that can be *sequentially* scheduled to meet the demands of *varying tasks*. This discretized approach has made the framework popular for computer vision studies that attempt to implement a general approach to visual cognition, in order to solve a variety of visual tasks for a wide range of image types. However, as indicated in the above section, the theory also only provides a framework and is largely undefined. Attempts at implementation of this framework have therefore led to various interpretations, emphasizing different properties of visual cognition. This section will review these studies by describing three classes of interpretations.

2.3.1 Interaction with low-level vision

The mere definition of Visual Routines requires that they operate on the result of the first stage of vision, the base representation, and therefore interaction with this stage is a prerequisite. Most models of this stage are based on representation construction as proposed by Marr (1982), and implement neuronal units that act as feature detectors and are organized in feature maps. A model that was designed to complement such a representation is the saliency map model (Koch & Ullman, 1985). This model illustrates how such a representation can be used to model attention, by constructing several feature maps and integrating them into one representation, and clearly was developed at the same time as the visual routines framework and designed as a supplement to it. Although this system has introduced a popular bottom-up attentional strategy, it details very little interaction with this representation in order to solve visual tasks. The visual routines

framework approaches visual problems by operations that interact with such feedforward representations. This section discusses systems that illustrate how such operations can be realized, and indicates that only slightly differing ways of interaction can make for a very different realization of visual routines.

One of the first robot heads built for tracking (Clark & Ferrier, 1988) processed its camera image in a way very similar to the first saliency map mechanism. It also produced a set of feature maps, indicating the extent to which a primitive (color, orientations, texture) is present at a certain location in the image. In this system however, each feature map i is then multiplied by a factor $k_i(t)$ before integration into a saliency map. The highest value within the saliency map again determines the focus of attention, which is the location the robot will fixate. Yet in this model, not the saliency of a certain location, but the weight $k_i(t)$ of entire feature maps decays over time, which causes attention to eventually shift to locations with other salient features. The authors emphasize the two different types of vision involved in such a task: active vision, which involves moving the robot's cameras or enhancing the zoom to obtain multiple images of a scene, and attentive vision, which concerns the selection of regions of interest to look at. Within this scheme, visual routines are the mechanism that realize attentive vision: the routine determines how the decay of each $k_i(t)$ is regulated, which determines the order and timescale of the salience of each feature. Active vision is then the result of the attentional selection process.

Conversely, it is stressed by Brunnström, Eklundh, and Uhlin (1996) that active vision operations such as fixation and accomodation need not just be the result of visual routines, but these motor operations provide information to be used as input information as well. They describe a system in which a visual routine is used to identify man-made objects in a natural scene. It uses both information obtained from a 'visual front-end', a feed-forward visual processor which constructs a very sparse base representation detecting junctions and edges within the image, and information from active vision, which provides information regarding depth and location as it establishes fixations on the points of interest. The routine takes one of the junctions in the images, classifies it, and then traces one of the legs of the junction to connect it with other junctions. Depth and fixation information can then be used to ensure that conjoined edges are part of the same junction, as well as to obtain richer object models.

Jeeves (Horswill, 1995) is a system that answers queries about the presence or absence of colored objects in a natural scene. Based on the saliency map model, Jeeves extracts low-level feature maps for edges, orientation and color which are then combined into a saliency map. However, the color maps are also used to construct another map, which can roughly segment objects based on color blobs. The segmentation map and the saliency map are then combined which results in an attended image region. This region is,

due to the preattentive object segmentation, centered at object with highest salience instead of simply at the most salient point in the image. Another important aspect introduced in the Jeeves saliency mechanism, is that the activation in the 'master map' is not solely based on image features, but also integrates a *position salience* map and a map for *inhibition of return*. These maps form the point where visual routines and task guidance operate. Consider for example the task of looking for "a blue block under a red block". After the red block has been found, a location under the red block wil be more salient due to activation on the position salience map, whereas the red object will become less salient through the inhibition of return map. This way, the visual routines can guide the search for objects in the image.

2.3.2 Task-driven visual routines

Although the work presented in the previous section provides insights how visual routines might use and manipulate different types of base representations extracted from the image, it largely ignores an important implication of the framework; that routines may be combined and scheduled in different ways to solve a variety of tasks. This is mainly because they were designed to solve a single task only. This section will discuss work on visual routines that addresses this more cognitive aspect of the visual routines framework, which would be ascribed to the visual routines processor: The selection and organization of the visual routines, based on the task requirements.

A significant part of the Jeeves system that has not yet been discussed targets this aspect of visual routines. The way Jeeves extracts instructions from task demands is by translating the task into a logical query that the visual routines processor will attempt to prove. The shape of the logical query will then determine the form of the visual routien. For example, if the task is to find a blue block on a red block, the system will look to satisfy the logical Horn-clause "$blue(x), above(x, y), red(y)$" by first attending a blue object by making it salient, then assigning position salience to objects below it, and finally check whether the attended object is red. If the query was formulated differently, say "$red(x), above(y, x), blue(y)$", Jeeves would have looked for the red object first.

Jeeves is flexible in the sense that it can answer a large number of queries by a number of strategies, combining different predicates in different ways. The way these predicates are *scheduled* by the task is however not addressed. Most other work on visual routines influenced by task instructions formulates the task interaction as just that – a scheduling problem. The approach in this work is that the system is involved in tasks composed of different behaviors, which make use of different visual routines. The visual routines here are recruited by the current behavior, and their results are used to provide feedback to the behavior. Examples are detecting traffic lights in a car driving system to signal that the car should stop (Salgian & Ballard,

1998) or estimating the distance to obstacles when walking on a sidewalk in order to avoid them (Sprague & Ballard, 2001).

This proposed closed-loop interaction between task-demands and visual routines plays an important role in a more recent formulation of the concept of active vision. According to this view, the visual system is always embedded in a task, which guides not only its visual processing strategies, but also motor processes. For example, in the process of opening a door, visual processing is not only aided by directing the gaze to the doorknob in order to foveate it, but the motor state of the eyes also provides a 'deictic pointer' to where the doorknob is (Ballard, Hayhoe, Pook, & Rao, 1997). The task at hand is then solved by 'visuomotor routines', which use the base representations from early vision, eye saccades and probably other motor actions such as movements of the head, and similar deictic pointers maintained in working memory. The argument for this view is twofold. First, it is argued that visual information is limited, and therefore can greatly benefit from embedding motor information similar to Brunnström et al. (1996) but also task knowledge. Aside from presenting a large body of theoretical support for this argument, Yi and Ballard (2009) illustrate this claim by designing a behavior analysis system. This system, which classifies the different subtasks in making a sandwich, manages to achieve up to perfect classification accuracy, using only very sparse visual information, but including information on how the hands move and the timing of the behavior with respect to the overall task. Second, Ballard et al. build on the idea that task demands have significant control over attention and gaze shifts, as opposed to the more popular approach where these processes are largely governed by properties in the image (as with the presented saliency map systems) (Ballard & Hayhoe, 2009).

This view on active vision has not yet gained wide acceptance, probably because it indicates a level of complexity of the interactions between the visual system and higher order processes which is largely ignored in purely visual models. If these processes can influence the way low-level information interacts, in the visual domain as well as in the sensorimotor domain and working memory, it may require a communication and control scheme within visual routines that stretch beyond the scope of the initial framework. A complete theory on visual routines will have to address this interaction.

2.3.3 Constucting visual routines

Aside from the interaction with visual input, and the task-dependent application of visual routines, an important aspect of the framework is the flexibility in their composition. Ullman proposed that in order to solve a variety of tasks, elemental operations can be combined in different ways to suit the task at hand. This section discusses work on how these combinations are acquired through different learning algorithms.

McCallum (1996) focuses on the sequentiality of visual routines and discusses how these sequences can be learned through reinforcement learning. He presents a simple virtual environment where an agent drives on a road with multiple lanes. The agent can perform one of five actions, four of which are visual routines directing the gaze to a certain lane or backwards, and one action that switches lanes. Gazing in a certain direction provides the system with sensory information such as whether another car is at the gaze point, or whether it's looming or approaching. The sequence of actions is learned via the U-Tree algorithm, that for each decision combines the sensory information together with state information from history. By assigning reward to making clear progress without obstructing other cars, the agent learns how to direct its gaze shifts to obtain the information it needs.

Although the U-Tree approach will learn the appropriate sequence of operations in different environments, it does not model the proposed idea of synthetic visual routines, where different operations are combined into an explicit routine that suits the task. Composing an algorithm or routine for a task by correctly combining a predefined set of operations is the main focus of Genetic Programming techniques, which have been applied to compose visual routines as well (Johnson, Maes, & Darrell, 1994). In this work, a visual system is presented with a binary image representing the contour of a human being. The system is then required to locate the left or the right hand, or the head. Genetic Programming solves this problem by combining operators such as detecting outmost locations and edges, placing markers and scanning along lines between markers. The elemental operations provide distinguishable incremental representations that are then used by later operations within the resulting visual routine.

One issue with both the Genetic Programming and Reinforcement Learning approaches is that the construction of visual routines will always be defined by the set of available elemental operators, which are manually predefined beforehand. Johnson et al. acknowledges this problem, and attempts to resolve this issue by focusing on elemental operations performing very low-level subtasks that are very abstract with respect to the eventual (hand-finding) task. Still, the final result will always be limited or biased by what was initially defined by the programmer. In order to overcome this, a reinterpretation of visual routines might be needed, either by redefining to which extent the elemental operations are predefined, or by redefining what actually constitutes a visual routine. An approach of the second type is proposed by Rao (1998). In this work, elemental operations are divided into two classes; those that move the focus of attention (FOA) by saccades or tracing, and those that establish primitive properties at the FOA. In solving a visual task, an 'attentional state' is composed of the set of properties at the FOA, the history of visuospatial relations between previous FOA's and the current, and the properties at previous FOA's. The proposed alternative approach to visual routines is then, that they are not preprogrammed

sequences of operations, but learned patterns within the attentional state space. When faced with a task, it is solved by executing a prototypical sequence of attentional shifts and extracting the information needed.

2.4 Summary

The history and the influence of the theory of visual routines illustrates several important aspects to the study of vision, and the way a model for visual cognition could contribute. Marr's theory illustrates the strengths of his computational approach to vision: by detailing what the system should be capable of and assigning a computational goal, one can derive a theory where experimental findings fit in how this goal is achieved. Ullman used a very similar approach, by expanding the goal of vision by considering a type of tasks that require operating on the result of feedforward visual processing described by Marr. A thorough analysis of the requirements to solve such problems has led to a framework that has been very influential to the study of visual cognition, and led to the design of various computer vision systems. The variety in these models illustrates the broad scope of the framework, and also illustrates new aspects of vision that require understanding for a complete theory of vision, including the interactions between the visual system and other brain mechanisms.

Despite this variability in these models, the interpretation of attention is relatively consistent in all of them. Attention is considered a selective mechanism that delimits detailed analysis of only a limited area of the image. The underlying mechanisms of attention are sometimes characterized by a saliency map mechanism that is controlled by the visual routines, or the routines have immediate control over the shifts of this attentional spotlight. Similarly, visual processing of the input image is consistently characterized as feature extraction throughout these systems. Since both these mechanisms play such a significant role in all these systems, it may seems crucial to the theory of visual routines that these interpretations still hold within current experimental findings. However, the 25 years since the theory was designed have seen a large body of work on vision that has greatly altered the view on visual processing and attention. The next chapter will therefore attempt to characterize the modern view on these aspects of visual processing, after which it will be discussed how these changes affect the theory of visual routines.

Chapter 3

Vision and attention

An important notion underlying the theories of vision and attention described in the previous section is that vision was seen as a two-stage process with relatively clear input/output relations. The base representation in the visual routines theory captures the first stage, a detection process that solely depends on the input image and which constructs a representation of present features. The second stage reflects interactions with this sensory representation and more higher-order processes. If any consensus is to be found in the vast literature on visual attention that has been produced over the years since, it will be that this division is not that clear-cut. Visual attention is not just a bottleneck mechanism that selects features from a preattentive stage for further processing, but it plays a modulatory role throughout the entire visual processing pathway, and thus manifests itself in many ways. These manifestations have led to a reinterpretations of what the function of visual attention is, where in the brain attention operates, and the timecourse of these processses. The following section will describe some of these reinterpretations and some of the observations that support them. First, however, some important changes in our current understanding of the visual system will be discussed.

3.1 The visual system

As indicated, the theories of vision presented above all rely on the assumption that vision is a two-stage process composed of measurement followed by higher order operations. The stage of measurement is based on the idea that the neurons in the visual system act as detectors tuned to very specific features. This idea of neurons as feature detectors has been dominant in the study of vision since the pioneering studies of Hubel and Wiesel (1962). They studied responses of V1 neurons in cats, and discovered very specific neuronal tuning reflecting orientation selectivity within small areas of the visual field which they called the neuron's receptive field (RF). Similar re-

sponses tuned for features such as motion and color have been found in V1, as well as neurons tuned to more complex stimuli in higher order visual areas, such as for complex objects and shapes in the inferotemporal cortex (IT) (Desimone, Albright, Gross, & Bruce, 1984). This has led to the view of neurons as detectors, signaling the presence of certain features, thereby constructing the base representation. However, Hubel and Wiesel also introduced another influential idea by dissociating two types of cells in V1: 'simple' cells that simply respond to the presence of simple features within small RFs, and 'complex' cells that integrated the output of these simple cells to represent more complex features within greater RFs. This hierarchical construct where different types of information are integrated to represent convoluted information has presented itself at many levels of the visual pathway (Orban, 2008), indicating a hierarchical organization of the visual system, where the sequential integration of information allows for efficient construction of detectors for these complex features. In fact, if detection of these complex features at all scales is to be computed in parallel, it would induce a combinatorial explosion of 'grandmother'-feature detectors for which the brain does not have the capacity (Tsotsos, 2011).

One of the implications of a hierarchical organization of information is that at higher levels the size of the RFs increases. This means that at the top levels of the hierarchy, the image representation has the coarsest resolution. This may seem like an inevitable weakness of the hierarchy, but an example by Orban (2008) indicates how this increase in RF might actually be necessary to extract reliable information. He describes how, in detecting the motion of an object, the small receptive fields of V1 cells induce an aperture problem (figure 3.1): they merely signal the presence of one certain motion at one specific point on the object, but this signal may be ambiguous (consider looking at a moving, slanted line through a very small hole for comparison. The actual motion of the line can not be concluded from the observed motion). In MT, the integration of information from the edges of the object into a larger receptive field allows to capture the motion of the object as a whole. Based on these principles, integration of information in the hierarchy in a coarser representation may be necessary to allow neurons to signal new types of feature information.

Also, it must be noted that a hierarchical organization of the visual system does not imply that all information is organized as one single single pyramid and integrated into a single top-level representation. The notion of a single pyramid for example, contradicts the well-established notion that the visual pathway is divided into two different streams: the ventral pathway that leads from V1 through V2 and V4 to Inferotemporal area IT, and the dorsal pathway leading through V1, V3, MT and eventually to parietal cortex. The coarse functional dissociation between the two pathways has for a long time been that they process different visual attributes, the ventral route being largely involved with the computation of identification, and the

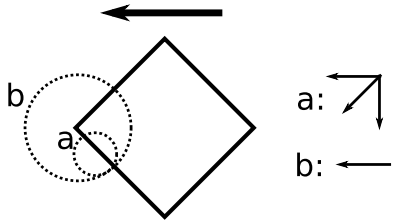


Figure 3.1: Illustration of the aperture problem. The diamond object moves in the direction indicated by the large arrow. Neurons (e.g. from V1) with a small RF would register motion in all directions ranging from leftwards to downwards, whereas neurons with larger receptive fields such as b (e.g. from V2) will be able to isolate the correct motion.

dorsal route with location (Haxby et al., 1991). Although recent findings indicate that there might not be such a clear-cut functional dissociation [1], this dissociation illustrates how a hierarchical organization does not immediately imply a single 'end result' of vision. Instead, multiple hierarchies may be identified within the structure, which would lead to different streams of information integration of different types.

Up to this point, the information flow through these hierarchies has been described as a unidirectional process: from the retina to V1, through either pathway up to IT or Parietal Cortex, integrating information at each level of the hierarchy. This conforms with a two-stage model of vision, as it implies that after a single feedforward sweep through the visual system, all information is available to be selected by attention for further processing. Although the feedforward activation of all levels helps to establish useful representations of the scene, a significant fraction of the information flow appears to originate from feedback and horizontal connections between different layers. These connections account for a variety of findings, some of which will be discussed here.

After activation by the feedforward sweep of activation, neurons remain active, affecting others via horizontal and feedback connections. The RFs of these neurons will likely not fully overlap, or in the case of feedback projections, the projecting neuron will likely have a larger RF. This causes the affected neuron to be reactive to stimuli outside its original RF. Therefore, the RFs described by Hubel and Wiesel (1962) might be better described as 'classical' receptive fields (cRF) emphasizing the difference with their larger counterpart at later processing times, which will be referred to as 'modern' receptive field (mRF) (Roelfsema, Lamme, & Spekreijse, 2000). As with feedforward integration of information, this is not limited to the

[1] there is increasing evidence for interactions between these two pathways. For example, attributes such as two- and threedimensional shape appear to be processed in both pathways, regardless of the eventual response, (Orban, 2008), and there is fast feedback from the dorsal pathway affecting object recognition in the ventral pathway (Bullier, 2001)

spatial domain, which means that not only the receptive field of a neuron might change after prolonged activation, but also its feature selectivity. For example, neurons in the Fusiform Face Area in IT might initially be globally tuned to respond to the presence of a face in the visual field, but might after an additional delay signal more selective information about the facial expression or identity. This happens even when the facial expression was not relevant for the task (Sugase, Yamane, Ueno, & Kawano, 1999).

Another function of feedback connections is to generate suppressive surrounds, a phenomenon that has been observed at various levels of the visual system. Suppressive surround mechanisms were first proposed by Tsotsos (1990) as a mechanism of attentional selection. Roelfsema et al. (2000) describes surround effects as *feedforward* inhibitory effects, but already at the level of V1, stimuli at the immediate surround of the cRF can inhibit a neurons firing rate. The fact that these effects occur in this early layer, make it unlikely that these are the result of feedforward interactions. At higher levels of the pathway, the integration of multiple features makes these surround interactions more complex. In MT for example, the surround for a moving stimulus is maximally suppressive when both speed and direction match the inhibited neuron's cRF. Conforming to the size of the cRF, the radius of suppressive surrounds at higher levels is also larger. Next to simple feature inhibition, the inhibitory surround effects in higher layers can also be used to construct new feature representations. For example, a speed-dependent surround inhibition can allow the neuron to compute a spatial derivative of speed, thus constructing a representation for acceleration (Martinez-Trujillo et al., 2005).

There is a close correlation between the effects from feedback- and horizontal connections described here and several effects of task-guided visual attention. Although the change in neuronal selectivity observed by Sugase et al. was evoked even without any relevant task instructions, it does follow the same pattern as proposed in classical models such as visual routines theory: after a certain delay following the initial response, the neuron is involved in computation of more detailed information that is not immediately available in the single feedforward pass. It seems plausible that the evaluation of facial expression or identity is evoked automatically, but that a task-guided and less automated search for information could use similar mechanisms, including feedback tuning. Similarly, surround suppression is a mechanism that not only manifests itself interneuronally in stimulus-driven data processing, but is also observed after attentional selection of an item in the visual field. The attentional modulation that induces surround suppression has been found in MEG- and physiological studies (Hopf et al., 2006; Boehler, Tsotsos, Schoenfeld, Heinze, & Hopf, 2009). Reynolds and Chelazzi (2004) also relate surround suppression to attentional modulation and describe a mechanism where attention enhances the firing rate of a neuron representing a selected feature, and thereby enhancing its inhibitory effect

on neurons representing the surround. Unlike Boehler et al. (2009), who explicitely relate attentional surround suppression to feedback modulations, this mechanism focuses on lateral inhibition. One of the shortcomings of the Reynolds and Chelazzi approach is that there is no description of how the 'winning' unit is selected and how its firing rate is enhanced by attention.

This section has introduced two key structural revisions to the classical two-stage model of vision, that mainly target the definition of the first stage. The hierarchical organization of the visual system, integrating information at higher levels in the processing stream at increasingly coarser resolution, and recurrent and horizontal connections modulating responses at lower levels, indicate that the description of a simple feedforward 'measurement stage' does not suffice. The interaction effects induced by feedback connections seem to play important roles in attentional modulations, which argues that a separate second stage of attentional selection can not be supported from both temporal and structural viewpoints, and it seems that a much more integrated view is needed. This concordance between attentional modulation and the effects of feedback connections has indeed been the reason for some to define attention as merely the result of these feedback connections (Lamme, 2005). The next section will indicate that a satisfactory definition of attention requires more elaboration than that. First, the functional role of attention will be discussed, followed by reconsiderations regarding where and when in the visual pathway attentional mechanisms operate.

3.2 Functional reconsiderations of attention

As Marr's approach to the study of vision illustrates, when studying a complex system or mechanism it may be worthwhile to start by introducing its computational goal. This section will argue that the underlying goal of attention, its functional role, is to implement various forms of information reduction. This goal for attentional mechanisms is illustrated by close inspection of two prototypical tasks for vision, visual search and recognition. For these tasks, it is argued that information reduction is an adequate description of the functional role of attention in these tasks. Furthermore, several other mechanisms that have been associated with attention – binding, working memory and consciousness – will be addressed, and their relation to visual attention as an information reduction mechanism will be discussed.

Tsotsos (2011) overviews various mechanisms and phenomena that have been attributed to visual attention. A general assumption that underlies all these mechanisms, is that the processing capacity of the brain is insufficient to completely process all the information that is continuously presented to our visual systems and reliably extract the necessary information from it at real time. Therefore, we need these mechanisms to reduce the informa-

tion load, and extract relevant samples to act upon. Defining the functional role of visual attention as information reduction, would imply that all attentional mechanisms ultimately serve this goal. This might seem counter-intuitive considering that attention is generally associated with facilitated processing of attended stimuli, which one might be inclined to associate with mechanisms that facilitate relevant information rather than reduce irrelevant information. Facilitation effects however, can be expressed in terms of information reduction. This can be illustrated by looking at the effects of the attentional mechanisms on the firing rates of individual neurons. When irrelevant items are suppressed, the signal-to-noise-ratio of neurons tuned to the attended stimulus increases, which would result in facilitated processing of this stimulus.

Information reduction as the computational goal of attentional mechanisms still is a very broad definition. In the same overview, Tsotsos introduced a coarse taxonomy, based on three means to achieve this goal:

- *Selection* – Selection processes are a significant topic in most studies of attention. As illustrated by the systems in the previous chapter, any attentional system, with the goal of reducing information, faces the problem of what to include in the selection. In the human visual system selection is seen in by both overt attentional shifts (e.g. saccades to points of interest) and covert shifts (those without gaze changes).

- *Restriction* – Whereas selection mechanisms choose one of many options, restriction mechanisms choose several out of many options. That is, when faced with many choices but with insufficient information to enable selection of the best choice, restriction permits the selection of the best few options to be further examined. Priming and negative priming are examples of the effects of these mechanisms; the visual processing is restricted which enhances processing of stimuli of the preferred type but can also cause unpreferred or purposefully ignored stimuli to be completely neglected.

- *Suppression* – The requirement to limit processing to the attentional selection illustrates the need for mechanisms to suppress information outside of this selection, in order to prevent these stimuli from interfering with the selection. Examples are spatial- or feature surround inhibition, which suppresses the processing of items close to a currently selected item, and inhibition of return, which is the suppression of items that have been selected before the currently selected item.

Although this taxonomy is important to illustrate what is meant by information reduction, it is still not a very satisfactory answer to the question what attention does, as these descriptions lack specificity by detailing the amount

or type of information reduction that would suffice, and how this can be realized while still making it feasible to complete visual tasks. (Rothenstein & Tsotsos, 2008). This issue shows that answering this question thus involves the goal of vision altogether, which is an issue of debate in itself. Although Marr and Nishihara (1978) proposed the object-centered 3D - representation as an end-product, this view has been heavily challenged throughout the 1990's as evidence accumulated that favored viewpoint-dependent representations in the brain (Peissig & Tarr, 2007). This has introduced many questions regarding the gray area between representations in the visual system and more abstract concepts stored in memory, and therefore, a clear computational goal for vision remains undefined and might even be considered nonexistent.

To illustrate the functional role of attention however, addressing the full computational goal of vision may not be necessary. Instead, it might be worthwile to consider a subtask solved by the visual system which involves attention and can be considered 'prototypical' for visual tasks. Visual search, in its most general form would match these criteria. A computational goal for this task can be defined as finding a subset of an image that matches a target that may be defined sparsely by a set of features (e.g. find a red horizontal bar). It can be proven that a purely data-driven approach to visual search is NP-complete and intractable (Tsotsos, 1990), and therefore it's unlikely the brain solves visual search that way. Instead, it is necessary to limit the visual search problem which can be done by, for example, introducing task-guidance in the search process. This is information reduction by attentional mechansims. Within the context of visual search, the following types of information reduction can be isolated (Tsotsos et al., 1995):

- *Region of interest selection.* Clearly, visual search needs to operate through the dimension of space. An efficient way of information reduction would thus be to bound the region where a match is to be found. Attention must therefore select this region.

- *Feature of interest selection.* Single features will only sufficiently direct visual search if these features are exactly that which discriminates the target from the distractors, that is when they characterize the pop-out singletons (Treisman & Gelade, 1980). Often, this will not be the case and targets are defined by a specific combination of multiple features. This would lead to a combinatorial explosion if the feature space isn't restricted. Therefore, attention is needed to select relevant features to be used in the search.

- *Control of Information flow.* The hierarchical structure of the visual system equipped with feedforward and feedback connections, allows the information flow to spread widely. If information were to flow through this network uncontrolled it would result in interference, blur-

ring and loss of resolution, which would cause the loss of the unique representation each unique image gives. Attention is thus associated with flow control, by reducing irrelevant or interfering information spreading.

- *Shifts of selection in time.* There is no guarantee that the first selection of region and features will actually match the target. Therefore, the system must be able to shift attention to process other locations, and maybe even other features. This doesn't only introduce the problem of how to let go of the established representation and construct a new one, but also what in particular determines the next region- and feature selection criteria, thereby determining the order of the search process.

- *Balance of task- and data-directed processes.* A visual search task is usually influenced by both a task directive (for example instructions on the properties of the target) and probably just as much by the data itself (which contains the target to be found). Related to the issue of flow control, attentional mechanisms needs thus to be able to balance the information flow from both ends to provide adequate guidance of the search process.

This analysis provides a computational argument as to why our visual system would be equipped with attention. Importantly, it indicates how attentional mechanisms do not seem to solely reduce search space, but also serve as a set of tuning mechanisms that guide the flow of information to represent the selected stimuli and construct its representation. From the difference in subprocesses indicated in this analysis it can also be seen that the influences of the attentional mechanisms are likely to be found throughout the entire visual system, and how visual attention and vision are therefore greatly intertwined processes. This is reflected in the variety of tasks where attention is believed to play a key part. Here, we will discuss the role of attention in another visual task that seems to highly contrast with visual search: visual recognition. Importantly, the role of attention in that task can be expressed as similar information tuning mechanisms as well. Then, three other visual phenoma that have been associated with attention will be discussed.

Visual Recognition is a visual task that seems to contrast with the proposed 'prototypical' task of visual search. Whereas visual search can be described as the locating the features of one model of the target, recognition deals with the identification of one of many possible targets, usually in a cluttered scene with multiple distractor objects (Macmillan & Creelman, 2005). Although recognition in natural scenes may seem like a practically unbounded task, experiments in humans (Thorpe et al., 1996) and macaque monkeys (Hung, Kreiman, Poggio, & DiCarlo, 2005) have shown that the

brain already signals robust and accurate identification responses to a visual stimulus after only 100 to 160ms. This indicates that a single feedforward pass through the visual system might suffice for recognition tasks, which might not be surprising given the highly specific tuning of certain neurons and groups of neurons in area IT that are immediately activated by the data-driven information stream (Peissig & Tarr, 2007). Indeed, feedforward models that extend Hubel and Wiesel's idea of simple and complex cells seem to account for the behavioral data found in rapid categorization tasks (Serre, Oliva, & Poggio, 2007).

May seem fast, but these don't capture the full breadth of visual recognition: - scene may be cluttered - localization of the target or feature configuration may be required - inclusion of more detailed features may be required. - brings the discussion close to 'binding'

Although these studies seem to indicate a highly automated and fast appearance of certain recognition tasks, they do not capture the full breadth of visual recognition tasks, and not all tasks can be resolved by simple feedforward feature detection. In various other recognition tasks significant roles for attention can be identified (Tsotsos, 2011; Tsotsos et al., 1995). First of all, the studies in the previous section only seem to target identification tasks involving one object in the scene. As was already indicated, these tasks may involved cluttered scenes, where feedforward detection of features that do not belong to the target object may disrupt the identification. Also, localization of the target object may be required in various tasks. Again, simple feature detection can not account for these tasks, as feedforward activation does not provide spatial detail. This may be a problem when recognition depends on the spatial configuration of the features within an object. Finally, these requirements may not be limited to the spatial domain, and detailed feature information may be needed to provide detailed categorization or the target object. To meet all these requirements attention can play a significant role. All these tasks require the inclusion of only the relevant features in the recognition process, which can be translated to the mechanisms of suppression, selection and restriction as outlined above. However, it is also important to note, that in these tasks, recognition is also related to binding the features that belong to a single object into one representation.

The binding problem poses the question how the distributed signaling from neurons, serving as simple feature detectors, can ultimately be combined to establish a unified representation of an object in a scene. How are different features of an object connected to form a single representation? (Malsburg, 1994) Feature integration theory (Treisman & Gelade, 1980) and the early saliency map model of attention (Koch & Ullman, 1985) address exactly that problem by recruiting selective attention. By aligning the different feature maps before integrating them into a single master map, the model can retrieve the feature values at the attended location. Selective attention then extracts these feature values to construct a representation for

higher order processing. Although appealing, this mechanism has important drawbacks. First, the mechanism assumes a spatial organization of the feature maps and master map, implying that spatial coherence is the single binding feature. However, already since the formulation of the Gestalt laws it is known that space is only one of the features that can constitute a unified representation (Koffka, 1999). Introducing other master maps integrating information based on other feature dimensions does not provide a solution, but would solely shift the problem from the integration of simple feature maps to the integration of the different master maps. Another important issue is that objects do not consist of single points in retinal space, and the integration of the different spatial points that belong to one object is again an essential part of the binding problem.

One of the difficulties these models face is that a feedforward pass is their only modus of integration. The introduction of horizontal and recurrent connections in the visual system allow for a different approach to solve binding issues, as proposed by Roelfsema et al. (2000). They propose that the initial feedforward pass through the visual system activates neurons that are organized in a sheet by means of horizontal connections. Once units are activated these form a so-called 'interaction skeleton'. An attentional label then spreads among interconnected neurons to form a unified labeled representation of the object. This spread of activation does not have to be restricted to the spatial dimension, and might also spread to different feature dimensions via a 'linking dimension'. This allows for a labeled representation of all the features that comprise a single object. The relation to the information tuning mechanisms is then that the binding problem is approached by limiting the information flow to those features that are included in the selection, and suppressing those that are not.

There are also aspects to vision where the involvement of visual attention is much more debated. One example is the complex concept of visual awareness or consciousness. It may seem natural to assume that visual attention results in a selection of visual stimuli for future processing, and that that would automatically imply conscious processing of these stimuli. This is illustrated by the famous definition of attention by James (1890), stating that "everyone knows what attention is (..) it is taking posession of the mind". This has led to many studies addressing consciousness and attention as different labels of the same process (O'Regan & Noë, 2001). Conversely, Lamme (2005) stresses that this relation is not necessary. If attention is viewed solely as a collection of mechanisms that modulate the processing of sensory input, then it would indeed result in some inputs having a higher chance at influencing the higher order processes such as descision making or memorizing, but it does not necessarily imply conscious processing. Instead, only when recurrent processing passes through the entire visual pathway, one has a conscious experience of the visual stimulus. This is seen in experiments with TMS and lesion studies, where impairments of the possibilities for re-

current processing cause stimulus inawareness, yet still can guide action and show effects of visual attention.

Another debated function of visual attention is its involvement in working memory and the relation between these two concepts. One popular interpretation is that attention acts as a 'gatekeeper' for working memory (Awh, Vogel, & Oh, 2006), where attentional selection determines which features and objects will enter working memory. This is the approach that is closest to the function of attention proposed by (Koch & Ullman, 1985), and stands closest to the role of working memory proposed by some systems in section 2.3. However, several findings have indicated that storage of working memory items may actually occur in the visual system using the feature detector units (for a convincing example, see Harrison & Tong, 2009). This would illustrate a different function for attention in working memory tasks, because it may have to avert interference between the bottom-up information flow and working memory items that use the same units. A third possible function for attention is proposed by Awh and Jonides (2001), who illustrate that a representation in visual working memory might actually be a manifestation of the attentional mechanisms themselves: for example, maintenance of multiple locations in visuospatial working memory could be realized by rapid shifts of attention directed at these locations. This view is supported by findings that forcing attention to a location during visuospatial working memory maintenance seems to cause interference. A fourth proposed role of interaction between attention and working memory suggests how working memory is actually involved in storing targets or prototypes that are used to direct visual attention, for example in visual search (Olivers, Peters, Houtkamp, & Roelfsema, 2011).

Although these four approaches illustrate very different interpretations of the relation and interaction between attention and working memory, it must be emphasized that they are not mutually exclusive; one can envision a combination of these interpretations, and section 4.4 will present a more elaborate discussion on the relation between working memory and attention, and will present an integrated view of these approaches. For this section it is important to note that these approaches together suggest widespread involvement of attentional tuning mechanisms in working memory tasks (and vice versa).

Similarly, an extensive review of the complete debates on the relations between attention and awareness stretch beyond the scope of this thesis, but the significant involvement of attention in these processes, as well as in visual search, recognition and binding is clear. From these examples we should deduce that attention can manifest itself in a wide variety of ways, thus indicating a very widespread functionality. This makes it difficult to obtain a conclusive definition of the functional role of attention. If a definition is to be found, it should be based on the overlap between the ascribed role of attention in these processes. One point of overlap is discussed in the begin-

ning of this section, which describes how mechanisms of attention all seem to illustrate one of three forms of information reduction, but the discussion of the visual tasks where attention plays a role illustrates the functional role of these mechanisms: attention acts as a set of information tuning mechanisms that are recruited so that at least the minimum requirements of a goal can be achieved (cf. Tsotsos, 2011)

3.3 Mechanisms of attention

A functional analysis for visual attention is important as it provides context to both models of attention and observations of attentional mechanisms. This section will discuss several mechanisms of attention, as new findings in attention research have also provided new insights in the way these mechanisms operate throughout the visual system. The previous sections on our current understanding of vision and the functional role of visual attention already indicated that the view of attention as a post-hoc operator selecting relevant features from an initial stage of measurement is infeasible: the recurrent nature of vision, exposed in the difference between cRF's and mRF's, as well as the proposed broad functional role of the attentional mechanisms, imply a more widespread impact of attention throughout the whole visual system. This section will address the view on when and where within the visual system attentional mechanisms operate.

The main structural dissociation of attentional mechanisms seems to be that of top-down versus bottom-up attentional processes. It is a dissociation that has issued long lasting debates between their different roles in information tuning, and which of these two best describe the nature of attentional mechanisms. However, from their definition it may be inferred their very nature indicates how attentional mechanisms need to be involved in both the processing of the data from the image and the task directives which define the relevance of every feature. The interplay of these two processes is one of the core unknowns in the study of attention.

The shift of the view of attention from primarily a data-driven process to a more top down process is well reflected in the evolution of saliency map models, where it was already proposed that data-driven 'preattentive' processing did more than the simple computation of features and feature contrasts. Data from visual search experiments with attentional capture by irrelevant distractors seem to agree with this bottom-up approach. The results of such experiments seem to challenge the idea that top-down selection of a relevant feature dimension, or even a single relevant feature, can completely restrict processing to that feature only; effects from an irrelevant but salient distractors can likely not be avoided, even when the exact target is known beforehand, and after extensive practice (Theeuwes, 1992). Regardless of the top-down imposed goals, attention seems to be intitially

captured by the most salient item.

As these and related findings marked the importance of data-driven information in visual processing, experimental psychology attempted to identify the conditions necessary for these irrelevant singletons to capture attention (Yantis, 1993), and gain deeper understanding in what makes a stimulus salient. As the concept of saliency has such a clear relation to the image itself, the topic provided an accessible approach to the study of attention in computational research. This is probably why saliency lies at the heart of many computational models of attention (Itti & Koch, 2001). Much like in the original saliency map model, they are based around a master map created from competition within or among feature maps that have passed through several filters to compute the conspicuity, based on the idea that saliency is defined by points of maximum contrast. This idea is probably best illustrated in the model of Itti, Koch, and Niebur (1998), where feature maps for orientation, intensity and color are constructed and convolved with Gabor filters of different scales. These filters compute points of maximal contrast at these different scales, which identify globally salient points of interest.

A different approach to the construct of saliency is described by Bruce and Tsotsos (2006). They postulate an entropy-based definition that saliency is defined by points and areas that provide maximum information. The idea is that a complete pointwise description of a scene will usually have a significant amount of redundancy because of the values of points that can be predicted by the neighboring area, like for example on a uniformly colored wall. Maximum information, and therefore saliency, occurs at points that are not easily predicted from their surroundings. Note that this approach encompasses the idea of contrast as a good marker of salient points, but that this definition is richer. In the AIM model (Attention by Information Maximization) this idea is applied by defining a set of learned independent components that best describe the variance in a database of scenes, and using these to compute the self-information of each point in the image.

Like all saliency models, AIM provides a way to transform the image data to a saliency map. Usually it is assumed that the filtering properties of neuronal tuning provide a way to implement this. For the idea of saliency defined by the maximum response of contrasting stimuli in various feature maps this may be clear, but the implementation of the trained independent components described in AIM might seem less transparent. A learned basis for points interest based on information maximization may be more feasible if neuronal plasticity in the visual system is considered. Indeed, it has been proposed, that saliency is not an active filtering process, but simply the manifestation of long-term memory enhancements in the visual pathway, that diverts the information flow based on our experiences with natural scenes and an acquired bias for certain points that we consider interesting (Lamme, 2005). These long-term memory enhancements could also account for the

phenomenon of 'novelty bias' (Desimone & Duncan, 1995), an attentional capture effect measured on the presentation of unfamiliar stimuli, also called 'temporal saliency'.

The above discussion addresses the strength of salient objects and attentional capture, and seems to suggest that salient items in a scene will automatically attract attention. However, there seems to be an important asymmetry in the studies on salient stimuli. Whereas expected but known to be irrelevant stimuli can still automatically capture attention, the opposite class of stimuli, those that are unexpected but potentially relevant, might fail to capture attention. These stimuli form the basis for the study of inattentional blindness (Simons, 2000). The nature of these stimuli is, however, heavily debated. Simons points out that although in these studies subjects often report not having seen the blinded stimulus, some behavioral data points out that they do act upon it. This shifts the debate from an information selection perspective to an issue of consciousness, and it fits well in the scheme proposed by Lamme (2005) that was discussed in the previous section. Even more distant from the study of attention, and more close to the nature of consciousness, is the proposed explanation that these stimuli might be consciously processed, yet immediately forgotten. Either way, these effects indicate that the hypothesis that attention is constituted purely by bottom-up mechanisms is not supported, and it emphasizes the role of task-guidance and the current goals of the observer. Top-down processes are assumed to play an important role in attentional selection, both voluntarily and involuntarily. The term 'top-down' itself might be misleading, as it implies that there has to be a definite end-result of visual processing, like the object-centered representation proposed by Marr (1982), but the term is usually only intended to mark the dissociation between the data-driven mechanisms, and those generated by endogenous signals. There doesn't seem to be one neural substrate that can account for all forms of top-down attention (Frith, 2005). The following discussion will indicate how this is not surprising, as the nature and implementation of different types of top-down selection is very different.

Top-down voluntary attention that would provide an explaination to the effects of inattentional blindness is a type of selection that is already established before the scene is viewed. The main idea is that the current state or task will impose a selection of relevant features or regions that are to be abstracted from the scene, while others are suppressed. The most common form in psychological experiments is task guidance. An involuntary variant is priming, where the presentation of a cue can facilitate processing of a subsequent stimulus of the same type or location, or interfere when the relevant stimulus does not match the cue. Since this type of attentional filtering is established before the processing of the image, it can establish early enhancements in processing of the relevant stimuli (Tsotsos, 2011).

Top-down attentional mechanisms that operate upon the result of the

bottom-up datastream after the scene is processed are more similar to the idea of attention introduced in the previous chapter, and these mechanisms are likely to implement the role of visual routines. The visual routines described in Ullman's original paper (1984), for example for tracing a curve or for marking points of interest, are likely voluntary forms of these selection mechanisms. Convincing evidence that these processes are not established by parallel, bottom-up processes, comes from neuroscientific studies that indicate that the effects of different subtasks of a single visual task, can be measured sequentially (Roelfsema, Khayat, & Spekreijse, 2003). This does not imply that execution of these attentional mechanisms is always voluntarily or consciously planned. Instead, these mechanisms can be executed in an automatic or ballistic manner. This lies at the heart of Rao (1998)'s theory of visual routines as learned patterns of attentional shifts, as well as Cavanagh, Labianca, and Thornton (2001)'s theory of sprites, which describes the ability of visual routines to analyze a motion by means of learned and chunked motion patterns. This can be done much faster (200ms) compared to the analysis of unfamiliar motion patterns for which no such chunk might exist (sometimes over one second).

Aside from attentional mechanisms that bias the processing of the scene before it has been presented, and those that operate upon the data extracted from the scene, there is another top-down mechanism that can have long-lasting effects after the required information is obtained, namely inhibition of return (IOR). This mechanism inhibits processing of locations and objects that have recently been attended, and discourages saccades to them. This is a type of top-down selection that is believed to counteract the effects of attentional capture and saliency from bottom-up selection mechanisms: without IOR, the most salient locations in a scene would probably dominate visual tasks, even when they are task-irrelevant. (Klein, 2000). In the context of visual routines, it seems important to note that this mechanism offers the possibilities of a tagging system, but this function should be adopted with care, as IOR has a suppressive nature and seems to be largely implemented by the oculomotor system, so it might not be suitable to aid in spatial reasoning tasks addressed by virtual routines.

One might argue that attentional positive and negative priming effects would also fall in the category of selective mechanisms that are triggered after processing of a scene, and that they resemble IOR mechanisms more than involuntary task-guidance mechanisms. In fact, many models of negative priming interpret these mechanisms as such, but although priming mechanisms are triggered by the processing of a scene, they do seem to establish a biased state of the visual system that resembles the effects of the observer's state. A strong argument that these priming effects rely on system state rather than reactively inhibiting mechanisms comes from studies that show very long-term negative priming effects lasting even a month after initial presentation. The proposed explanation is that instead of inhibitory

mechanisms that last for a month, these effects are caused by the triggered memory of initial stimulus presentation, which brings back a previous state where a certain stimulus was to be inhibited. However, these findings do not exclude that active inhibitory mechanisms, resembling those of IOR mechanisms might play a role in priming aside from that, on a much shorter timescale (Tipper, 2001).

These different categories of attentional mechanisms indicate a coarse division of when they operate. To summarize, the order in which different attentional mechanisms operate when percieving a stimulus would be as follows: (1) the current goal or state might impose a preliminary attentional bias towards certain stimuli or features, (2) mechanisms based on saliency determine the initial point of fixation (3) upon scene presentation, the image data passes through the visual system, filtered by the aforementioned bias. (4) top-down processes are triggered by the resulting information flow, either to change behavior accordingly or to achieve richer information extraction. (5) after selection of a location, IOR processes inhibit the attended location, which possibly combined with other reactive inhibitory processes affect the processing of the next scene. This ordering does not impose that these attentional mechanisms have to be executed sequentially. On the contrary, it can be assumed that task or state bias can persist throughout the entire process until the observer might decide to, for example, switch his or her strategy. The effects of top-down selection mechanisms will likely last until the task is solved and the required information has been extracted or constructed. IOR effects are known to last for several seconds, but they are diminished when the scene is removed or drastically changed. The parallel execution of these mechanisms makes it an almost impossible task to attribute any observed effect of attentional selectivity, either behavioral or psychophysical, to a single process. Therefore, attempts to identify the various attentional mechanisms must isolate them temporally and functionally.

In the formulation of vision as a two-stage process, the preattentive stage of vision was proposed to take approximately 150 to 160ms. As indicated, this is currently a relatively well supported estimation of a single pass throug the visual system (Thorpe et al., 1996). This implies that recurrent modulatory effects of attention can not be observed untill after 150ms. This knowledge can be used to make a temporal dissociation between bottom-up and top-down attentive processes. For example, Lamme and Roelfsema (2000) overview the neuronal responses in V1 to different attentional subtasks. They find that the signaling of a cell's preferred orientation, a cell's cRF at a texture boundary and a cell's cRF in a texture figure can be temporally dissociated at 55ms, 80ms and 100ms after stimulus onset respectively, thus indicating that these are the result of bottom-up processes integrating information from an increasing number of cells through horizontal and feedback connection. However, a more task-related attentional effect – in this case whether a cell's cRF lies on an attended visual object (a curve) or

an unattended one – occurs at a much longer latency of 235ms, thus consistent with the idea that these top-down endogenously generated signals occur after 150ms.

In this section, the complexity of attentional mechanism has been demonstrated by a discussion of the various types of attentional mechanisms and their different sources and effects. Due to this complexity, any study of attention should take caution in attributing attentional effects to one of these sources. This might be the reason why most models of attention tend to focus on a limited set of attentional mechanisms only, and why it is hard to compare or integrate different models. Especially computational models, which suffer from the need to balance a significant level of detail versus implementability, require important abstractions and assumptions that may define and dominate the model. The next section will describe the four main types of computational models of visual attention, and discuss the core assumptions these models make.

3.4 Computational models of visual attention

Just as there are numerous different facets to the study of visual attention, there are also numerous computational models that try to capture the mechanisms involved. This makes for a wide variety of models that rely on very different mechanisms and representations which makes them hard to compare. Still, a coarse taxonomy can be derived based on the assumptions and predictions these models make, to derive four main hypotheses of how visual attention is implemented in the brain: Selective routing, Saliency maps, Temporal tagging and Emergent attention (Tsotsos & Rothenstein, 2011). These four hypotheses will be detailed here, illustrated by some of the models that follow these hypotheses.

3.4.1 Selective routing

The selective routing hypothesis considers visual attention as the result of mechanisms that control the information flow. More specifically, most of these models focus on how a selected subset of information from the visual scene gets transmitted to higher order areas, and what problems arise in this process. For example, when assuming a pyramidal hierarchy in the visual system, uncontrolled information flow would lead to issues such as blurring and cross-talk (Tsotsos et al., 1995). Several different architectures have been proposed to resove these issues. One of the most influential ones is the Shifter Circuits mechanism by C. Anderson and Van Essen (1987). Originally it was intended to account for how the brain solves what they named the 'registration problem' in stereopsis: that the two eyes are never focussed in perfect accordance, and the brain will need to compensate for the difference. This is resolved by introducing a feedforward network of

units that all project to two units in the next layer, governed by a shift control mechanism that will inhibit either of these projections. The shift control mechanims operates over an entire array of neurons, so that an entire layer has its projection shift to the right or left in the next layer. This control mechanism is in turn triggered by disparity cues in the image that persist at a higher level. An extension of the mechanism with more than two projections could account for stabilized motion processing, or directed visual attention. For the latter they propose a pyramidal structure where an attention control mechanism governs the shift at each level, resulting in coarse-to-fine shift control with higher levels controlling larger shifts.

The shifter circuits model provides a clear insight into the principle of selective routing, but the routing process is completely controlled by the shift control mechanisms that affects the information flow af all the connections in the network, which constitutes a biologically implausible mechanism. The SCAN model (Postma, Van Den Herik, & Hudson, 1997) provides an alternative approach, introducing a 2D network layer as a triangular lattice. In this lattice, every unit can be either open or closed, but every unit is competing with its neighbours for being open in a WTA-fashion. This results in a fixed set of three states the network can be in, where in every state a different subset of the information is selected to pass. The state of the network depends on the presence of information at each unit, and a top-down controlled bias. The SCAN-architecture consists of a multitude of hierarchically organized lattices, covering a subset of the input pattern or image. The top layer integrates information from all lower layers, covering the entire input pattern but still selecting from a minimized number of units selected from lower layers, which renders the mechanism easy to scale up. They also emphasize that their architecture addresses the spatial organization of the input pattern and thus should resolve the binding problem.

The computational benefits of a hierarchical coarse-to-fine organization as found in both these systems plays a key role in the Selective Tuning model of attention. The above argumentation for the computational need of information tuning mechanisms in vision (section 3.2) was originally formulated to support this model (Tsotsos et al., 1995). The model consists of a pyramidally organized sheets of neurons that can be thought of as feature maps, where lower layers are reciprocally connected to higher layers in a many-to-one fashion. The initial state of the units can be affected by task bias. The bottom-up activation by the input image results in a low-level resolution image at the top layer of the pyramid. At this top layer, the units engage in local WTA competition, after which all the units of the layer below it that did not contribute to the winner are pruned. This is repeated for every layer downwards, resulting in an attentional beam of selected units with an inhibited surround.

There are many other selective routing models, but the above three describe the key properties of this hypothesis. By definition, all selective rout-

ing models have to select a set of units from a much larger pattern (the retinal image), and most models acknowledge the computational issues that arise in this task. The models described here all use a hierarchical coarse-to-fine approach to effectively face these issues. Upon first glance, it may seem that this requires extensive control over the information flow, but as these models illustrate, a correct path emerges mostly from the local interactions in the system. This makes selective routing a very plausible concept for being implemented by the brain, but only if the routing is not completely constituted by an external mechanism, but rather by local interactions of the units in the network.

3.4.2 Saliency maps

In the above discussion on bottom-up attention mechanisms the concept of saliency has been extensively discussed, and some of the characteristic models have been introduced as well (Koch & Ullman, 1985; Itti et al., 1998; Bruce & Tsotsos, 2006). Due to the focus on low-level attention in the context of these discussions however, the focus has been mostly on how these models try to capture a definition of saliency and how it is computed from the image. There are many follow-up models however, that integrate these saliency maps in a much larger vision system and use the maps to guide other visual tasks. Therefore this section will describe several of these models, illustrating the modularity of these saliency map models.

The dominant interpretation of most saliency map models is that they provide a spatial representation of points of interest in the scene, which is usually interpreted as an automatic drive guiding attention to those locations. Combined with an inhibition of return mechanism they can be used to simulate scanpaths that are often validated with eye-movement data from human experiments. Based on this idea, most models that incorporate a saliency map use it for spatial preferencing of interest points. A rather simple but effective example is the recognition model by (Walther, Itti, Riesenhuber, Poggio, & Koch, 2010) They describe an architecture that combines a simple saliency map model with a feedforward recognition network. The saliency map model has an inhibition of return mechanism, and a WTA mechanism at the top determines the focus of attention. This location is used to construct an attentional modulation mask that highlights the locations that contributed most to this location. This modulation mask in turn boosts the value of the features at that location in the recognition system. This enhanced processing of interesting locations results in rapid extraction of interesting object regions, and thus facilitates recognition.

The use of saliency as a cue for spatial guidance of attention and eye-movements is also acknowledged in the integration of saliency in the ST model (Tsotsos, 2011). There it is proposed that the AIM algorithm is used to compute saliency based on information from the lower layers of the pyra-

mid, representing maybe V1, V2 and MT, but excluding the foveated area. This results in a peripheral priority map, which is used to determine the next point of foveation. Note that this introduces a clear distinction between mechanisms of covert attention (the attentional beam in the pyramid) and overt attention (the mechanism describes here). This is a distinction that has recieved significant attention in the experimental psychology literature, but that is neglected in most models.

A more complex model that combines working memory, long term memory and task guidance with the saliency map model for visual search and recognition has been proposed by Navalpakkam and Itti (2005). They propose an extended idea of the saliency map called a Task Relevance Map (TRM). This map is modulated before scene presentation by symbolic working memory encoding task instructions such as 'look at the center', or more complex instructions such as 'what is the subject in the scene eating?'. This would rely on symbolic information from long term memory relating eating to a hand-to-mouth action, which would bias for hand- and mouthlike features. After scene presentation, a normal 'saliency' map is constructed, which is combined with the TRM into an Attention-Guidance Map (AGM) where WTA selection determines the attended location. After attentional selection, the features at the location are extracted and bound, and used for object recognition, after which the WM updates the TRM based on symbolic information. For example, if the attended location is identified as a finger, it indicates the presence of a hand. IOR mechanisms are implemented in similar fashion. The process is repeated untill the task is complete.

As noted before, saliency map models provide a biologically plausible way to transform an image into a map of interest points without losing the spatial relations. This section shows how these properties allow for insightful mechanisms to direct attention to these points, of which three succesful applications have been introduced. Because of the image-based approach of saliency map models they are easy to integrate into recognition models. However it remains a subject of debate to what extent the brain relies on such mechanisms. Multiple areas in the brain have been found with retinotopic organization that show enhanced processing of salient locations, but that does not provide a clear account of how it is computed in the brain.

3.4.3 Temporal tagging

An property largely overlooked in models based on the saliency map and selective tuning is the behavior of units over time. The firing behavior of neurons in the visual system is usually abstracted to 'activation values' or an 'open or closed state'. The Temporal Tagging hypothesis suggests that these abstractions may obscure the actual mechanisms underlying attention, as these mechanisms may be best described as the result of interactions of units over time. An important predecessor of this view is the Adaptive

Resonance Theory, which details how pools of neurons activated from top-down and bottom-up signals interact, causing oscillations until the system reaches a steady state. Carpenter and Grossberg (1987) proposed how such interactions could play an important role in recollection from memory or representing attention. Indeed, neuroscientific evidence for the role of oscillations has been found in many different tasks, including attentional selection and flow control (Sejnowski & Paulsen, 2006) where firing rates of neurons representing the attended features tend to synchronize.

Crick and Koch (1990) have proposed an architecture where such oscillations in the γ-band (40 - 70 Hz) play a key role in attention, binding and visual awareness. They suggest a saliency map architecture where upon stimulus presentation all neurons initiate firing at roughly the similar frequency, but not necessarily with the same phase. Attentional modulation then 'boosts' the units representing the most salient item which causes synchronized, phase-locked firing. This results in a bound, conscious percept of this item, which is strong enough to activate working memory until the attentional spotlight passes on to another item on the map. Although this model collapses a lot of complex concepts such as consciousness, binding and working memory into the single notion of γ-oscillations, and some of its assumptions are largely outdated, it does provide a relevant theory of what attention actually establishes *after* either saliency map mechanisms or selective routing mechanisms select the attended location.

Other models based on oscillatory mechanisms abandon these approaches altogether and focus on how the dynamics between excitatory and inhibitory pools of units actually establish attentional selection. Deco, Pollatos, and Zihl (2002) simulate reaction time data from feature search and conjunction search by a model that consists of excitatory feature maps, but explicitly dismisses a centralized locus integration. Instead, each feature map is composed of leaky integrate-and-fire units that engage in competition with a pool of inhibitory neurons. Units representing features that compose the target are boosted by top-down bias, so eventually only these units will win the competition. There's no explicit integration map, except for synchronized activation of winning units. The time needed for the entire system to converge, that is reach a steady state of simultaneous activation of the unit representing the target, matches the patterns found in reaction times in humans for different search conditions.

Although the Temporal Tagging hypothesis seems to provide adequate tools to resolve the binding problem, these two example architectures indicate a dissociation in the models based on this hypothesis, posing the question whether the synchronization of firing rates actually represents the computations of feature binding, or simply the resulting representation of already bound features. In a critical review, Shadlen and Movshon (1999) point out that it is unlikely that the observed synchronizations early in V1 represent binding computations because due to their limitations in recep-

tive fields they suffer from the aperture problem. So even if synchronization represents bound features, the real binding process has to occur otherwise. However, they propose that even this functional role of representation is unlikely, because cortical neurons seem to lack the required temporal resolution, plus it would also require 'cardinal' cells to detect this synchronization, which would defeat the purpose of synchronized oscillations after all.

The role of oscillations and synchronized firing in the mechanisms of attention and binding thus remains elusive. However, its discussion does point out the importance of an explanation for these mechanisms that seem to operate after attentional selection. Selective routing and saliency map models largely dismiss these processes; a complete model of attention should also encorporate how these signals are further processed.

3.4.4 Emergent attention

The three hypotheses presented here do not simply emphasize different aspects to attentional mechanisms, but they all propose an underlying mechanism that is dedicated to attention, which operates on the visual information flow. The last class of models deviates from this pattern, as the emergent attention hypothesis explicitly denies a dedicated mechanism for attentional selection. Instead, it postulates that attentional phenomena all arise from the internal dynamics in the visual system. These dynamics are established by, for example, the competitive lateral interactions in the visual pathway, but may be mediated by top-down influences that can bias the competition. An important finding that contributed to this view was divided attention, where attention appears to be spread out over multiple objects. This would then be established by multiple activation peaks in the visual pathway (Duncan, 1979).

Because this view relies entirely on the dynamics of the system, there is large overlap with oscillatory models of attention and emergent models, as both these classes heavily rely an the temporal aspect of attentional computations. For both hypotheses, the Adaptive Resonance Theory model (ART) has been very influential (Carpenter & Grossberg, 1987). Various models have been developed based on this principle, mostly differing in the amount of attentional computations that are considered emergent and and those that are controlled. For example, Deco et al. (2002)'s model that was discussed in the previous section, is largely based on local competition and can therefore be considered an emergent attention model, but one might argue that the true attentional computation is realized by biasing certain feature maps. The same can be argued for Desimone and Duncan (1995)'s biased competition model.

Thefore, it is posed that unlike the other hypotheses, the emergent attention hypothesis does not provide clear instructions for the basic design of models of visual attention. Instead, it emphasizes the amount of attentional

computation that can be realized *without* such dedicated mechanisms, and might be embedded in the system.

3.5 Summary

In this section, our modern understanding of vision and visual attention has been reviewed, using four different approaches. The first section illustrated our modern understanding of the visual pathway itself, the 'stage' for visual attention to operate on so to speak. The most important revisions include a hierarchical organization of the visual system, and lateral and feedback connections. The next section illustrated the updated functional role of attention, which showed that the classical interpretation as a simple selective mechanism doesn't seem to suffice. Instead, 'information reduction' or 'tuning' seems to capture the computational goal of attention, which should include various subgoals such as flow control, binding, and involvement in working memory maintenance. The discussion on the functional role of attention indicated that attentional mechanisms are likely to be found throughout the entire visual system at various moments during visual processing. This was illustrated in the next section, where various attentional mechanisms were discussed. It showed the timecourse of various attentional processes during the perception of a visual scene. The models and their underlying hypotheses discussed in the last section similarly illustrate a similar notion: attention is probably best not described as a single process, but a collection of mechanisms that operate upon various aspects of visual processing in a variety of ways.

Although visual attention is such a broad and complex concept, this is certainly doesn't make it impossible to grasp. However, the study of attention does require a more versatile approach than the simple selective definition of attention that was dominant in classical visual models and still persists today. The visual routines framework is certainly no exception, as it is completely based on the assumption of a two-stage model of vision where attentional operations utilize a simple measurement stage. The next chapter will therefore discuss how our current understanding of the visual system affects the framework, whether the framework can still be used within this new context, and how an updated architecture for visual routines should be formulated.

Chapter 4

A new theory of visual routines

This section will point out how our current understanding of visual attention and the visual system as it was presented in the previous section invalidates some of the assumptions that the framework of visual routines was based on. These assumptions and the implications of these conflicts will be thoroughly discussed, and will be used to deduce a new theory on the functional properties of visual routines as a framework for visual cognition. Based on the vast scale at which our understanding of vision and attention has changed since the theory was originally devised by Ullman in 1984 one might be surpised by this attempt and be more inclined to discard the theory of visual routines altogether. Instead, there are many reasons to preserve a large part of the framework. First of all, it should be noted that the visual routines framework is one of the few serious attempts to model the interaction between low-level visual signal processing, and higher order cognitive processes that guide us to solve the task at hand. As a recent review by Cavanagh (2011) suggests, the lack of a robust theory to model this interaction is one of the great gaps in vision research today. Also, it should be noted that visual routines provide a means to explain subtask sequencing in visual tasks, a phenomenon that still finds evidence under modern interpretations of visual attention. (Roelfsema et al., 2003; Cavanagh et al., 2001). Another important consideration is the need for such a framework in computer vision research. When the theory was devised, its appeal would lie mostly in that it provided a means to interface with low-level vision and to control selective visual attention to extract information relevant to the task at hand through a fixed library of elemental operations. Currently, Computer Vision systems could still benefit from such a system that allows them to integrate high-level symbolic reasoning techniques with noisy low-level techniques in order to process visual input in a robust fashion.

The following discussion will point out the main gaps and inconsisten-

cies in the old understanding of the visual routines framework. This will be largely based on the original formulation by Ullman (1984), but since this framework is at times relatively incomplete compared to some implemented systems built on similar principles, other models will be referred to as well. The issues with the old formulations of the base representation, the interpretation of attention, the concept of an elemental operation and the ways these are organized into composite visual routines will be discussed, and by addressing each issue with the modern understanding of visual attention a new theory on the functional properties of visual routines will be outlined. From this functional description an architecture will be derived to implement this functionality. Finally, some visual tasks from earlier visual routines studies will be used to illustrate how this new architecture might solve those.

4.1 The base representation

The 'starting point' in the classical theory of visual routines is the base representation, which represents the result after immediate and automatic feedforward processing of the retinal image. The base representation described by Ullman is an immediate, accurate and complete representation of all features created by a single feedforward pass, that will always be the same when an identical image is presented. As the discussion of the visual system has illustrated, the existence of such a static and complete base representation is disputable, as the activation pattern resulting from scene presentation is dynamic and dependant on a variety of top-down and bottom-up influences that will be considered here.

First of all, the nature of the visual system does not allow for a static base representation, due to its feedback and horizontal connections. Not taking top-down influences into consideration for now, one can describe the initial neuronal response to a visual stimulus as the feedforward activation of feature detectors, but as activation persists after this feedforward sweep, horizontal and feeback connections can alter the response profile of a neuron to encorporate context information, and thereby changing the feature detection characteristics of the visual system. An illustrative example of the potential extent of this change is the change in neuronal selectivity for face-detection to expression-selectivity (Sugase et al., 1999). The effects of these interactions are relatively long-lasting. In the case of Sugase et al., the selectivity for facial expression in monkeys seems to arise at about 50ms after the selectivity for face detection. It is generally assumed that a single feedforward pass in humans takes about 150ms, but as this feedforward pattern is created, it will immediately be altered by horizontal interactions. At roughly 300 ms after stimulus presentation the feedback effects from the top of the hierarchy will have reached the bottom affecting all layers on the

way down, but as these effects will influence every layer on the way down, it will alter the feedforward processing of these layers as well. Therefore, a purely bottom-up activation pattern that constitutes a base representation would not be immediate, and transient at best.

One could argue that this does not necessarily conflict with the definition of the base representation. Although the visual information that is available will change over time due to the dynamic interactions, effectively the base representation would still have the potential to represent all the features present at all locations of the visual field, and the notion that a base representation is transient may therefore be an issue of information accessibility, not information representation. However, the dynamic interactions in the visual system would still challenge the defining claim that all the information is *immediately* available after a single feedforward pass. Many feature representations in the visual system simply can not exist immediately after only a feedforward pass, as they actually *require* horizontal or feedback modulation.

Other issues with the base representation come from the notion that it should be complete, implying that all over the visual field all features are detected with the same amount of detail. Again without considering top-down influences on the system, this is challenged by the organization of both the retina and the visual hierarchy. Along the retina, the distribution of receptors, the rods and cones, is not uniform: the cones are mostly found densely distributed over the fovea, a retinal region that covers only the central two degrees of the visual field; the rods are much larger in number, but they are relatively widely distributed along the periphery (Steinberg, Reid, & Lacy, 1973). From the differences in properties between rods – more sensitive to light and motion, but insensitive to red colours – and cones – high in resolution and sensitive to all colors – one can infer that the representation of foveal and peripheral vision will be very different, and will certainly not allow for a complete representation. The hierarchical organization of the visual pathway provides another challenge for the notion of completeness. Some image features are only represented in higher layers as they require information integration from lower levels, and as a consequence the resolution of the representations in that layer will be lower (Orban, 2008). This implies that these higher order features can only be represented at a lower resolution,

The notion of incompleteness, especially the kind that follows from the receptor distribution on the retina, sheds new light on the use for eye movements and fixations in solving visual tasks. Under the assumption of completeness, the base representation would include all information in both the periphery and at the fovea, which would render gaze changes only necessary if the location to be attended would lie outside the visual field. Some models acknowledge this gross overestimation of the sensory capabilities of the retina and counteract it by restricting the visual image to a small area that

is to represent the fovea (Ballard & Hayhoe, 2009; Rao, 1998). This assumption, on the other hand, clearly underestimates the involvement of peripheral cues in directing eye movements and visual cognition. It is therefore important to acknowledge the fact that eye movements or fixation changes do not simply provide spatial shifts along the base representation, but also can alter the way a visual stimulus is processed within the scene as it shifts from the foveal region to the periphery. Due to the low acuity of peripheral vision and the small size of the fovea, many eye movements may be needed in order to fully explore the scene, which would enhance the effect of receptor anisotrophy on scene processing.

As indicated, the issues described up to this point all arise while still holding the assumption that only bottom-up processing influences the base representation. However, the discussion on the timecourse of attentional mechanisms described a set of mechanisms that would influence the processing of information even before the scene has been presented: positive and negative priming, task-guided attentional bias and inhibition of return all impose top-down constraints and effects on the processing of the scene, that at the same time conflict with the ideas of completeness and accuracy. Moreover, these effects could be different everytime the same scene is presented. A classical example of task-dependent scene processing comes from the early eye movement studies by (Yarbus, 1967), where different questions about the same scene elicit widely different eye movement patterns. One could argue that this reflects a difference in the routine applied in the task rather than a difference in base representation. However, results from the puzzle task described by (Ballard et al., 1997) show that even within the same task, participants displayed repetitive gaze shifts to the same location when determining either the shape, color or location of the puzzle piece, indicating that the features extracted from the same location were (sub)task dependent. This is the main reason why also the assumption that the base representation will be the same everytime the same scene is presented, its third and last characteristic and defining property, will not hold under our current interpretations of vision and the visual system.

This section has challenged the idea of a base represention as presented by Ullman by pointing out that the assumptions that it is based on – of an immediate, accurate, complete and constant representation of all features created by a single feedforward pass – do not hold. Without these functional and computational properties, the idea of a base representation may seem obsolete in the theory of visual routines. Nevertheless, the base representation also has a terminological function: to indicate the representation that elemental operations act upon. Therefore, the next sections might still use the term base representation, when referring to the activity pattern throughout the visual system that these operations could both modulate and interpret. However, it should only be regarded as a dynamic and limited mode of representation that varies over time, and allows only partial

control by attentional mechanisms. Within the context of a model for visual cognition, the base representation provided a model for the sensory mechanism of the visual system. A new architecture for visual routines will also require such a sensory module, but including the properties that challenged the classical base representation:

1. A hierarchical converging organization of feature detectors, such as a visual pyramid of layers,
2. Top-down modulations of the pathway before stimulus presentation
3. Feedforward, feedback and horizontal interactions,
4. Recurrent modulation of the pathway after stimulus presentation
5. Receptor anisotrophy,

Not many modern models of vision and attention have a visual system that meets these criteria. Many classification models acknowledge the hierarchical organization but dismiss attentional mechanisms (e.g. Serre et al., 2007), whereas most saliency models use features at different scales and resolutions to direct attention, but neglect the hierarchical organization of these feature detectors (Itti et al., 1998). Several selective routing models do model attentional mechanisms by a hierarchical system, but like saliency models and classification models, they rely too heavily on feedforward processing and do not implement any lateral and feedback interactions. Any form of lateral interaction is usually implemented by a WTA-algorithm at the output level, and any form of feedback interactions tend to be implemented by a separate control mechanism that is triggered by the outcome of the feedforward pass only. Models of emergent attention and synchronous firing on the other hand do rely heavily on lateral interactions, but tend to rely on simple feature maps instead of a hierarchical organization (Deco et al., 2002), and thereby ignore the importance of feedback interaction in neuronal modulation.

The Selective Tuning model of visual attention includes many of the properties of the visual system that seem necessary to implement the functionality that is described here (Tsotsos et al., 1995). The model assumes a pyramidal organization of the visual system, where higher levels integrate information from the lower levels to construct new feature representations (Figure 4.1). At every layer of the pyramid, lateral interactions are implemented by means the θ-WTA algorithm that allows multiple winners in one layer. The initial activation of the system is determined by the feedforward pass after which feedback tuning is recruited to inhibit unattended units. This process will after a feedforward and feedback pass result in an attentional beam, which allows the pass of all attended features and inhibits the surrounding units, providing an attentional mechanism that supports

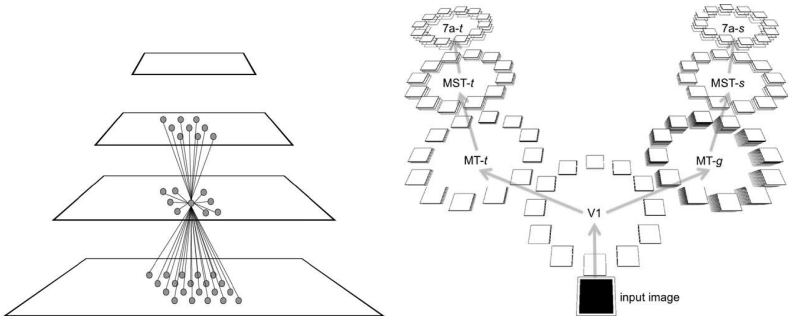


Figure 4.1: Pyramidal representation of the visual pathway. (a) The abstraction of the feedforward representation as a pyramid. Multiple interpretive units from lower layers project onto a smaller number of units in the layer above it. (b) The wide range of feature maps establishes a large number of such pyramids in the visual system. Here a possible organization of motion sensitive feature maps is given, with their associated cortical layers. The circle of 12 sheets in every area represents 12 directions of motion, which are all represented at 3 different speeds. The left branch of the pathway largely maintains this organization, the right branch integrates information to represent more complex motion patterns such as rotation, expansion and contraction at different angles and speeds. (Both images taken from (Tsotsos, 2011); For more detail on the motion model, see also (Tsotsos et al., 2005))

binding. As the following section will illustrate, the attentional process can be partially guided by a set of task-based parameters, thus allowing for top-down attentional influences.

One of the properties of vision listed above however is not by default accounted for by the original formulation of the model: the issue of receptor distribution along the retina. One potential solution to this problem would be to apply a simple filter to transform the input image into the retinal image, which is then processed by the visual pyramid. There is however an important problem with this approach, which stems from the assumption in the model that the units in a layer that engage in the θ-WTA competion represent similar features at similar scale and resolution, and not an image with great differences in the representation of the peripheral and foveal areas. Resolving this issue might require a much more complex connectivity scheme that as of yet has not been devised. Instead, an alternate approach has been proposed (Zaharescu, Rothenstein, & Tsotsos, 2005; Tsotsos, 2011). Peripheral vision may be too limited to be used for detailed attentional processing and information extraction, it can still provide cues to guide the eyes to potentially interesting and relevant locations in order to eventually extract detailed information there, following the principles of active vision. This is why ST includes a separate pathway to include the functionality to determine whether and where a gaze change would be necessary, which uses

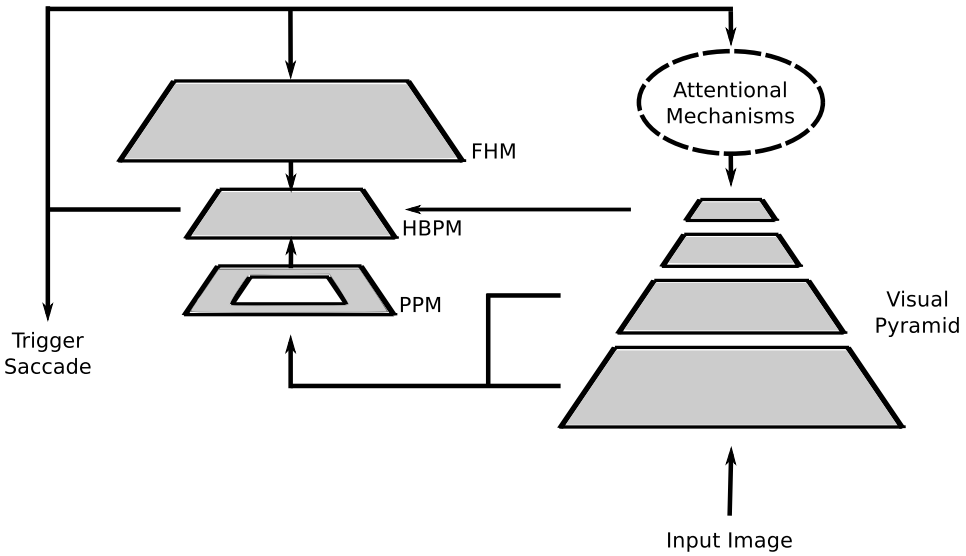


Figure 4.2: The sensory pyramid with the peripheral vision system. The oval illustrates the feedback of the HBPM output onto the attentional mechanisms that operate on the pyramid (covert shifts), which are detailed in section 4.2

peripheral vision as its input (Figure 4.2).

Like many models of gaze shifts, this system is saliency-based, and it uses three maps.

1. The first map, the peripheral priority map (PPM), is associated with the sensory processes in peripheral vision. It draws input from the peripheral units ($> 10°$) in early layers in the pyramid, and uses a saliency mechanism (e.g. AIM, Bruce & Tsotsos, 2006) to compute a conspicuous locations. This map has been associated with the parieto-occipital area (PO), an area activated by peripheral sensory stimuli.

2. The second map is the fixation history map (FHM), which represents both the visual field and a large extra-retinal area. In this map, fixated locations are registered in order to provide a simple form of memory to influence selection. This map has been related to the frontal eye fields (FEF), and it is the proposed area for computation of inhibiton of return (IOR).

3. The sensory information from the PPM, and the fixation information from the FHM are integrated to compute and update the activation pattern in the History-based Priority Map (HBPM), which will eventually trigger signals either to the motor systems to trigger a gaze shift, or to a visual executive to impose top-down spatial bias on the sensory pyramid, which could result in an covert shift.

In conclusion, the structural organization of the Selective Tuning model provides the tools to deal with issues that were incompatible with the classical formulation of the base representation. Although the issues from receptor anisotrophy are not immediately addressed within the visual pyramid, the perpipheral priority map can be used to prevent overestimation of peripheral vision capabilities, as it will enforce gaze shifts when interesting and relevant locations are not the center of fixation. Another issue for the base representation, the absense of top-down influences of attention on the base representation, was only briefly covered in this section by stating that Selective Tuning provided the mechanisms for this modulation. The next section will describe these mechanisms in more detail. First however, the notion of attention requires a more extensive discussion.

4.2 Attentional focus

The discussion on visual attention has firmly established that the approach of attention as a spotlight-like spatial filter, selecting all information at a certain location for further processing, can no longer be sustained. The discussion on the base representation supports this conclusion: when the base representation is transient and incomplete, attentional focus of an object or item can not be established by simple spatial selection. Instead, the attentional focus refers to a configuration of features that belong to a certain object, region or group of objects that are selected to influence further information processing.

The attentional focus is therefore a much more complex construct than the classical assumption, but within the context of visual routines this also means it is more versatile and more powerful. First of all, the attentional focus is not only a construct of attentional selection, but will at the same time implement a form of feature binding, as the focus provides a set of features that are processed similarly, modulated by the same attentional effects, as opposed to those not included in the sample. The consequences of this binding for visual routines have been illustrated by the work of Roelfsema et al. (2000). As they describe, this form of a binding process allows for object-based attention which introduces a new approach to several elemental operations. For example, whereas curve tracing with only a limited spatial sampling operator would require actual spatial traversal of this spotlight along the curve, object-based attention allows attentional sampling to grant an entire curve or curve segment attentional focus, and derive visuospatial conclusions from the resulting representation. The next section will discuss the relationship between attention and elemental operations in more detail.

However, this merely addresses the spatial configuration of the result of attentional mechanisms, which would only be a partial update of our current understanding of attention. As the previous chapter illustrated, attentional

mechanisms can be subdivided into categories of operations of selection, restriction and suppression, which operate not only over the visuospatial dimension, but also other feature dimensions, or over cognitive operations such as interpretations of a scene, world models and search space (Tsotsos, 2011). For now, the focus will be on the effects of these mechanisms on sensory processing in the visual system (the higher order influences on attention will be discussed later). The previous section indicated that a model of this system should include a method to implement top-down modulations of the attentional pathway, and this refers to the attentional mechanisms of selection, suppression and restriction on the visual pathway.

For these reasons, it is here proposed to regard the attentional focus as an *attentional sample* to replace the spotlight metaphor. The attentional sample is the representation constructed by the sensory activity at a point in time, manipulated by the attentional mechanisms to manage the amount and shape of the information being processed – the sample – and to which extent. The implementation of the attentional sample in the Selective Tuning model, the attentional beam, can be used to illustrate this concept. By constructing this beam by means of selection, suppression and restriction mechanisms, the construction of the beam binds selected units into a single representation, but also inhibits the surrounding interpretive units. Here, the attentional mechanisms of selective tuning that construct this attentional sample are described in more detail.

1. The initial activation pattern throughout the visual system is established by feedforward processing of the image signal. After this bottom-up signal has reached the top layer of the pyramid, attentional mechanisms are recruited to transform this activation pattern into a detailed and fine-grained representation. This is done by a feedback sweep of *selection* and *recurrent localization* mechanisms: 'Winning' units are selected whereas others are inhibited, and only the selected units are involved in the next step in the feedback sweep. These mechanisms are implemented by running the θ-WTA algorithm on the layer. In this algorithm winners are determined by lateral inhibition within a sheet, where unit A will only inhibit unit B if

   $r_A(t) - r_B(t) > \theta$,

   where $r(t)$ indicates the unit's firing rate, and task-specific parameter $\theta \geq 0$. As the impact of every inhibiting unit A on unit B is defined as $r_A(t) - r_B(t)$, the end result will be that all units will be fully inhibited, except for a bin of units with firing rates lie within θ of one another. These units form the attentional selection and are used for further processing, while the other units are fully suppressed. The feedback traversal then progresses by pruning all the units from the layer below it that did not contribute to the selected units in the feedforward pass,

which implements restriction of the search space. This process then repeats for all layers all the way down the pyramid.

2. The feedback sweep of the attentional mechanisms in the Selective Tuning model acts as a search process through the hierarchy of interpretive units. From the top layer down each selected unit branches out to a number of units that feed to it, out of which again winning units are selected and the other units are pruned. This is an instantiation of *branch-and-bound*, a mechanism for optimized search through such a hierarchy (Lawler & Wood, 1966). To implement branch-and-bound through recursive pruning, the Selective Tuning network is equipped with two gating networks. The network of gating units ς controls per unit whether they are involved in lateral interactions and competitions, the network of γ-units controls each feed-forward input to the unit. During the feedforward pass the ς-neurons are by default switched off, the γ-units are all switched on. When the feedback pass begins the ς units are switched on, to start the θ-WTA process. When the competition has converged, only the ς controlling those units that have contributed to the winners are switched on, and the γ units are switched off accordingly.

3. The bottom-up activation of the network (and therefore the eventual attentional sample) is not just determined by the input image, but is be affected by *top-down priming effects*. To realize this, the Selective Tuning model has a network of Bias units, that can be used to suppress task-irrelevant units, by inhibiting their spike rate through multiplication with a factor $0.0 \leq B(t) \leq 1.0$. This bias is not necessarily spatially uniform throughout the sheet of features, as it can also be used to prime for certain locations within in the visual field.

4. Selection and recurrent localization in selective tuning is realized without embedding any spatial information. Therefore, it is possible for spatially disjoint objects or regions to have their representations included in the winning bin. Although *spatial continuity* of an attenial sample might not always be required, it may be an important selection criterion in some tasks. Therefore, Selective Tuning provides an other selection mechanism that makes it more likely to select the largest, strongest responding spatially contiguous region from the initial selection. This is implemented by a competition mechanism works much like the initial θ-WTA competition, but it includes a factor for the distance of competing units: the inhibitory effect of unit (x, y) on unit (w, z) in the same sheet is defined as

$$\Phi(x, y, w, z) = \mu(r_{wz}(t) - r_{xy}t)(1 - e^{\frac{\delta^2_{wzxy}}{\varsigma^2}}),$$

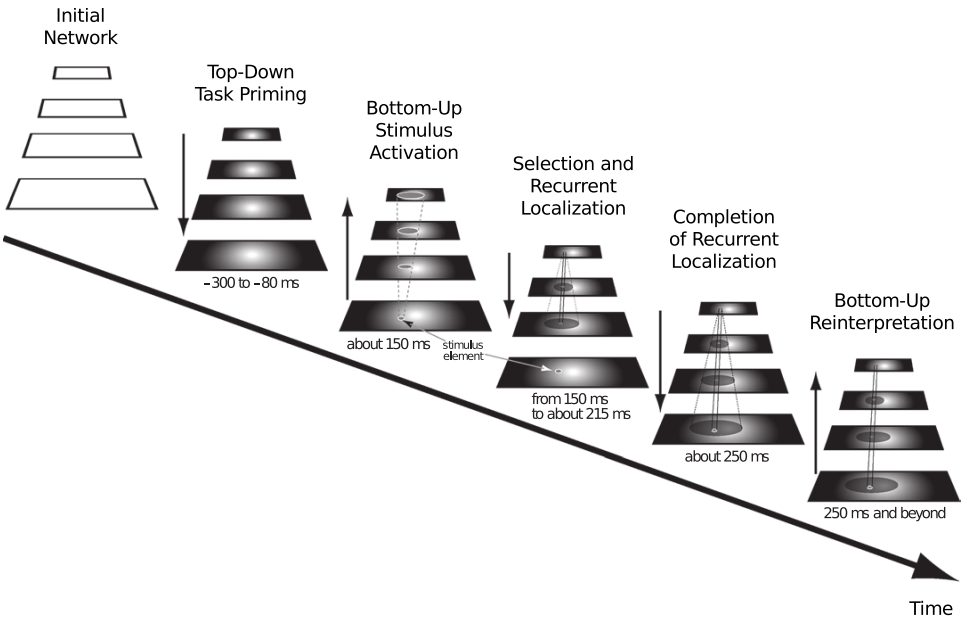


Figure 4.3: The steps of the selective tuning algorithm illustrated. Dark shades indicate suppression of those units, the illustrated timecourse indicates the approximate timecourse of corresponding mechanisms brain. Adaptation from (Tsotsos, 2011)

Where μ indicates the effect of this competition, δ_{wzxy} is the retinotopic distance between two units, and ζ can be used to control the spatial variance in this process.

The algorithm of these attentional steps is depicted in figure 4.3.

These processes describe the attentional mechanisms that are implemented by selective tuning to operate over the hierarchical visual pathway. As can be seen, there is a variety of parameters – B, θ, ζ, μ – that govern these attentional processes, and have to be determined by higher order processes. The processes controlling these parameters will be discussed shortly, but the next section will first discuss a concept at the level between the sensory system and higher-order processes: the elemental operations.

4.3 Elemental operations

In the classical formulation of visual routines, elemental operations were defined as the steps that composed the routines. Ullman, however, did not provide a restricting definition of the basic operations. Instead, he illustrated some possible candidates based on the reasoning steps in a set of visuospatial tasks, and assumed these were part a much larger set of elemental operations. This approach permitted other visual routines models to formulate any type

of function or operation that would suit the task at hand, and justify this as an elemental operation, to the extent of even including complex motor functions such as obstacle avoidance (e.g. Sprague & Ballard, 2001).

An important exception is the work of Rao (1998), which made an attempt at a structured definition of elemental operations, and defining the relation between visual attention, elemental operations and visual routines. He describes a visual system where attention is modeled as saliency computation combined with color- and motion blob detectors to select blobs of interest. Elemental operations are defined as the steps needed to interact with attention and the visual system: *shifting* the focus of attention, *establishing properties* at the focus of attention, and *selecting* new locations to attend to. Visual routines are then composed by cycling through these classes of operations in this order. Rao later emphasizes the close relation between elemental operations and attentional focus by postulating that visual routines should not be thought of as hard coded programs, but simply patterns of attentional shifts that originate from experience. Within the context of the Selective Tuning model, the distinction between these three types of elemental operations of attention is less strict. The processes described above implement mechanisms to simultaneously select and shift attention, and the extraction of information at the attentional focus is assumed to be the immediate consequence of attentional tuning. Important though, is the emphasis on how elemental operations are in fact realized by control over attentional focus.

As was briefly indicated in the previous section, the studies of Roelfsema et al. (2000) provide an example how attentional sampling could be used to realize some of the elemental operations described by Ullman. Roelfsema et al. describe a visual hierarchy where the feedforward pass established activation of an 'interaction skeleton' of activated interpretive units. Attentional mechanisms then realize an 'attentional label' that spreads along the activated units that belong to the target object or region, binding them into a single representation. Ullman's operations can all be reinterpreted by this process: *region filling* and *curve tracing* are both forms of task-guided spread of the attentional label, either along a curve or over a region. *Shifting the attentional focus* should, as in Selective Tuning, not be interpreted as a spatial shift of a spotlight, but as the construction of a new attentional sample, which might even be at the same location but involving different features. Similarly, *marking* could be realized by applying an attentional label to feature units, which would again not only operate on a spatial level.

There are, however, several issues with this scheme of elemental operations defined solely as attentional sample construction through attentional labeling, which largely stem from a lack of detail in the definition of 'attentional labeling' and its underlying mechanisms (figure 4.4). First, it seems that the attentional label determines whether units are included in the attentional sample or not by selecting a starting point and including contiguous

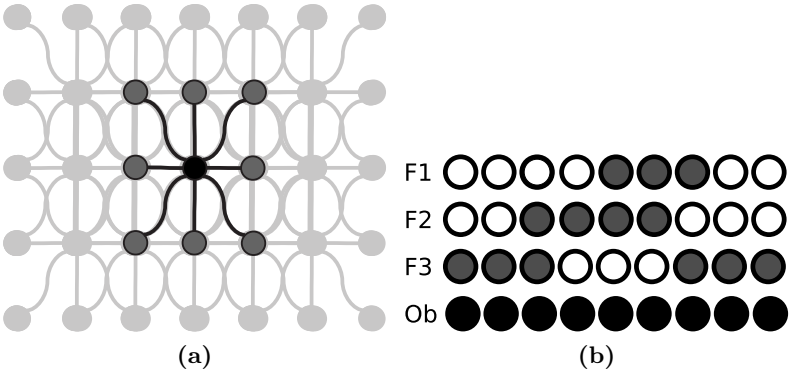


Figure 4.4: Problems with attentional label spreading illustrated. (a) A 2D illustration of lateral inhibition within feature maps (exerted by the central unit): a darker shade of gray implies a more heavily activated unit. Due to the competition that results from lateral inhibition, this activation pattern will most likely cause the central unit to remain active after horizontal competition, whereas the less activated would be completely inhibited. The contingency condition in attentional label spreading would, however, imply that these nine units were included in the sample, as they are activated and contingent. (b) 1D abstraction of label spreading over three activated feature planes within the same layer (F1-3). The attentional label theory would allow the object-based attention pattern (Ob) to spread throughout the entire visual field due to the spatial overlap between these units, whereas the pattern in F3 might indicate 2 distinct objects.

units that were also activated in the feedforward sweep. The inclusion criterion of contiguity however seems to ignore the well established mechanism of competition within layers of the visual cortex implemented by lateral inhibition. Lateral inhibition especially affects units that represent spatially contiguous locations. Second, as Roelfsema et al. states, label spreading is not not limited to contiguous units representing the same features. The attentional label can also spread across layers via activated units representing the same location but different features. Since there are no boundary conditions implemented however, this could easily allow the attentional label to spread across various feature maps, including an implausible amount of feature units with overlapping or contiguous receptive fields. Moreover, there are no boundary conditions that prevent the attentional label from spreading out to the periphery, thereby overestimating the capacity of peripheral vision by constructing detailed attentional representations. Retinal anisotrophy would not allow for such detailed attentional representations that could potentially span the entire visual field. In short, by only detailing inclusion criteria and lacking exclusion mechanisms, the theory of the attentional label seems to overestimate the capacities of the visual system and visual attention altogether.

At this point it should be noted that section 4.2 has already presented an alternative to contiguity-based attentional label spreading theory that

is part of the Selective Tuning model. A stage of competition among the winning bin of the initial selection process was described, which allows for parameters to enforce construction of an attentional sample that reflects a spatially contiguous region. Because this form of competition operates after the initial selection process, the issues presented here will be resolved, while still providing a detailed approach to the 'object-based attention'-proposal attentional label spreading was designed to implement. At the same time it provides a useful flexibility, as objects are not necessarily defined by spatial contiguity.

In short, the studies of Roelfsema et al. and Rao illustrate how a reinterpretation of elemental operations as controlled steps of attention provides a fruitful definition to approach problems of visual cognition. Although one should be careful not to overestimate the capacities of attentional mechanisms in constructing representations, abandoning the simple spatial definition of attention, and emphasizing its binding properties as a sample construction mechanism allow for a much more flexible approach to visual problems. Nevertheless, some tasks and operations will require information from multiple attentional samples. Also, within the broader perspective of visual routines, it will be required to have a mechanism to store multiple visual representations and draw conclusions from several of these items. To these ends, a visual Working Memory system (vWM) will have to be defined.

4.4 Working memory

Interestingly, the classical visual routines theory does not discuss a module for storage of representation or other types of information. Instead, a visual routine is assumed to be realized as a chain of incremental representations: every operation recieves one as input and passes the transformed representation as a result. Other systems based on the visual routines framework, especially those involving gaze changes, illustrate the importance of a way to store information, usually implementing a set of buffers where extracted information is stored to be consulted at later steps in the analysis (Horswill, 1995; McCallum, 1996; Rao, 1998). Remarkably, none of these models seem to explicitly link this storage facility to the study of (visual) working memory. For a complete and integrated theory of visual cognition however, the findings of visual working memory studies should be included in this model to illustrate the underlying mechanisms of visual storage. Recent findings in indicate that visual working memory should not be viewed as simple feature buffer, and instead illustrate more complex mechanisms that closely interact with the visual sensory system. Here, three arguments for this view are presented, which will be used to illustrate a modern view on visual working memory that can be embedded in the model.

First, it appears to be impossible to identify a single separate working

memory system or buffer. One of the most influential models in the field of working memory is the multiple components model (Baddeley, 1992), which introduced a dissociation between visual and phonological working memory. When studies suggested that the dissociation in the visual pathway–'where'- versus 'what'-information–can also be found in working memory, the model was adjusted accordingly by separating the visual buffer in two distinct components. In an extensive review, Postle (2006) points out that this dissociation in working memory types is no rarity. Instead, it seems that separate working memory storage systems are used for every type of information that can be dissociated in its sensory mechanism as well, which is feasible if working memory storage is achieved by the same systems as those that process the information initially. This is one of the arguments that led to the view of *emergent working memory*, where working memory items are a product of all brain mechanisms representing the information it contains. For example, visuospatial information of a working memory item could be represented by both the visual system, and motor systems coding for the gaze direction of the item. Another convincing finding that supports this argument illustrates how the entire visual system, including V1, seems to be involved in working memory storage comes from fMRI decoding experiments (Harrison & Tong, 2009). In this study, the fMRI activation pattern in V1 during a working memory task could be succesfully used to classify which of two differently oriented sinusoid gratings was held in working memory.

Second, there appears to be a very close correlation between working memory mechanisms and attention. For the early models of visual routines, the metaphor introduced by (Awh et al., 2006) seems most apt, where attention acts as a 'gatekeeper' for working memory. Only items that occupy the attentional focus have the potential to enter working memory. The effect of visual attention on visual working memory items is not limited to the encoding stage. Awh and Jonides (2001) review studies where visual attention tasks seem to disrupt visual working memory representations when they operate along the same dimension. For example, performance on a working memory task of spatial rehearsal was compromised when participants were given a task that required a spatial shift of the attentional focus.

However, the correlation between attention and working memory is not just a one-way interaction, as there is evidence that working memory content and load influences attentional performance as well. ERP-studies have been used to elicit the relation between search targets in working memory, their load, and search performance (Carlisle, Arita, Pardo, & Woodman, 2011). Using the Contralateral Delay Activity (CDA) as a measure for working memory load (Ikkai, McCollough, & Vogel, 2010), they find evidence that templates for search targets are stored in working memory. If the search target is persistent over trials however, the CDA disappears within seven trials, which seems to indicate that the search now no longer relies on guidance from a working memory template but from a long-term memory template.

This experiment describes the influence of working memory on attentional selection when working memory content is dedicated to the task. However, working memory load will even affect attentional performance when the working memory task is unrelated to the attentional task (De Fockert, Rees, Frith, & Lavie, 2001). They describe an fMRI-experiment where participants were required to contain digits in working memory, during an attentional selection task with stimuli consisting of faces with superimposed names. Participants were instructed to only respond to the text, but when working memory load was high, performance on the attentional task decreased and there was more activity in the fusiform face area. This indicates that participants were not able to inhibit the response to the facial stimulus as much as with low memory load. Although this study does seem to indicate that selective visual attention and general working memory rehearsal share similar resources, this would likely concern a more general issue of cognitive control over both these processes rather than conflicting representations.

Third, although working memory capacity is one of the most extensively studied facets of working memory, it has proven hard to provide an accurate quantization of it. Initially, capacity studies seemed to rely on the assumption that working memory provides several 'slots' where memory items are stored. The classical quantification of working memory capacity is seven plus or minus two items (G. Miller, 1956), but this number has been heavily disputed since, as it was proposed that this number didn't account for the process of chunking multiple items into single representations. More recent experiments that attempt to prevent chunking point towards three to four items (Luck & Vogel, 1997; Cowan, 2001). These studies, however, could not provide a consistent theory of *why* working memory capacity was limited. Several studies point towards a relation between working memory load and oscillations in neuronal firing (e.g. Lee, Simpson, Logothetis, & Rainer, 2005) which would cause interference when four or more items are stored (a mechanism that again illustrates similarities between working memory representations and attentional representations as proposed by temporal tagging and emergent attention models). However, this raises the question whether interference is as likely to occur when representations share the same resources, or whether a model that limits capacity in feature space instead of object space (e.g. Baddeley, 2003) is more accurate.

In a recent review, Brady, Konkle, and Alvarez (2011) gather (occasionally contradicting) evidence from studies of visual working memory capacity and fidelity and identify several important charactheristics of visual working memory: (a) the fidelity of working memory representations depends on the number of items; (b) objects can not be stored independent of their information load, but using objects more features can be stored in total; (c) working memory items interact, both spatially and contextually (d) knowledge and expertise influence working memory by potentially increasing capacity, but also by biasing towards certain features during recall. They conclude that no

accurate estimates of visual working memory capacity can be made without considering the structure of the stored items, and propose that items are organized into 'hierarchical feature bundles'. This may not be surprising, given the evidence that working memory representations are stored using the associated sensory systems, and given the hierarchical feature organization throughout the visual pathway. This would indicate that working memory items are not too different from attentional samples.

These three categories of findings can be combined to outline an architecture for visual working memory. After an attentional sample has been created, it provides a representation that can be stored as a working memory item. Like the metaphor of attention as a 'gatekeeper' suggests, the attentional sample forms a selection of features that are selected to be stored. A hierarchical organization of these features is inherent to the organization of the visual system. However, the visual pathway is not only involved in encoding the memory item, but is used for maintenance and retrieval as well. This illustrates a model of visual working memory as a dual component system: a passive component used used for representation and storage of the items, which overlaps with the visual pathway, and an active component that influences these representations via mechanisms of suppression, restriction and inhibition, very similar to those used to guide visual attention. This dissociation between representations and maintenance control would also provide an explanation to the different types of interference between attention and working memory. The influences of spatial and temporal context from visual information on working memory items occur in the visual system at the level of representation, whereas influences of expertise or cognitive load would occur at the level of control, thus affecting item maintenance and inhibition of interference from other items, sensory input and noise.

The proposal of working memory distributed over a storage system and a control system raises several questions regarding representation: are working memory items maintained in the storage system (i.e. the visual hierarchy) while the control system supports maintenance by inhibiting interference and dissociating between the various items in this storage system, or are the items actually represented in the control system which can actively re-visualize these items using the storage system? Various studies on lesions in the PFC, as well as on amnesic patients with intact sensory functioning, seem to indicate that simple storage of items without distracting stimuli is still intact (Postle, 2006; Dewar, Della Sala, Beschin, & Cowan, 2010). This would imply that storage is a property that arises in the sensory systems. However, this again raises the question how working memory control manages to avert the interference between working memory items and feed-forward input, as they would simultaneously be represented by the same units. An interesting new perspective regarding this question comes from evidence that indicates a functional and physiological dissociation between working memory items that are actively used in the perception task, and

so-called Accessory Memory Items that are maintained to be applied later (Olivers et al., 2011). It seems that only the relevant working memory items affect perceptual processing, indicating a mechanism where working memory control activates only this item in the perceptual system. How these active memory items are represented with respect to the accessory memory items, and how working memory control prioritizes and deprioritizes items would probably also resolve the debate between revisualization versus maintenance of items.

Nevertheless, the separation of working memory control from the storage facility provides an interesting perspective on visual working memory processes, which is reflected in several other models of working memory and perceptive processes. For example, a similar duality is present in the popular multiple components model of working memory. There, several components are characterized by a static component (e.g. the visual buffer or the phonological store) and a more dynamic component (e.g. the visuospatial sketchpad or the phonological loop), which has led to interpretations that simply summarize the model as a set of static storage components and dynamic rehearsal components (Reisberg, 2006), and is similar to the duality presented here. The architecture also illustrates how working memory control is a generic process that interacts with the systems and likely the motor systems as well. Within this interpretation, *visual* working memory is then defined as the manifestation of these control mechanisms within the visual modality, but this does not imply that operations of visual working memory will solely rely on information from the visual areas. The definition of working memory control as a higher-order generic process indicates that it would be able to integrate information from various other sensory or motor areas. A similar notion resonates in theories that postulate that computations at this timescale ($> 300ms$) are characterized by computations of embodiment (Ballard et al., 1997) and the integration of sensory and motor information, but also in theories that approach memory items not just as a simple feature set, but a large-scale hierarchical representation of stimulus features, combining contextual features from all modalities, 'greedily' recruited during encoding (Postle, 2006).

The architecture presented here only implements the visual modality and therefore the discussion of memory items is restricted to visual information. Nevertheless, the variety of components could be used to illustrate how such a 'rich' memory item could be organized and utilized. When an attentional sample is stored, it will not only include the features that have been selected in the visual pathway and are included in the sample. As an object is fixated, it's fixation will be also registered, for example in the FHM in the peripheral vision system.

It must be noted that the proposed working memory control module does not explicitly represent a designated brain area, although various studies point towards frontal areas such as the PFC as a structure for working mem-

ory (Braver et al., 1997; Goldman-Rakic, 1987). The discussion whether a designated area that implements this functionality exists or whether it is an emerging property from various interactions involving higher-order areas is beyond the scope of this review. However, this question does introduce an important issue of how detailed sensory information is communicated between the visual system and both higher-order areas and other sensory areas. The visual pathway only allows for interaction with the highest layers, and this is not sufficient for detailed working memory items. Tsotsos (2011) adresses the issue of communication between the visual hierarchy and proposes the thalamus, and in particular the pulvinar for this functionality. The thalamus has often been suggested as an area of multisensory integration, and the visual system is known to project onto the pulvinar, especially areas V1 and V2. The pulvinar is known to be involved in attentional tasks, and is one of the areas that implements a map representation of the visual field. The proposed role of the pulvinar is to act as a passive 'blackboard' that the visual system 'writes' on, projecting the information from the attentional sample which is then readily available for other areas to read (for a similar proposal, see Cavanagh, 2011). The blackboard could then provide detailed information from lower layers to tune higher order units, but more importantly it could provide the same information to other sensorimotor areas. Regardless of how working memory control is implemented, it would then very likely draw information from this blackboard representation as well. The organization of the components that establish visual working memory in this model is depicted in Figure 4.5.

As is illustrated by the figure, the function of the blackboard is mostly targeted at conducting information from the lower layers in the hierarchy, although the higher layers in the hierarchy contribute as well albeit to a lesser extent. In the visual pathway, these units from different layer are hierarchically organized, and it seems reasonable to assume that this hierarchical organization is preserved in the blackboard, to assure maintenance of its functionality and properties associated with the connectivity in the pathway. Therefore, it seems that the blackboard is abel to establish a similar hierarchical organization. However, it must be emphasized that the blackboard is not a straightforward structural copy of the visual pathway, as only the attentional sample needs to be represented, and units in the blackboard can represent different features between different attentional samples that are represented.

So far several modules have been illustrated to describe an architecture for visual cognition. Although these modules interactively manage to resolve subtasks of visual cognition, they all have been described to be subject to task guidance from higher order areas. The next section will provide further detail on the way task influence acts affects the modules to allow for task-specific elemental operations to be realized.

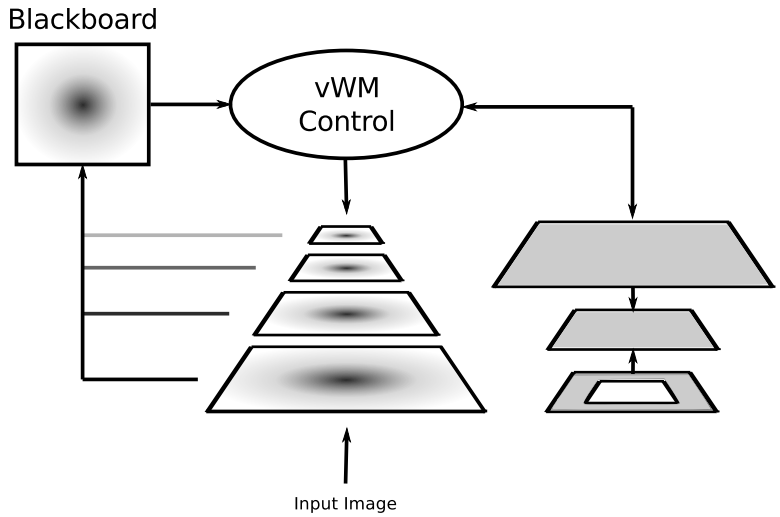


Figure 4.5: The proposed organization of visual working memory control. After an attentional sample has been constructed, it is 'written' to the blackboard mechanism. This representation mostly relies on the otherwise inaccessible detailed representations in the lower layers of the pyramid (indicated by the darker shades of their efferent connections). Working memory control uses the blackboard representation to represent the memory item in the pyramid. On the side of the peripheral vision system, the FHM can be interpreted to link spatial fixation information from the eyes to the memory item. To revisualize this representation, the FHM can be used again. For simplicity, the earlier discussed connections (e.g. between the pyramid and the peripheral vision system) have been omitted from this graph (but see Figure 4.2)

4.5 Task guidance

One of the strengths of the classical visual routines framework is that approaching tasks using elemental operations allowed for a generic modular solution for a large variety of visual cognition problems. The only form of top-down task guidance that is implemented in this framework is the visual routines processor, that combines the correct elemental operations into ballistic routines. The new framework that has been laid out so far, requires a more elaborate scheme of task-guidance which interacts with the sensory visual pyramid, the peripheral visual system, and working memory representations. This section will detail the mechanisms that can be used to exert this top-down influence, but first, an attempt will be made to identify this executive module.

In the previous section the similarities between visual working memory and visual attention have been stressed, partially illustrated by evidence of interference between these two processes. The proposed organization of visual working memory would support how these interactions seem to occur at two levels. One is at the sensory level, where attentional representations and working memory items can disrupt one another as they share the same representational resource. This form of interaction is illustrated in the review by Awh and Jonides (2001). There is, however, also evidence of interactions between working memory items of other modalities and control of visual attention (De Fockert et al., 2001), as well as control of visual attention originating from working memory (Ikkai et al., 2010). Because these interactions occur across different modalities and have a more cognitive than sensory nature, it is here proposed that they occur at a higher order level involving more widespread and complex structures than just the sensory systems. The similarities between functionality and resources of working memory control and attentional control could be explained by a model where generic cognitive control is implemented throughout the brain and manifests itself in multiple ways including control over (visual) attention and working memory.

The wide range of research on these higher order processes will not be fully reviewed here, but findings from this field can be used to illustrate the way the visual executive realizes visual routines. Like working memory control, higher-order cognitive functioning is usually associated with PFC, but also the Anterior Cingulate Cortex (ACC) – for conflict monitoring and error processing – and the limbic system – to regulate rewards for learning (E. Miller, 2000). Two fields of study of these higher order processes provide an indication of how visual routines could be implemented within this architecture. First, cognitive functioning in the PFC is associated with goal-directed functions through discrete IF-THEN rules, which is reflected in cognitive architectures such as ACT-R and Soar (J. Anderson & Lebiere, 1998; Laird, 1987). It has been shown that PFC-neurons show properties that

can be used to implement discrete rule-based 'steps' in cognitive processes despite the massively parallel processing of sensory information (Zylberberg, Dehaene, Roelfsema, & Sigman, 2011). Parallel influx of sensory evidence from multiple systems can be used to integrate evidence that triggers certain output processes. Positive feedback triggers reward that can be used in learning, i.e. to strengthen the relation between gathered evidence and the resulting action. Evidence suggests that the PFC is largely involved with goal maintenance and keeping to a task, which is an important requirement when tasks are divided into subtasks, as is the case in visual routines. With sufficient practice, the execution of such a sequence of subtasks becomes automatic and PFC involvement decreases. Two possible reasons for this could be that antecedent-consequent relations have been strengthened and less evidence needs to be gathered before a production rule is triggered, or that there is less need for active goal maintenance (E. Miller, 2000).

The second field of study targets the output of these mechanisms and rules. Most cognitive architectures emphasize the cognitive process within the frontal structures, and initiate motor programs as output, implicitly following the classical perception-cognition-action pipeline. However, aside from limited control over eye movements, visual executive functioning largely involves influence on the sensory systems, which is more similar to the conflict monitoring and control mechanisms as modeled by Botvinick, Braver, Barch, Carter, and Cohen (2001). Their study consists of extending a set of classical neural network models with a single unit that measures conflict in activation between different output nodes. In the case of conflict, the node triggers top-down enhancements of the task-relevant input neurons. Findings from neuroimaging studies support this model, indicating a clear role of conflict monitoring in the ACC and task preparation and control in the Dorsolateral PFC (MacDonald, Cohen, Stenger, & Carter, 2000), combined with higher PFC activation after trials with high conflict, followed by a higher response in relevant sensory cortical areas (Egner & Hirsch, 2005).

Visual executive processing can be identified by the application of these mechanisms to the visual domain. Information extracted by the sensory mechanisms could be combined with the current goal and objectives, which would trigger production rules. These rules, acquired through learning, could feed back into the sensory mechanisms and guide the sensory process to suit the task at hand. The next chapter (5) will describe how these influences can be organized into discrete steps as proposed in classical visual routines using the rule-based theories of cognitive processes described here, but first the 'tools' to exert top-down influence on the sensory systems will be discussed. The cortical amplification found by Egner and Hirsch (2005) seems to indicate an effect of top-down bias to relevant feature units, but the discussed components in the architecture allow for several other modes of top-down task influence, in order to realize suppression, restriction and selection. These modes will be discussed here for the visual pyramid, the

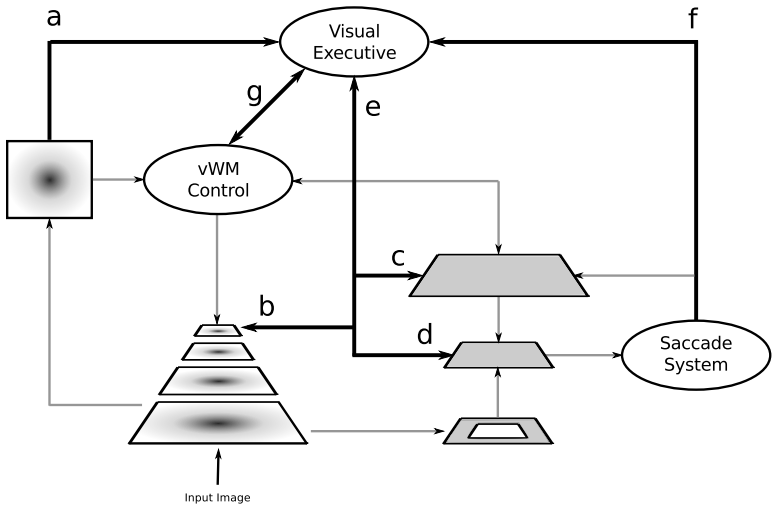


Figure 4.6: The visual executive, and its interactions. The lighter gray connections have already been discussed. (a) Reading the attentional sample from the blackboard (b) Modulating the parameters for attentional tuning in the pyramid (c) Modulating IOR shape in the FHM (d) Modulating the HBPM by biasing (e) Reading the top of the pyramid, the HBPM output and the current fixation on the FHM. (f) Recording feedback from motor systems when saccade has been triggered (g) Reading and modulating the organization of working memory items. See sections 4.5.1–3 for more detail.

peripheral vision system and visual working memory. The interactions between the executive and these systems that will be discussed are depicted in figure 4.6.

4.5.1 Top-down influence in the visual pyramid

The attentional mechanisms in the visual pyramid that are implemented by the selective tuning model have been detailed in section 4.2. The formalization of these mechanisms allows for easy identification of parameters that are dependent on higher-order cognitive functions. These are the following:

1. Before the stimulus image is presented, the visual executive can already exert top-down task influence to prime the system. In Selective Tuning, this is implemented via the Bias-subnetwork B, which can be used to model various forms of priming. The classical form for this bias network to operate is on a feature-basis, e.g. by inhibiting certain colors or types of motion, simiar to the bias applied in many saliency map models (cf. Koch & Ullman, 1985) or the guided search model (Wolfe, Cave, & Franzel, 1989). To accomplish this, task instructions can be used to inhibit feature maps that are irrelevant to the task. However, the bias network operates on feature units, not entire maps, which allows for modeling spatial bias as well. This can be based on

a spatial cue or simple task instructions – e.g. "the target will be at the top-left of the image", but also gained from experience from previous trials – e.g. the target tends to appear at a certain location – or experience and long-term memory, such as center bias on computer screens, or other forms of compositional bias (Tatler, Baddeley, & Gilchrist, 2005). Finally, the bias network can also be based on a template, for example to combine the spatial - and feature properties of a target that has been analyzed before. This form of bias can be based on the attentional sample of the target template which has been transferred to working memory, or it could originate from long term memory after sufficient practice (Carlisle et al., 2011).

The bias network is not only used to prime the system before the stimulus has been presented, but can also be recruited in response to the feedforward activation pattern, when interpretation of the feedforward activation contains information about the target. Also, if the feedforward activation indicates that certain features are simply not present in the scene, the bias B can be set to 0 for these channels, to inhibit any potential noise. Similarly, the competition control units ς can be set to 0 for these maps.

2. After the stimulus has been presented and the feedforward activation throughout the pyramid has been computed, the next step in the Selective Tuning algorithm is recurrent localization. The implementation of this step, the θ-WTA competition, makes use of the task-dependent parameter θ which determines the maximum difference in feedforward activation between included and excluded units, and thereby it affects the size and shape of the attentional sample. The value of this parameter would heavily depend an the task instructions or the target template, but the resulting attentional sample could provide feedback to adjust for a more selective (smaller) or a more liberal (larger) θ to obtain an appropriate attentional sample.

3. As illustrated in section 4.2, the recurrent localization mechanism is not solely defined by the initial θ-WTA competition. The model provides another competition mechanism in order to refine the recurrent localization process and impose a certain amount of connectedness between the units in the 'winning bin'. The effect of this second competition stage is dependent on the task requirements that impose a certain amount of spatial contiguity in the attentional sample, which could for example be based on a target template. There are two task-dependent parameters in the formalization of this competition stage. ζ is used to determine the spatial variance: a larger ζ implies a larger tolerance for spatially disjoint units in the attentional sample. μ defines the influence of this competition stage, where increasing μ from

1 implements an increase of this inhibitory effect.

It is worth to briefly note that although these parameters allow one to embed task-guidance in different steps in the selective tuning algorithm, the execution of the algorithm itself may be task-dependent as well. The algorithm consists of multiple passes through the pyramid, but certain tasks may be solved only by a single feedforward pass, e.g. in simple detection tasks (Tsotsos, 2011), or by an incomplete feedback tuning pass, e.g. when the attentional sample would not require this level of detail. This way attentional selection is not a ballistic process, and the visual executive can interrupt it based on the current task requirements.

4.5.2 Top-down influence in the peripheral vision system

The detailed formalization of the sensory pryramid in the ST model makes the modes of task-guided influence on this system easily identifiable. The peripheral vision system does not have such a formal definition, and thus requires a more functional approach to identify which processes in this system rely on top-down influences from the visual executive. There are multiple processes that use this system, but this section will illustrate that relatively few of these require top-down task guidance, or even benefit from it.

As stated, the PPM integrates input from the lower layers of the visual pyramid. This means that many of the top-down effects on the sensory pyramid are transferred on to the activation pattern in this map. Therefore, there seems to be no need for additional task-dependant biasing. Then, in this map a saliency based mechanism can be used to compute interesting locations. AIM (Bruce & Tsotsos, 2006) has been proposed as a suitable mechanism. One could introduce top-down influences to attempt to guide the saliency computation, for example to bias for task-relevant feature components. Also, the input to the PPM could be weighted, which would emphasize saliency features at certain scales more than others. However, by omitting any higher order influences in this map, it provides a suitable location for attentional capture, which was initially the phenomenon that saliency models attempted to address (Tatler, Hayhoe, Land, & Ballard, 2011). As has been discussed, attentional capture seems to be largely if not completely a bottom-up process. Although the definition of a bottom-up processes does not exclude priming effects, influences of the visual executive on the PPM are not included in this architecture.

While the PPM recieves input from the sensory pyramid, the FHM recieves input from the motor processes involved in steering saccades, and is used to keep fixated locations in memory in a fixation-centered map. Whenever a saccade is triggered, the map will have to update accordingly to retain the fixation centered representation. This computation is completely determined by the saccade and does not require any top-down influences. The

FHM allows for an implementation of mechanisms of inhibition of return (IOR). Research in IOR still faces many questions and is looking for suitable methodologies to answer them (Wang & Klein, 2010). Testable models can be used to formulate new research questions and suggest ways to answer them.

There are multiple ways in which top-down guidance could influence IOR, some of which touch on research questions that have not yet been studied (Wang & Klein, 2010). (1) Task-guidance could affect the duration of the inhibiton. Findings indicate that the inhibition effect lasts for at least 1000ms, but task guidance could lengthen or shorten these effects to suit the temporal dynamics of the environment and the task; (2) task reasoning could be used to affect the shape of the inhibition tags. For example, different task demands could determine whether, after fixating an object, it is appropriate to inhibit the entire object or only a fixated fragment or component of it. The size and shape of the attentional sample would provide suitable guidance to determine the shape of these tags; (3) It appears that IOR-effects disappear when the search display is removed. This implies that a sufficient change in scene properties, interpreted by the visual executive, could trigger removal of the inhibitory tags; (4) Similarly, when the task requirements change, for example after the target has been found, the executive could trigger removal of the tags; (5) It appears that no studies have investigated the properties of IOR with dynamic search arrays so far, but if IOR is more object-based than spatially based, the FHM would need to be updated according to the motion in the scene, which could be regulated by the executive.

The HBPM is a map representing the current field of view, which integrates information from the FHM, the PPM and the top of the visual pyramid. Based on this input it permits competition among 'interesting' locations, either in the periphery or in the central field of view. This map is interpreted by the visual executive to select potentially interesting locations in the sensory pyramid, but if activation in the periphery is strong enough it will select a saccade to fixate the interesting location. There is a large involvement of interaction with the visual executive in this process: For example, it could influence the integration process by weighting the input from the three maps. Similar to the proposed influence on the sensory pyramid and the FHM, the executive could determine the size and shape of what would be considered a target point or region. Also, the task demands could influence the acuity of the output of the HBPM. For example, when temporal demands are heavy (e.g. during driving tasks) fast saccade triggering may be more important than accurate triggering. Deprioritizing accuracy could account for phenomena such as the center-of-gravity-effect, a finding from eye movement studies that a saccade lands between two salient target locations (Tatler et al., 2011).

Aside from these effects from the executive on the HBPM, it has already

been mentioned that the HBPM can also provide bias to the sensory pyramid which would be mediated by the executive. This has been identified as the descision of a covert shift opposed to an overt one, made in this map. In the case of an overt shift however, various parameters and top-down influences that have been ascribed to the executive, will also have to be adjust. For example, the attentional sample that has been constructed will need to be released, and the spatial bias needs to be relocated accordingly. Several attentional mechanisms that prepare the visual system for a gaze shift have been identified. However, prominent theories relate these mechanisms to corrolary discharge of the impending eye movement. This would thus relate these mechanisms to the muscles involved in the eye movement rather than a structure such as the HBPM (Cavanagh, Hunt, Afraz, & Rolfs, 2010).

In conclusion, the definitions of the components of the peripheral vision system allows one to identify various ways in which the visual executive can influence processes in these subsystems. Together with the control over the visual pyramid this defines the executive control over visual attention and the processing of visual input. This leaves one component of the architecture that is also affected by task influences: working memory.

4.5.3 Top-down influence in visual working memory

Section 4.4 has introduced visual working memory in this architecture as a system consisting of two components. The first is the storage components which is believed to be located in the visual pyramid, the second is the working memory control component, closely related to the control exerted by the visual executive, but targeted at working memory representations. An explicit dissociation between these two control components may not be identifiable in the brain, as they might originate from the same frontal processes described in the beginning of this section on task guidance. In this architecture the control components are separated to illustrate the difference in function more than an underlying structural difference: the working memory control mechanism has the function of organizing, maintaining and revisualizing working memory items in the visual pyramid. In that, it is controlled by the visual executive which may influence these representations in several ways, and the visual executive thus does not directly operate upon the memory items represented in the pyramid.

Unlike the representations in the other mechanisms in the architecture, working memory items are largely task-independent. They reflect the information gathered in past sensory experiences that can no longer be adjusted. There is therefore relatively little influence of the visual executive on working memory. A notable exception is the process of chunking, when multiple working memory items are combined in order to increase the capacity. To prevent interference, the items in a chunk shouldn't overlap too much in all feature dimensions. However, task influence could determine whether inter-

ference in certain dimensions is more important than other irrelevant feature dimensions. Another example the need for multiple working memory representations to be combined into one. For example, multiple samples of the corners and line segments of a triangle could be used to construct a single object representation of the triangle in working memory. Such operations are heavily task dependent but it is not yet clear how this is realized. The process is closely related to the notion of visual imagery, and the extent to which such operations actually rely on the perceptual systems is strongly debated (Farah, 1988; Roland & Gulyas, 1994). It seems therefore likely that such transformations may stretch beyond the scope of simple working memory mechanisms described here, and rely on higher order cognitive reasoning processes instead.

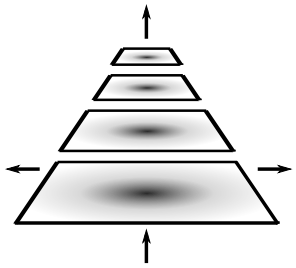
Although the task influence on individual items in working memory items may be limited, the organization of the items themselves could easily be affected by task influence. For example, section 4.4 indicated a representational dissociation between memory items that are currently being used in the task and those that are simply stored for later use. The underlying mechanisms of this prioritization is largely unknown, but they are certainly influenced by task guidance. Due to the similiarities between working memory and attentional mechanisms, it seems that similar mechanisms can be used to select, suppress, and restrict the memory items in order to prioritize the relevant item.

Thus, the precise amount of influence of the visual executive on working memory represetation and control remains largely unclear. However, as stated, working memory items would be much less dependent on task influences than the perceptual systems, and therefore it is here assumed that the influence is limited to guidance in the chunking process.

4.6 A new architecture

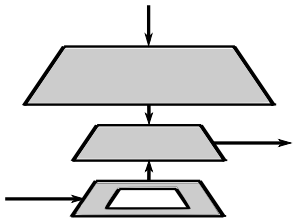
A global overview of the complete model organization as it has been described is depicted in figure 4.6. Below, the model components are reiterated, with a brief description summarizing their connections, their function, and the way these dynamic systems are influenced by the task requirements.

Selective tuning pyramid



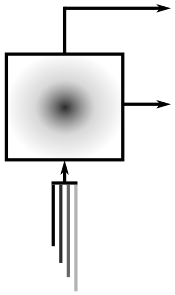
Subcomponents	feature sheets
Function	provide base rep, implement attentional tuning via recurrent localization and branch-and-bound mechanisms, vWM storage
Input	Input image
Output	Lower layers: blackboard and peripheral vision system. Top layer: VE
Task Influence	parameters that guide tuning

Peripheral vision system



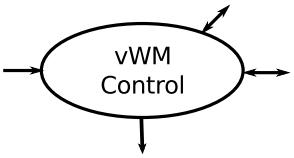
Subcomponents	Top to bottom: FHM, HBPM, PPM
Function	Implement IOR, trigger saccades, find interesting peripheral locations, vWM storage (FHM)
Input	Lower layers of the pyramid (PPM), Saccade feedback (FHM)
Output	Saccade trigger, spatial bias (HBPM)
Task Influence	Spatial bias, speed vs accuracy trade offs (HBPM), IOR shape and size (FHM)

Blackboard



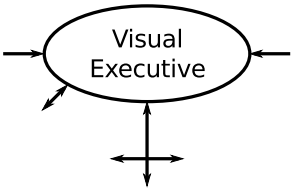
Subcomponents	No structural subcomponents are identified, but functionally information from the different layers of the pyramid should be dissociable
Function	Record detailed low-level attentional sample information for use by higher order areas (and perhaps other sensory areas)
Input	The pyramid, lower layers moreso than higher
Output	Visual executive, vWM Control, Higher layers of the pyramid
Task Influence	None identified

Visual working memory control



Subcomponents	None identified
Function	Binding-, chunking-, organizing- and (de)prioritizing working memory items; inhibiting interference
Input	Attentional sample in Blackboard, gaze information from FHM
Output	Pyramid, FHM
Task Influence	item priority, chunking candidates

Visual executive



Subcomponents	None identified
Function	Exerting task influence, sequencing operations.
Input	Attentional sample in Blackboard, gaze information from FHM, spatial 'suggestions' from HBPM, image information from top layer of the pyramid, working memory item organization information from vWM controller, Saccade information from eye movement mechanisms
Output	Task influence on the pyramid, FHM, HBPM and vWM controller
Task Influence	Method: coarse description of the task/strategy

Chapter 5

Visual problem solving using the new framework

Although the architecture illustrated in the previous chapter would provide the tools to implement visual problem solving, it still has not addressed how this is realized, as it is not yet discussed how the processes in each component are recruited and cooperate to solve the present visual task. The task-influence discussed in the previous section is an important aspect in visual routines theory as a framework for visual cognition, as a flexible architecture that can be adjusted to meet task demands allows for this system to provide a general solution to problems of visual cognition. However, it only addresses single 'steps' in the cognition process, whereas visual routines theory is characterized by the notion that they are composed of multiple operations that solve components of the task. To that end, it should be detailed how the architecture (a) combines the information extracted or constructed in multiple steps, in order to draw conclusions based on the available representations; (b) organizes the sequence of operations in order to lead to a solution of the main task. In the classical formulation, the first issue was addressed by introducing the concept of incremental representations, the second issue defines visual routines themselves. Here it will be discussed whether the same approaches can be applied to the current architecture.

5.1 Combining information: incremental representations

Although the classical visual routines framework did not include an explicit working memory model, it did acknowledge the issue of communicating the result from one elemental operation to the next. To this end, the concept of an incremental representation was introduced. The first elemental operation in the routine would construct an incremental representation using the base representation as input, which would then be provided as input to

the next operation in the routine. This mechanism largely originates from the assumptions made about the base representation: a static, accurate and complete representation. The problems with the base representation were discussed in the previous chapter and formed one of the most important reasons for the design of the new framework. This raises the question whether there is room for incremental representations in this framework, and whether they are still needed for visual problem solving. Their role in the classical framework can be summarized as: (1) representing the result of elemental operations, (2) transferring these results throughout the task, (3) triggering the appropriate operation given the results so far. The current architecture explicitly abandons the concept of a base representation as presented by Ullman (in section 4.1), and for the same reasons incremental representations that suffer from the same fallacious assumptions must be abandoned. However, it seems that their functionality is implemented sufficiently by the other components. First of all, during the attentional tuning process feedforward activation is modulated. Although the result – an attentional sample – is very different from the visuospatial markers that govern Ullman's incremental representations, it provides an isolated representation of the information in the scene that is relevant for the task. The classical formulation doesn't detail how the information is maintained or communicated next, but it appears that the information is available as a generic representation which is accessible for all following operations to read and write on. In the new framework the attentional sample is transcribed to a blackboard and from there on communicated to higher order areas to trigger working memory control and visual executive functioning and perhaps influence other sensory systems. A key difference however is that this is a passive system that can not be modulated by other operations. At later stages of the task, the relevant attentional samples are maintained as working memory items. The way that both working memory items and current attentional samples can be used to infer reasoning steps in visual cognition provide another similar functionality as an incremental representation, albeit in a decentralized fashion. A final tool of this architecture to implement incremental representations is by chunking in working memory, which allows information from multiple working memory items to be combined into one.

Thus, although the current architecture does not provide an explicit mode of implementing incremental representations, it seems that the functionality involving these representations is provided through the multiple components of the model. Therefore, the notion of an explicit incremental representation seems obsolete.

5.2 Sequencing into visual routines: methods and scripts

An important advantage the visual routines theory provides for modeling studies and computer vision, is that the general approach to any visual task is to string together *discrete* operations. Provided with the appropriate operations, visual cognition can be described by linear execution of this sequence of operations, much like the instructions in a Turing machine (Turing, 1950). One of the major complications considering the biological plausibility of this approach however, is that the brain is not organized as a Turing machine, which means that operations will not be discrete and allow for simple linear execution. This leads to the following questions: how are these 'steps' implemented by the brain, and how is their sequence determined?

The previous section has already introduced the recent proposal of Zylberberg et al. (2011), which describes how cognitive steps can be implemented by the properties of neurons in the PFC. There are neurons with sensory properties, which can be used to indicate the simple presence of a sensory stimulus. However, mnemonic neurons will, after being triggered by these sensory events, persist their firing for a longer amount of time. This property can be exploited to use these neurons to code 'collecting evidence'. When enough evidence is gathered, these fire production rules, a process which is implemented by the triggering executive neurons. This could then lead to motor actions, sensory adjustments, or new evidence gathered for another rule. It may seem that this scheme would require an implausibly large number of neuron ensembles to code for all possible production rules that are part of our cognitive abilities. However, what PFC neurons code for seems to be heavily dependent on the current task requirements. In a related fashion, with sufficient practice these neurons can learn general rules that are applicable to certain tasks, so that less evidence needs to be gathered and the active involvement of PFC neurons becomes less (E. Miller, 2000).

These properties of PFC neurons are closely related to a scheme for problem solving – both in the visual domain and in general – proposed by Tsotsos (2010) of *methods* and *scripts*. A method can be seen as a universal approach to a task or problem, much like the universal routines introduced by Ullman, which can be applied without any task information, but can be tuned into a script by task influence and sensory- or motor feedback. Tsotsos describes four classes of methods and scripts; task, sensing, motor, and reasoning. The task-class recruits methods and scripts from the other three classes and thereby describes a strategy to solve the task. A task method describes the general approach to solve a task using sensory-, motor- or cognitive processes. These methods accept information about the task to form an appropriate task script which in turn tune the other methods into appropriate scripts. A sensory method describes the default approach to the

sensing aspect of any task, whereas the script applies the appropriate bias or other tuning mechanisms. A motor method describes a ballistic movement of an agent or a component (for example, an eye) from point A to point B, whereas a script could impose a path or avoid obstacles. Reasoning methods describe how to answer questions about stimulus size, location, shape or spatial relations given the sensory- or motor information, and again scripts are tuned to the task at hand.

This scheme can be used to define an organization of elemental operations in order to define strategies similar to visual problem solving much like visual routines, but with two important advantages over the classical approach: the scheme is not limited to visual operations – so it can be used to describe visuomotor routines – and the strategies in problem solving are not necessarily rigidly defined by the task. When applying this scheme to the visual problem solving domain this scheme can be used to control the components in the framework proposed above as follows:

- *Sensing* The sensing processes in the framework are implemented by the the visual pyramid and the peripheral vision system. The different task influences that can be used to tune the method into a script have been illustrated in section 4.5. When they are recruited in a method there is no explicit task influence and these parameters need to take on their 'default' value. The following table suggests default values for the parameters in these components.

Parameter	**Default value**
Bias and Gating	No bias or gating inhibition
θ, μ, ζ	Must suit the scale of the stimulus: either gained through experience or last used values
IOR in FHM	IOR at the fixation location, gaussian-shaped with size related to θ, μ, ζ, which lasts for about 1 second
HBPM computations	unbiased integration of information from FHM and PPM, direct translation of map coordinates to suggested saccade (would lead to center-of-gravity effects)
Working Memory Control	No active maintenance of memory items, so all attentional samples are stored but will interfere and items are easily lost. The most recently attended item is prioritized.

It must be emphasized that these values are only used to illustrate the concept of a default value in sensing methods. Most factors, in particular the properties of IOR, will require further research to find which values best capture the default sensing properties.

- *Motor* The framework is targeted at visual problem solving, and although the importance of motor processes in active vision is acknowledged, motor processes are only sparsely represented by the framework. The system as depicted in figure 4.6 does illustrate a saccade system, but an extensive discussion on eye movement control stretches beyond the scope of this thesis. Still, motor methods and scripts can be dissociated within the context of eye movements to the extent as discussed here. In a task where eye movements are required, a motor method can be defined as straightforward execution of the HBPM output as a ballistic saccade. In a tuned method however, eye movements can be suppressed when necessary, or guided by reasoning processes instead of the sensory output. For example, in an antisaccade task, the cue from the sensory system can be translated into a saccade in opposite direction (e.g. Roberts, Hager, & Heron, 1994). Another example is when tracking multiple objects, where the accuracy of the representation of each moving object in working memory might impose fixating an object that might otherwise be lost.

- *Reasoning* For visual cognition, reasoning methods and scripts are essential as they are involved in two functions. The reasoning methods describe how the the sensory and motor processes could be used to solve a task, thus how the collective of these processes translates to a solution to the task. For example, it describes that a single feedforward pass through the system should be enough for a simple detection task. The reasoning script however, also details how the sensory input may require further tuning in order to appropriately guide the sensory or motor scripts. For example, a reasoning script can use the size of the stimulus as determined by the initial feedforward pass to adjust the parameters (θ, μ, ζ or IOR shape and size) and solve the task. In the new framework reasoning is implemented by the visual executive.

  As section 4.5 indicated, these reasoning processes may be best modeled by production rules which may integrate information from the current task (for example: find the red cross, track the blue ball), the current motor scripts (for example: the fixation location, the planned trajectory) and the current sensory output (for example: an attentional sample in the blackboard constructed using the current parameters and peripheral cues in the HBPM) in order to produce an appropriate action (for example: return the location of the red cross, adjust θ, change the planned eye trajectory). Another important source of information for the reasoning processes that has not yet been explicitely mentioned is the content of working memory. As the previous section illustrated, working memory items form a crucial component in modeling incremental representations. For example, in reasoning about visuospatial relation, the location of previously attended components

of the stimuli are needed to infer the answer to a visual task.

For a better understanding of how these methods and scripts can be organized in order to detail the approach to visual problems using the framework proposed here, section 5.3 discusses three example problems that apply the scheme discussed here.

5.3 Examples

5.3.1 Curve tracing

One of the characteristic tasks presented in Ullman's visual routines paper is the study of curve tracing. Tasks such as the one presented in Figure 5.1 were used to illustrate the use for a curve tracing operator in visuospatial analysis. Since, interesting characteristics of such curve tracing tasks have been discussed (Jolicoeur, Ullman, & Mackay, 1986; Roelfsema et al., 2000). The reaction time in determining whether two markers (here a and b) are on the same curve seems to depend on the euclidian distance between the markers, but also on the length of the curve between them, and on the properties of the curve. Curve segments with high curvature or proximity to other image elements (such as the curve segments near the c in figure 5.1) seemed to slow down the search. Jolicoeur et al. (1986) attribute this to the size of the attentional spotlight, which needs to be narrowed when other line segments would otherwise be included. The object-based attention approach of Roelfsema et al. (2000) explains this by the scale and level of detail at which the attentional label can spread: when no other curve segments are close, a much less detailed analysis of the curve is necessary. However, the issues with these approaches to curve tracing have already been discussed in sections 4.2 and 4.3. Here it is described how the curve tracing task could be solved using the new architecture and present these characteristics. Then, a description of a possible task approach is given in terms of methods and scripts.

Before the curves are presented, the observer is assumed to know the characteristics of the markers which allows for easy search. This would impose bias on the pyramid for features matching the search template of a cross. When the trial is presented, the first step in the analysis will depend on the configuration of the display. When one of the crosses falls within central vision, it will be attended and construct the first attentional sample. If the crosses fall in the periphery, they will capture attention due to their inherent saliency, boosted by their task relevance. Although it may seem sufficient to pick and attend only one marker, it is more likely that, before curve tracing starts, both markers are attended and stored in working memory, so that the end point can be maintained during the task to monitor progress. Attentional samples (and therefore also working memory items)

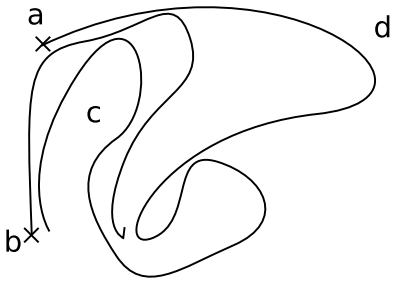


Figure 5.1: A curve tracing task as described by Ullman (1984) and Jolicoeur et al. (1986). The task requires determining whether the two markers are on the same curve. (Letters are not part of the actual stimulus)

of the markers would involve extracting their spatial location, as well as a segment of the curve they are on. After these have been stored, one of these markers will be selected to start tracing. This choice could be arbitrary but is likely influenced by experience from other trials, or forms of compositional bias, such as the bias to start at the top left of an image or page originating from conventions in reading in the observers culture. Here, it is assumed, curve tracing will start at marker a.

Tracing a curve without any interfering segments, such as curve section d in figure 5.1 should be relatively fast. In the architecture, the combination of the functions from central and peripheral vision could establish quick analysis. The feedforward activation in the pyramid established by these segments should provide cues that there are indeed no interfering curve segments there, so that no complete top-down tuning pass of the visual system would be required to extract detailed information of the curve. This is very similar to the proposal of Roelfsema et al. (2000) for tracing such sections at map with a coarse representation. At each fixation point, the PPM would highlight two interesting locations for a gaze shift; at both ends of the currently fixated curve. The HBPM will trigger a saccade corresponding to the tracing direction, based on top-down bias from interpretation of the FHM indicating the path traced so far, and perhaps also by simple IOR from the FHM itself (even though this is not a standard search array usually associated with IOR). A potiential sequence of samples is depicted in figure 5.2. In curve tracing, detailed storage of these samples will not be necessary. It is more likely that the entire curve segment is chunked into a low-resolution representation of the curve segment with several orientations present in the segment and its average spatial information (Zhang & Luck, 2008; Brady et al., 2011).

Tracing a curve segment such as the one indicated by c (figure 5.3a) requires more intensive involvement of the attentional mechanisms. As figure 5.3b illustrates, before attentional tuning the activation at the top level of

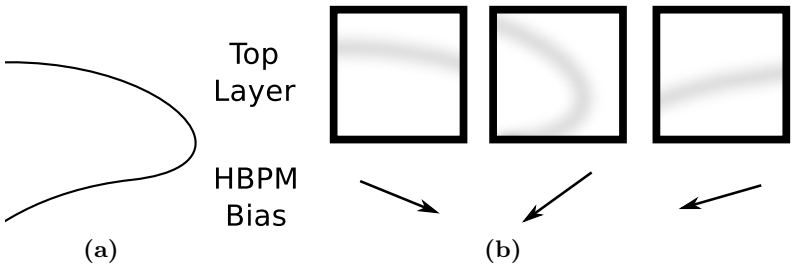


Figure 5.2: (a) curve segment d from the trial depicted in figure 5.1. (b) Sequence of steps needed to trace this curve segment. Low resolution representations at the top layer of the pyramid provide enough information to signal no extensive tuning is necessary. The arrows indicate the bias on the HBPM for the next fixation. This bias does not originate from analysis of the fixated curve segments, but from the saliency of the rest of the curve in the PPM (not depicted) plus the fixation history information in the FHM.

the pyramid does not provide enough information to reliably trace a single curve. Also, peripheral vision does not provide sufficient information to guide tracing. Attentional mechanisms can however be used to select the curve, suppress the distracting curves, and extract the information from the curve needed to select the next fixation location. The attentional sample that focuses on the right curve can be constructed using the information from the last fixated location to bias for the correct curve segment, and impose connectivity constraints for the sample to not include the other curves. This will most likely involve a complete feedforward- and feedback pass through the pyramid. Once detailed information has been extracted from the pyramid it can be used to infer the next direction of the gaze shift.

Although these two mechanisms describe very different ways of curve tracing, there may not be an all-or-nothing dissociation for either mechanism; The initial top-down activation of the scene may determine the amount to which detailed analysis of the curve is necessary, and to what extent simple saliency suffice to accurately trace the curve. The alterations between the two mechanisms may thus be of a more dynamic nature, depending on the amount of reliable evidence that can be gathered. After the curve has been traced there are several ways the system can derive a conclusion. The most straightforward way is when the second marker may not have been 'encountered' on the curve before the end-point is spotted. However, in some cases fully tracing the curve may not be the only strategy that can be used, as an earlier conclusion can be drawn based on reasoning – for example, when the curve directs away from the stored location of the second marker, and there are no signs that it will eventually connect with the curve of the second marker. Another potential strategy is to attempt to trace both curves and as soon as one of them connects up to the marker one can draw a conclusion. However, whether such strategies are applied seems a matter

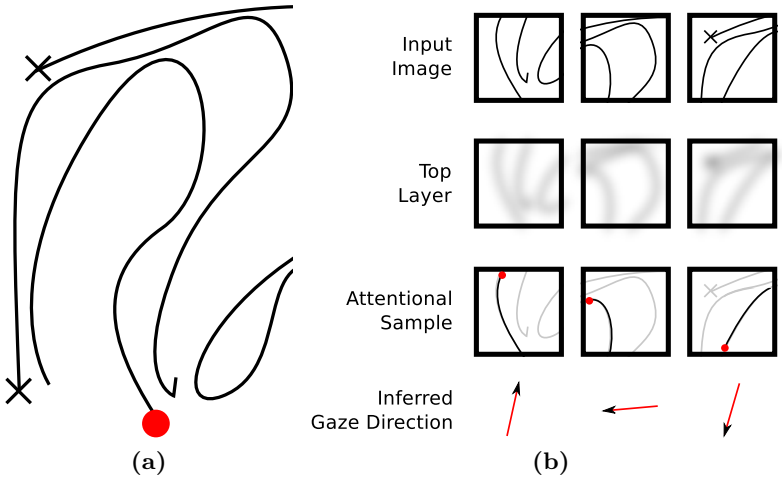


Figure 5.3: (a) curve segment c from the trial depicted in figure 5.1. The red dot marks the latest fixated location. (b) Sequence of steps needed to trace this curve segment. The input images at certain fixation locations are too incoherent to provide enough information at the top layer of the pyramid about the current curve segment and others, hence attentional sampling is needed. Also, the cluttered scene analysis makes it impossible to rely on the peripheral vision system to determine the next fixation. Therefore, the orientation at the locations marked by red dots in the attentional sample is used to determine the next gaze shift.

of higher-order cognition and not of this architecture.

From these steps, the approach to the curve tracing task in terms of methods and scripts, can be derived. These will be described in terms of a qualitative description of the task method, and a more formal definition of the production rules that govern these methods. The whole task can be divided into two subtasks: visual search, which is used to find the two markers, and the curve tracing itself. Visual search is a process that would normally require little involvement from reasoning influences and is only governed by the target properties, yet here it is necessary to monitor whether a marker has been found and store its location in working memory as a prioritized item that should not be forgotten, as well as to count the number of found markers to determine whether both have been found. This can be modeled by a production rule as follows:

```
IF blackboard sample matches the marker
  markers found + 1
  prioritize location in working memory

IF markers found equals 2
  stop search
```

The actual tracing process itself requires more rules for reasoning. The task method will include three reasoning steps: to the initiate of the tracing

process, during the tracing process itself, and one to infer a conclusion. To initiate the tracing process, the marker samples in working memory need to be compared and the most suitable starting location must be defined. As indicated, there are various factors most likely based on experience that govern this comparison, but the production rule can be as follows:

```
IF working memory sample 1 >* working memory sample 2
  move eyes to location of working memory sample 1
ELSE
  move eyes to location of working memory sample 2
```

Here, >* indicates a comparison as described above. In this rule, the instruction to move the eyes refers to a motor script which is tuned to direct the gaze to the appropriate location. The tracing process itself requires reasoning about to let the appropriate representation guide the fixation, and to construct attentional samples where necessary. Also, it requires constant monitoring whether the target marker or the end of the curve have been reached or not.

```
repeat
  IF top layer ~ marker
    AND FHM ~ location that equals the marker location
  OR top layer ~ endpoint
    construct attentional sample
    stop tracing;
  ELSE IF top layer ~ one curve
    move eyes based on HBPM activity
  ELSE
    construct attentional sample of the traced curve
    move eyes based on the extracted direction
end repeat
```

The symbol ~ included in this rule can be read as "indicates", where the presence of features in the top layer of the pyramid can signal whether the marker or an endpoint is found, or whether there are multiple curves or just one at the current fixation point. As can be seen, the process will repeat untill an endpoint has been reached or the target marker has been found. At that point, the final reasoning method will be able to draw a conclusion, based on the current attentional sample of the curve and the stored sample of the target marker.

```
IF attentional sample ~ end point
  return ''different''
ELSEIF attentional sample matches working memory sample
  return ''same''
ELSE
  return ''different''
```

This example illustrates how the approach to the curve tracing task illustrated above can be modeled as a methods. The execution of the taks including sensory feedback will lead to a script that will apply to the task at hand. The task method indicates the structural organization of the task

into its separate subtasks. The rules in the reasoning methods indicate that the visual executive interprets the sensory components and based on their activation the appropriate sensory and motor processes are initiated. In this particular task, all sensory and motor processes are 'wrapped' in reasoning methods they require interpretation and guidance from the executive in this task, but this does not need to be the case.

5.3.2 inside/outside relations

Another slightly more complex task that Ullman describes as prototypical for the visuospatial domain, is identifying inside/outside relations. The goal of these tasks is to determine whether a marker lies inside or outside of a contour (figure 5.4). At first glance this task seems very similar to the curve tracing task of the previous section, where the difficulty (and therefore response time) of the task is determined by the curvature and length of the curve. Similarly, the trial depicted in figure 5.4a is relatively easy compared to the one in figure 5.4b, and the curve in 5.4b is defined by similar properties as hard curves in the curve tracing task. However, as the trial in figure 5.4c illustrates, the placement of the marker in the contour may be much more important than the boundaries of the contour. Because the task involves complete contours, visual analysis of the heavily curved part of the contour is unnecessary to determine whether the marker is inside or outside.

Ullman proposed two alternatives to the curve tracing approach. The first may be called 'ray intersection', which he described as shooting a ray from the marker to a 'point at infinity' outside the contour, and counting the number of intersections. An odd number implies that the marker was outside the curve, an even number means outside. This method would lead to a similar dissociation as with the trials depicted in figure 5.4, but Ullman stated that this method was implausible as it would only solve the problem correctly if the curve is isolated, and no other interfering lines would be present in the stimulus (figure 5.6). Instead, the 'coloring' or region filling method was introduced, where an elemental operation can spread activation through the object, bounded by the curves that constitute the contour. Once the contour is completely activated, it can be determined whether the marker location is also activated, which means that the marker lies inside of the curve. Ullman considered this approach not only the most plausible solution to the inside/outside task, but also considered it such a characteristic operation for visuospatial relations that it was defined as an elemental operation. A possible implementation for this elemental operation was also described in Roelfsema et al. (2000)'s proposal by means of attentional label spreading.

Important issues with the coloring method and its implementation via attentional label spreading, have been discussed in section 4.3. To summarize, an attentional sample seems incapable to fully represent an object through

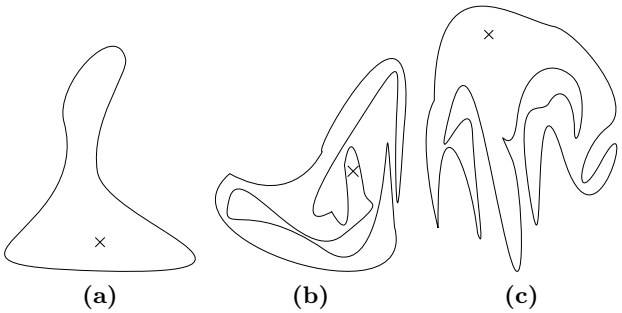


Figure 5.4: Example trials of the inside/outside relations task where the observer is to determine whether the marker lies inside or outside of the depicted contour. Interestingly, this can be determined relatively fast in trials (a) and (c), whereas the task seems much harder in trial (b).

the simple spreading of activation because of inhibitory effects in the visual pathway and the limited resolution of visual information. Nevertheless, this does not imply that the idea of a coloring method to obtain an object-based representation of the contour should be completely discarded, as long as its capacities are not overestimated. Within the new framework this could be implemented as follows. Upon stimulus presentation, a very coarse representation of the contour is extracted, to obtain spatial information about its outer boundaries. Next, the marker is searched for by means of visual search strategies. Next, as with curve tracing, attentional samples are constructed of contour segments, in order to deduce the next region of fixation in the 'outwards' direction. The shape of these segments is determined by the boundaries of the curve, which indicate the directions of the next trace. This process continues until the outside boundary has been reached (figure 5.5). Important in this strategy are a basic understanding of the concept of analyzing a contour in 'outwards' direction, and the knowledge that the outer boundary of the contour marks a clear distinction between the outside and the inside area. This is knowledge that is most likely not part of visual cognition but a product of higher-order reasoning and experience, as none of the components of the visual system has the capacity to represent this knowledge

The notion that higher order cognition can provide knowledge to interpret the shape analysis from the visual system as an indicator for what areas are 'inside' or 'outside' is interesting, as it would not only support the theory of inside/outside analysis through shape analysis, but it would also support an approach very similar to the ray intersection approach. A very effective strategy for this task would simply be to focus on an outside point, preferably close to the marker, and than gradually shift attention towards the marker, and interpreting every detected boundary as a switch

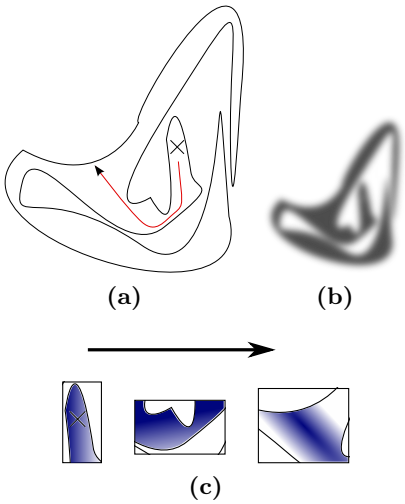


Figure 5.5: Solving the inside/outside task using a shape analysis approach. (a) The spatial trajectory of the sequence of attentional samples, from the marker to the outside boundary. (b) Very coarse representation of the shape of the contour. This is only used to obtain an initial gist of its outer bounds. Gray shading has been use to indicate the inside area of the contour. (c) Attentional samples used to analyze whether the marker is on the inside or outside of the contour. The blue shading is used to indicate the information deduced of the shape. The gradient indicates that not the entire width of each segment needs to be included in detail: if the attentional shifts move outwards, largely based on a single curve only, information about the other side of the segment is largely irrelevant. The gradient in the final attentional sample indicates how, as with curve tracing, shape analysis can be based on coarse representations when no obstructions are present and no detailed information is needed.

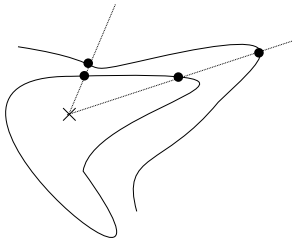


Figure 5.6: An inside/outside trial that can not be solved by a straightforward ray intersection approach. The additional curve makes the number of intersections with the ray unreliable to deduce a conclusion.

between 'outside' and 'inside'. Although some of Ullman's objections would still hold, if the constraint is met that every line in the image is part of the contour, this would provide a failsafe approach to the task.

However, there are reasons why human observers might tend to use the strategy based on shape analysis instead. First, shape analysis may be a strategy that is chosen largely automatically as this task is much more prevalent in everyday life. For example, shape analysis plays a significant role in object recognition or discrimination, but also in planning motor actions for interaction with objects. Moreover, ray intersection without any further analysis of the image is as mentioned only guaranteed to succeed when all the lines in the image are part of the contour, which will only hold for such artificial stimuli (see figure 5.6). Second, perhaps the most crucial difference between the shape analysis approach and the ray intersection approach, is that shape analysis is largely solved by the visual system alone, as it is only marginally influenced by higher-order cognitive processes to direct attention towards the outside boundaries and for deducing the final conclusion. The ray intersection approach however, highly relies on cognitive reasoning, where either all intersections need to be counted and categorized as odd or even in number before a conclusion can be drawn, or every intersection requires reinterpretation of the figure observed so far, switching between an 'inside' or 'outside'. As studies of reinterpretation of ambiguous stimuli indicate – e.g. the Necker Cube, (Pelton & Solley, 1968) – such operations might be expensive which would also render a purely visual approach preferable.

Unlike for the curve tracing task, there appear to be no studies targeted to elicit the properties of the inside/outside task, which makes it difficult to make predictions on the underlying strategies used by observers. Here, two approaches were illustrated based on Ullman's discussion of the task, but there might be other plausible strategies involved. Also, one could envision a mix between the two strategies, where depending on the observers proficiency with this particular task they might switch from shape analysis to a ray intersection approach once the 'outer parts' of the contour are reached. Here, a task method will be defined based on the mixed approach. This

strategy illustrates a task method consisting of four subtasks. (1) extract a coarse representation of the stimulus and store it into working memory, (2) find the marker and fixate, (3) extract attentional samples of the contour and infer stepwise gaze shifts, (4) infer a conclusion based on the number of curves.

The first two tasks require very little reasoning and can likely be resolved by task-directed sensory- and motor methods. A purely sensory method can extract a coarse contour representation and store it in working memory. Task influence can direct these processes, for example to establish prioritization of spatial features of the working memory representation of the contour and assure that they will be remembered during the task. The second task is a simple visual search process which can be solved without any reasoning by biasing the visual system for the target features. Untill the marker is found, the eye movements can be governed by HBPM activation. The only point in the search process where reasoning could be introduced is to deduce when the marker has been found and can be fixated. The rule governing this descision is relatively straightforward:

```
IF blackboard sample ~ marker
  fixate
  stop search
```

The next subtask is largely governed by reasoning processes. The reasoning process largely resembles the curve tracing process; again, a path of gaze shifts needs to be determined, either from the coarse feedforward representation or by interpretation of a detailed attentional sample, until a stop condition is reached. However, this stop condition is less rigidly defined than in the curve tracing task, as it is the point where the system switches to the ray intersection strategy as discussed above. The appropriate conditions to start the ray interesection method will largely depend on training and experience. However, it appears that two important factors in this rule would be the distance to the outside boundary of the contour in the coarse representation, and the number of curves separating the fixation from the outside region – which can not be known without attentional sampling but is indicated from the amount of activity in the HBPM. These rules together would indicate that the production rules that govern the tracing process can be organized as follows:

```
repeat
  IF the distance fixation location to outside of the
  contour sample in working memory < threshold
  AND HBPM ~ suitable
       stop tracing
  ELSE IF  top layer ~ little curvature in the shape
    move eyes based on HBPM activity
  ELSE
    construct  attentional sample of one boundary of the contour
    move eyes based on the extracted direction
```

```
end repeat
```

The final subtask is to derive a conclusion based on the number of curve segments that separate the current fixation point from the outside region. It must be noted that there might also be none when the rule to stop tracing might not fire until a point has been reached where the outside boundary is missing (which would indicate the marker and every point on the traced path has been outside of the contour). However, when there are many curves still separating the fixation point from the outside region, every curve will require attentional sampling in order to be counted, and gaze shifts may be needed to reach the outside region. This illustrates a reasoning method organized like the following

```
set count 0
repeat
  IF fixation location falls in the outside region
    IF count is odd
      return ''inside''
    ELSE
      return ''outside''
  ELSE
    construct attentional sample
    add the number lines in the attentional sample to count
    move the eyes outwards in a straight line
end repeat
```

Again it must be emphasized that this task method is constructed for illustratory purposes, as the lack of a study that has thoroughly investigated the task characteristics makes it hard to determine the actual underlying strategy for human observers. Nevertheless, this example illustrates a strategy for the inside/outside task that again would account for the differences in difficulty or response time between trial types. The difficulty appears to be largely determined by the share of the problem that can be resolved without attentional tuning, and the approach to solve both tasks is relatively similar. Therefore, the next example will focus on a task that is not from the class of visuospatial problems as illustrated by Ullman.

5.3.3 Object tracking with occlusion

Possibly inspired by the set of visuospatial tasks presented by Ullman, many systems based on visual routines focus on visual tasks in static images. A notable exception is the system of Rao (1998), which solves several tasks in dynamic natural scenes. As an example using moving real-life images, Rao constructed a visual routine to solve the problem of tracking a rolling ball while its trajectory is partially occluded (figure 5.7). When the object is occluded by another object, the focus of attention moves to the other side of that object where the ball is expected to reappear, and when it does the system continues tracking from there. In Rao's framework, all visual

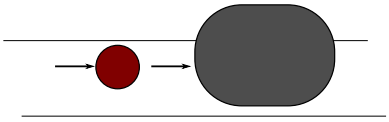


Figure 5.7: Rao's ball tracking task. The red ball in the image rolls to the right. The gray object is in front of its trajectory and will temporarily occlude it untill it reappears on the other side.

routines are composed of cycles of *selecting* the focus of attention – *moving* the focus of attention to that location – *establishing* the features at that location. The visual routine with this approach is described as follows:

1. *tracking the ball* - With the ball at the focus of attention, tracking is a mechanism that allows to move the focus of attention along with it.

2. *Lose the ball* - As the ball gets occluded, the focus of attention is at the edge of the occluding object. At this point, the occluding object can be attended, and its properties can be extracted.

3. *Saccade to the opposite edge of the edge* - In order to get the ball back at the focus of attention, a saccadic shift of the focus to the opposite edge of the object is necessary. This can for example be achieved by extracting the edges of the object, and deducing the opposite edge from the motion of the ball.

4. *Wait for motion* - As the focus of attention is at the edge where the object should be reappearing, the system can wait untill its motion is detected again to continue the trace

It is stressed by Rao that this approach to the task is a mere illustration of the capabilities of the elemental operations that were defined for the system (saccades, tracing, establishing edges), and how simple mechanisms such as the routine could underly complex cognitive constructs such as 'expectations'. However, the approach of attention as a purely spatial selection mechanism may have limited the system and may have forced a more complex approach to this problem compared to a strategy allowed by the framework presented here.

A remarkable aspect of the routine illustrated by Rao is that the first step of the tracking task seems to be absent from the routine, which is the initial search for the object. Visual search has been extensively discussed for the new framework, and the same mechanisms and cues that can be used for search may be of use for the tracking task. The term 'tracking' has several interpretation. A simple interpretation of 'tracking' would imply that the object is to remain at the focus of attention, although it is moving. This could be achieved by a constantly updated search process. A more

complex interpretation of tracking would imply that motion information is extracted from the stimulus and is used to construct a rich representation of it. The motion pattern of the rolling ball can be extracted using both motion features in the visual pathway, and the fixation history in the FHM. This pattern can be used to extrapolate its trajectory, and can be used to facilitate search whenever it may have lost attentional focus; the power of representations using motion was already illustrated discussing Cavanagh et al. (2001)'s sprites in section 3.3.

In the particular task illustrated here, the object *will* indeed lose attentional focus due to the occluding object. The framework then allows for three ways to facilitate tracking after the object has been lost (see figure 5.8). First, visual search should be relatively easy as it can be biased using a rich object model including a spatial bias for the expected location of the occluded object that can be inferred from the extracted motion pattern. Second, an approach similar to Rao's routine can be used by attending the occluding object and inferring the edge or boundary where the ball is expected to reappear. Third, even without using any of the knowledge about the ball's features extracted before occlusion, the sudden onset of motion at the location where the ball reappears is likely to capture attention of peripheral vision.

The tracing task was introduced by Rao to illustrate the power of his relatively simple framework of directing spatial attention. The new framework implements Rao's operations, albeit somewhat differently, and it is therefore not surprising that the same strategy can be realized using the new framework as well. However, the new framework also allows for much richer representations, and the above section shows that this allows the task to be solved using only the very basic operations of visual search and binding object representations to implement tracking, even with the extra difficulty of partial occlusion of its trajectory.

Because the task is governed by mechanisms of visual search and will only recruit mechanisms of reasoning extensively once the object is out of sight, the definition of its method is fairly straightforward. It consists of two subtasks: 'searching and tracking' for finding and tracking the ball while it's visible, and once it is no longer visible 'expectation' in order to easily resume tracking. In searching and tracking, the only aspect of reasoning is that the motion in the attentional sample can be used to guide the eye movement, and to signal whether the object is hidden or not. The rules in this method are as follows

```
IF blackboard sample ~ target
  extract motion from blackboard sample
  move eyes according to the motion
IF top layer ~ target lost
  prioritize last sample in working memory
  stop searching and tracking
```

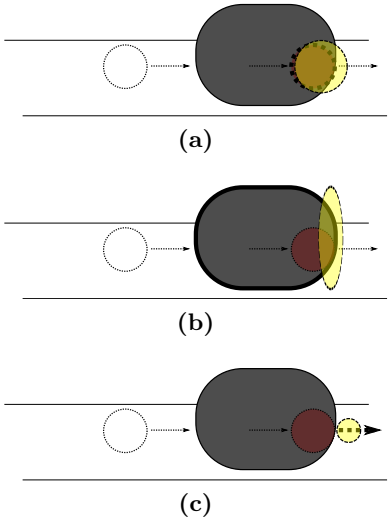


Figure 5.8: Three cues that can be used to guide attention in order to resume tracking of the red ball. In these images the location of the ball behind the gray object is indicated by the transparent red circle, and it is about to reappear. The yellow shade is a simplified spatial representation indicating the attentional guidance by each cue, but it is again emphasized that the mechanisms establish more than simple spatial bias. (a) The trajectory of the ball (stored in the FHM) and its motion features (stored in the working memory representation) can help predict when the ball is supposed to be during occlusion. (b) The size and shape of the occluding object can be extracted by attending it, which can be used to determine the location where the ball is supposed to reappear. (c) Once the ball reappears, it produces a sudden onset of motion in the retinal image, which will capture attention in the periphery due to its inherent saliency.

Note that as the object is no longer visible, the last attentional sample is to be maintained in working memory, as it is used during the expectation subtask. This task is organized with two rules that implement the three cues discussed above

```
IF working memory item ~ target
AND top layer ~ target lost
  extract motion pattern from working memory
  bias pyramid and HBPM for this motion

  extract location from working memory
  construct attentional sample of occluding object
  bias pyramid and HMBM for the reappearing edge of the object

repeat
  IF working memory item ~ target
  AND top layer ~ target lost
    extract motion pattern from working memory
    extract location from working memory
    bias pyramid for location + motion * time
  ELSE
    stop expectation
end repeat
```

This example would be interesting for the same reasons pointed out by Rao: what may appear to be largely cognitive mechanisms – such as reasoning about an occluded object and expectations regarding where it may reappear – can actually be modeled using relatively few cognitive mechanisms and largely relying on the attentional mechanisms. Moreover, even tracking the object can be implemented using the same mechanisms as for visual search. However, it should be mentioned that by attending the object and constructing attentional samples, the system implements both the simple and the complex interpretation of tracking. One of the uses for a complex interpretation – constructing a rich object model can immediately be seen when it is used to infer the cues that will guide attention to the location where the object will reappear.

Chapter 6

Discussion

In this thesis, an elaborate modern interpretation of the visual routines framework has been presented, and it has been shown how this can be used for visual problem solving. First, the classical visual routines framework was introduced within the context of the classical views of vision and attention. The potential of the framework as a visual problem solver and as a mode of implementing task-directed vision was illustrated by a review of various implementations of the framework that provided more insight in its components. Next, the modern understanding of vision and visual attention was reviewed, illustrating how several classical assumptions and hypotheses on how these mechanisms work in the brain do not hold. Furthermore, a modern functional role for visual attention was discussed, and its various underlying mechanisms were detailed. The classical framework and these components were then subjected to a functional analysis and critically reviewed with respect to this modern view on vision and attention. For the separate functions of the different components of the model revisions were proposed that would comply with modern theories but provide similar functionality. This has been used to adjust the model appropriately, and thereby proposing a new framework for visual cognition. Finally, it has been discussed how this framework could be controlled to solve various visual problems, organizing executive operations using methods and scripts.

This section will illustrate how this thesis affects vision research. First an overview of the important changes to the new model with respect to the classical theory will be given. As will be illustrated, this does not only affect the theory of visual routines, but these changes also illustrate a new definition for visual cognition. Then, the predictions provided by the new model will be summarized. Finally, this is used to illustrate how the new framework provides guidelines for future work in vision.

6.1 A new framework for visual cognition

The new framework differs in many ways from the classical theory, but the alterations can be roughly divided into three classes: (1) reconsiderations of the components, (2) the structural organization of these components, and (3) the way visual routines are formulated.

The reconsiderations of the different components in the framework is perhaps the most natural change that follows from new developments in vision research. The classical model described only two components: one for sensory processing, which was assumed to be largely similar to Marr's primal- and $2\frac{1}{2}$D sketches, and one to implement visual routines and connect higher-order cognitive processes with the low-level sensory processing. Novel findings indicate more elaborate sensory components, and thus the updated framework consists of several components which can account for these properties. The visual pyramid, the peripheral vision system and the blackboard together model a broad variety of findings in vision research that were discussed in chapter 3. Moreover, the working memory control component imposes additional tasks for these components as they are used to represent memory items as well. Because these components provide such a widespread representation of the sensory information that is to be used for visual cognition, the visual executive has a more integrative role than the visual routines processor in the classical model; it should no longer only execute the visual routine based on the task demands, but account for the integration of signals coming from all these various components, as well as provide appropriate feedback to all areas to realize elemental operations.

Because visual processing is distributed over various components instead of one sensory system, the structural organization and interactions of these components is an important aspect of the new framework. The connectivity pattern illustrated in figure 4.6 illustrates closely connected system, where many components are dependent on eachother's output. However, it should be stressed the design of this framework in chapter 4 was incremental, which illustrates how the architecture could easily be expanded by including additional components that would capture different aspects of vision, should the components introduced here not suffice. More importantly, the framework is also suitable to model interactions with other sensory modalities or motor programs. Aside from the the low-level representation of the visual input that is communicated from the lower layers of the visual hierarchy to the other visual areas, most connections within the framework represent bidirectional communication between the components and the visual executive. The description of visual executive functioning, reflects this communication pattern: the executive integrates information from the various sources and in response provides feedback to the sensory systems in the framework. If one were to add components from other modalities this would most likely use the same connection scheme for executive control. One could also envi-

sion additional connections. For example, these components could interact with the blackboard representation to implement multisensory integration, which could occur early after stimulus processing (Schroeder & Foxe, 2005).

These two adjustments to the classical framework already indicate that the implementation of visual routines in this framework differs significantly from the original interpretation. In the classical framework, visual routines are programs that consist of ballistic sequences of operations executed by the routines processor in order to solve a particular task. Visual executive functioning as presented here is a more dynamic process. The executive gathers information from the sensory systems and the higher order task influences and triggers production rules. The behavior of these rules can be described following the scheme of methods and scripts as presented in chapter 5. In this design, the strategy to solve a visual task is translated into a *method* for solving the task, but not a well-defined *script*, because the execution largely depends on the sensory input and task-specific instructions. One could therefore argue that the formalization according by means of methods is less powerful than the classical approach because the generic methods do not allow to express as much detail about how the task is solved as a classical routine. However, one of the main inspirations for the design of visual routines was a class of problems – visuospatial ones – that required a generic approach, as their nature caused them to take on infinitely many forms. Visual cognition should be able to solve such problems, and should moreover also be able to do so in changing or unpredictable environments and swiftly changing task requirements as can occur in our daily environments. This illustrates that a generic approach to visual cognition and visual routines seems preferable.

Nevertheless, despite its largely different implementation visual executive functioning still allows for implementation of sequences of operations that are characteristic for the visual routines theory. Although the execution of these sequences can not be fully expressed beforehand, as they are dependant on the stimulus properties, they can be expressed in a method illustrating the strategy that is used in the task. In this aspect – but not exclusively this aspect – the model differs from Rao (1998)'s attempt to update the visual routines theory. Rao expresses visual routines as sequences of attentional shifts that result through experience and are almost impossible to express as a formal strategy. The model presented here combines this with the classical view: the execution of visual problem solving is indeed largely dependant on patterns of attentional modulation that arise from experience and may seem impossible to express. However, the strategy and the sequence of subtasks that largely shape the elemental operations can be expressed as a method.

Another notion from the classical model that still holds for the new framework is that the control of this architecture by the visual executive still operates as an intermediate layer between the sensory process and higher order cognitive processes. Section 4.5 addressed this, and indicated how ex-

ecutive functioning relates to other cognitive processes; visual cognition can be defined as the set of production rules that interact with the visual system. This definition is important, as it also illustrates the boundaries of visual cognition and by extension this thesis. Although the design allows to easily expand this model with various different components implementing other sensory and motor components – which relates the framework to the theory of embedded cognition as described by Ballard et al. (1997) – this study is explicitly limited to the visual domain, and does not address mechanisms from other modalities or higher order cognitive processes, even though they might indirectly influence visual problem solving. For example, when the task is looking for a car in an urban scene, the framework as discussed here illustrates properties of the target are used to influence visual search, but reasoning about plausible locations for the car (e.g. looking at roads or in parking spaces) is not considered. Nevertheless, it must be noted that such mechanisms could be easily integrated in this framework. For example, one could envision a mechanism where the gist of the scene (for example analyzed cf. Oliva & Torralba, 2006) could activate representations of world knowledge that can be accessed by the visual executive. These representations could then be translated into appropriate top-down signals in the sensory components.

This summary of the most important changes to the visual routines framework and the scope of this research illustrates how the new framework provides a clear conceptualization of visual cognition, the way it operates and the way it relates to both sensory processes and higher order cognition. Aside from a detailed conceptualization of these complex mechanisms, the new framework allows to make predictions on visual processing and visual problem solving in particular. The next section will discuss some of these predictions.

6.2 Predictions

It has been stressed that the analysis of the classical framework and the design of a new model has primarily taken a functional approach and has focused on the requirements of a visual problem solving mechanism in general rather than constructing a model for visual cognition based on empirical findings. However, the need for an updated theory for visual routines was marked by our modern understanding of visual and visual attention, which is shaped by empirical findings about the mechanisms and brain areas involved. Therefore, the components that were used to construct the model have all been related to brain areas based on the proposed functions and properties of these areas. The table presented below shows the associated brain areas that were introduced throughout chapter 4. The areas in the top half of the pyramid describe the sensory components, and the framework

predicts feedforward activation patterns in these areas upon presentation of visual stimuli. When elemental operations affect visual processing, this will alter the activation pattern in those areas that represent the relevant information.

Component	**Related brain area(s)**
Visual Pyramid	Feature sensitive units in striate - & extrastriate cortex
PPM	Parieto-occipital area (PO)
FHM	Frontal Eye Fields (FEF)
HBPM	Superior Colliculus (SC)
Blackboard	Pulvinar
Working Memory storage	Distributed (All of the above areas)
Working Memory control	Prefrontal Cortex (PFC)
Visual Executive	Prefrontal Cortex (PFC) and Anterior Cingulate Cortex (ACC)

As is illustrated in the table, this framework also predicts a dual organization of working memory. As discussed in section 4.4, the exact underlying mechanisms of both storage and control are not clear, but a coarse dissociation is that working memory items are represented in the sensory areas, and either reconstructed or maintained there by working memory control mechanisms regulated by prefrontal areas. The section also discussed how this would imply two different types of working memory interference; that which disrupts the item representation by sensory information that use the same areas, and that which is realized by 'overloading' working memory control which hinders maintenance. Working memory control and Visual Executive functioning are proposed to rely on similar mechanisms, so working memory control is not only affected by overdemanding working memory tasks, but also by attentional tasks, which can also be realized in different modalities. Studies that address the capacity of working memory will therefore need to consider this duality in order to measure storage capacity or control capacity.

Aside from the predictions about representation of visual information, the framework – as a model for visual cognition – makes predictions on how visual problem solving is achieved. One of the most important differences between this model and other visual routines models is that it relies on distributed processing. As the ball tracking example in section 5.3.3 illustrates, the steps in visual problem solving are not necessarily based on cues from individual components as information from all areas can be used. However, the task demands and the sensory input could determine the extent to which each component is used to solve the task: the curve tracing example (section 5.3.1) illustrates how different stimulus properties determine whether tracing can rely on cues from peripheral vision, or requires attentional tuning

in the pyramid to construct a detailed representation that may need to be communicated to visual executive functiioning via the blackboard.

This way, the framework makes predictions on how visual stimulus properties affect response times and performance. When the task can be solved by coarse representations that do not require attentional mechanisms this will be preferable because it will be faster. Similarly, attentional tuning does not always require complete tuning of the entire hierarchy to construct a representation as fine-grained as possible. However, the 'amount' of attentional tuning is not the only factor that affect response times. The 'shape' of attentional properties is determined by the influence from the visual executive, and is based on the task, as well as expectations on the stimulus properties. When attentional tuning based on these expectation does not result in the expected attentional sample, new tuning parameters will need to be used until the subtask has been completed. This illustrates how unexpected stimulus properties would affect response times. Finally, it must be emphasized that these issues all affect response times, but that this is not the only factor that define the difficulty of an attentional task. If the task heavily relies on working memory representations, interference in the storage areas with other information – either from the visual input or from other working memory items – could lead to errors and disrupt performance as well.

These predictions also illustrate the involvement of control areas for working memory control and the visual executive in visual cognition. The elemental operations are defined as influences from the visual executive onto different areas, which implies involvement of the executive either when a new subtask requires changing this influence, or when the current influence is insufficient or incorrect. This illustrates that the involvement of these components is less in tasks that follow expected patterns, where visual executive functioning may be largely automated. In that case, it would be hard to discriminate the different reasoning steps that are used to solve the problem from just the brain activity in the PFC.

To summarize, the proposed model makes several predictions on brain activity, representations of visual information, task difficulty in visual cognition and the involvement of cognitive control, that needs to be validated by future work in vision research. However, the proposed model also indicates other aspects of vision and visual cognition that are still largely undefined or lack a strong theory to explain them. The next section will address some of these points.

6.3 Future work

As the first section of this chapter illustrates, the framework presented here provides an approach to visual cognition that allows it to be easily integrated

into other models for cognition, and can be expanded to include processing of other sensory modalities or motor programs. Although this would cause the model to shift from a model for visual cognition (following the definition used here) to a more general model, it has already been illustrated that higher order reasoning could be used to enhance performance in visual problem solving as well. For example, such a system could reason about environmental cues that are present in the scene. Similarly, the framework does not yet address the issue of recognition in much detail. Recognition systems could, for example, be used to match the visual information that is extracted to concepts stored in long-term memory, and assign meaning to the information present in the scene objects. However, expansion of the model does not have to be restricted to models of human cognition. The framework could, for example, be used in computer vision systems or robotics as well. Especially in robotics, it would be interesting to see how the design of this framework could be used to realize computation of embodiment, by integrating information from different sensory modalities.

However, before a completely implementable system of this framework can be constructed, the definition of the mechanisms in several components in the model will need to be more formally defined. One of the predictions of the model is that the elemental operations that are used to implement steps in visual problem solving can be expressed as the influence of the visual executive on parameters that define the behavior of the sensory components. The selective tuning model was used to provide a formal definition of these influences in the visual pathway, which formally defined the extent of the influence of visual executive functioning in this component. Although section 4.5 illustrated what aspects of task guidance could affect the behavior of the other components, a formal definition which defines the parameters is still lacking. As the discussion on the different implementations of visual routines in chapter 2 illustrated, a formal definition of these components could elicit new issues that have not been considered here, and eventually enrich our understanding of how the different mechanisms in vision operate, but also how visual cognition is expressed in the visual system. One notable exception of a component for which implementable models do exist is the PPM for which it was mentioned that different saliency map models would provide suitable implementations. Future work however will need to elicit the extent to which this mechanism is affected by task guidance. Most saliency map models have been used to model visual attention as a whole, including covert and overt shifts, which may have led to an overestimation of task influences in these models (Itti & Koch, 2001; Navalpakkam & Itti, 2005). The framework presented here illustrates how saliency models only one of many attentional mechanisms, and suggests that it may be better to use it to model phenomena such as attentional capture.

The analysis also elicited several questions about components that will need more empirical findings before formal models can be constructed. One

example is illustrated by the issues that persist in defining inhibition of return (IOR) despite a multitude in recent findings. Again, the framework presented here could be used to guide future research in IOR, to isolate findings that can be attributed to IOR with respect to those that arise from different mechanisms (such as negative priming or inattentional blindness). Another example of a component that still needs more extensive research to construct a formal model is visual working memory. Conflicting findings make it still difficult to dissociate between aspects of working memory that are governed by mechanisms present in the storage facilities provided by the sensory components versus those that rely on working memory control from frontal areas, as is illustrated by the discussion on reconstruction versus maintenance of working memory items (see section 4.4).

Lastly, one aspect of the new framework that requires formalization is the translation from methods into the elemental operations. A formal definition of every elemental operation is a set of many parameters to control the components, and these parameters which define the scirpt for solving a particular visual problem are impossible to formulate manually. Several alternatives have been propsosed: the framework of production rules as presented by (Zylberberg et al., 2011) provides a biologically plausible guideline that have been shown to resemble more general models of cognition, but this framework has not yet been fully implemented. Interesting formal alternatives have already been used to implement visual routines, reinforcement learning scheme proposed by McCallum (1996), or the Dynamic Bayes Networks described by Ballard and Hayhoe (2009).

6.4 Summary

The classical visual routines theory provided an interesting approach to visual problem solving that has inspired many models and systems, but the theory relied heavily on assumptions on vision and visual attention that are challenged by our modern understanding of visual processes. This thesis has attempted to update the visual routines theory, by using a similar functional approach as was used in its original formulation. The resulting framework does not only provide an contemporary model of visual problem solving, but also acts as a framework for visual cognition and provides a detailed definition of what visual cognition is and how it can be realized. The different components in the framework can be related to several brain areas which illustrates predictions on brain activity during visual problem solving. Also, the framework can be used to derive predictions on the capabilities of visual processing and the difficulty of tasks, and how visual information is represented. Moreover, it provides a guideline for future work in vision and illustrates important issues that have not yet been sufficiently addressed. These issues all need to be addressed to construct a complete definition and

thorough understanding of visual cognition and visual processing in general.

Bibliography

Anderson, C., & Van Essen, D. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(17), 6297.

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126.

Awh, E., Vogel, E., & Oh, S. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Bajcsy, R., Computer, U. of Pennsylvania. Department of, & Electrical Engineering, I. S. M. S. of. (1985). *Active perception vs. passive perception*. University of Pennsylvania, Department of Computer and Information Science.

Ballard, D., & Hayhoe, M. (2009). Modelling the role of task in the control of gaze. *Visual cognition*, *17*(6), 1185–1204.

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(04), 723–742.

Binford, T. (1971). Visual perception by computer. In *Ieee conference on systems and control* (Vol. 261, p. 262).

Boehler, C., Tsotsos, J., Schoenfeld, M., Heinze, H., & Hopf, J. (2009). The center-surround profile of the focus of attention arises from recurrent processing in visual cortex. *Cerebral Cortex*, *19*(4), 982.

Botvinick, M., Braver, T., Barch, D., Carter, C., & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Brady, T., Konkle, T., & Alvarez, G. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5).

Braver, T., Cohen, J., Nystrom, L., Jonides, J., Smith, E., Noll, D., et al.

(1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, *5*(1), 49–62.

Broadbent, D. (1958). *Perception and communication.* Oxford University Press.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in neural information processing systems*, *18*, 155.

Brunnström, K., Eklundh, J., & Uhlin, T. (1996). Active fixation for scene exploration. *International Journal of Computer Vision*, *17*(2), 137–162.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2-3), 96–107.

Carlisle, N., Arita, J., Pardo, D., & Woodman, G. (2011). Attentional templates in visual working memory. *The Journal of Neuroscience*, *31*(25), 9315.

Carpenter, G., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, *37*(1), 54–115.

Cavanagh, P. (2011). Visual cognition. *Vision Research*.

Cavanagh, P., Hunt, A., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in cognitive sciences*, *14*(4), 147–153.

Cavanagh, P., Labianca, A., & Thornton, I. (2001). Attention-based visual routines: sprites. *Cognition*, *80*(1-2), 47–60.

Clark, J., & Ferrier, N. (1988). Modal control of an attentive vision system. In *Proceedings of the international conference on computer vision* (pp. 514–523).

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(01), 87–114.

Crick, F., & Koch, C. (1990). Some reflections on visual awareness. In *Cold spring harbor symposia on quantitative biology* (Vol. 55, pp. 953–962).

Deco, G., Pollatos, O., & Zihl, J. (2002). The time course of selective visual attention: theory and experiments. *Vision Research*, *42*(27), 2925–2946.

De Fockert, J., Rees, G., Frith, C., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, *291*(5509), 1803.

Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, *4*(8), 2051.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

Deutsch, J., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological review*, *70*(1), 80.

Dewar, M., Della Sala, S., Beschin, N., & Cowan, N. (2010). Profound

retroactive interference in anterograde amnesia: What interferes?. *Neuropsychology*, *24*(3), 357.

Duncan, J. (1979). Divided attention: The whole is more than the sum of its parts. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(2), 216.

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature neuroscience*, *8*(12), 1784–1790.

Farah, M. (1988). Is visual imagery really visual? overlooked evidence from neuropsychology. *Psychological Review*, *95*(3), 307.

Frith, C. (2005). The Top in Top-Down Attention. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *The neurobiology of attention* (pp. 105–108). Academic Press, Elsevier.

Goldman-Rakic, P. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory.

Harrison, S., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635.

Haxby, J., Grady, C., Horwitz, B., Ungerleider, L., Mishkin, M., Carson, R., et al. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(5), 1621.

Hopf, J., Boehler, C., Luck, S., Tsotsos, J., Heinze, H., & Schoenfeld, M. (2006). Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), 1053.

Horswill, I. (1995). Visual routines and visual search: a real-time implementation and an automata-theoretic analysis. In *International joint conference on artificial intelligence* (Vol. 14, pp. 56–63).

Hubel, D., & Wiesel, T. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106.

Hung, C., Kreiman, G., Poggio, T., & DiCarlo, J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863.

Ikkai, A., McCollough, A., & Vogel, E. (2010). Contralateral delay activity provides a neural measure of the number of representations in visual working memory. *Journal of neurophysiology*, *103*(4), 1963.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

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

James, W. (1890). *The principles of psychology* (Vol. 1). New York:Holt.
Johnson, M., Maes, P., & Darrell, T. (1994). Evolving visual routines. *Artificial Life*, *1*(4), 373–389.
Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, *14*(2), 129–140.
Klein, R. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, *4*(4), 219–27.
Koffka, K. (1999). *Principles of Gestalt psychology.* Psychology Press.
Laird, J. (1987). *Soar: An architecture for general intelligence* (Tech. Rep.). DTIC Document.
Lamme, V. (2005). The difference between visual attention and awareness: A cognitive neuroscience perspective. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *The neurobiology of attention* (pp. 167–174). Academic Press, Elsevier.
Lamme, V., & Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.
Lawler, E., & Wood, D. (1966). Branch-and-bound methods: A survey. *Operations research*, 699–719.
Lee, H., Simpson, G., Logothetis, N., & Rainer, G. (2005). Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. *Neuron*, *45*(1), 147–156.
Luck, S., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–280.
MacDonald, A., Cohen, J., Stenger, V., & Carter, C. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835.
Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide.* Lawrence Erlbaum.
Malsburg, C. Von der. (1994). The correlation theory of brain function. *Models of Neural Networks II: Temporal Aspects of Coding and Information Processing in Biological Systems*, 95–119.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* Henry Holt and Co., Inc. New York, NY, USA.
Marr, D., & Nishihara, H. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *200*(1140), 269–294.
Martinez-Trujillo, J., Tsotsos, J., Simine, E., Pomplun, M., Wildes, R.,

Treue, S., et al. (2005). Selectivity for speed gradients in human area mt/v5. *Neuroreport*, *16*(5), 435.

McCallum, A. (1996). Learning visual routines with reinforcement learning. In *Aaai fall symposium 1996.*

Miller, E. (2000). The prefontral cortex and cognitive control. *Nature Reviews Neuroscience*, *1*(1), 59–65.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*(2), 205–231.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, *155*, 23–36.

Olivers, C., Peters, J., Houtkamp, R., & Roelfsema, P. (2011). Different states in visual working memory: when it guides attention and when it does not. *Trends in Cognitive Sciences*.

Orban, G. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological reviews*, *88*(1), 59.

O'Regan, J., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(05), 939–973.

Peissig, J., & Tarr, M. (2007). Visual object recognition: Do we know more now than we did 20 years ago? *Psychology*, *58*.

Pelton, L., & Solley, C. (1968). Acceleration of reversals of a necker cube. *The American Journal of Psychology*, *81*(4), 585–588.

Postle, B. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38.

Postma, E., Van Den Herik, H., & Hudson, P. (1997). SCAN: a scalable model of attentional selection. *Neural Networks*, *10*(6), 993–1015.

Rao, S. (1998). *Visual Routines and Attention.* Unpublished doctoral dissertation, Massachusets Institute of Technology.

Reisberg, D. (2006). *Cognition: Exploring the science of the mind.* WW Norton Nueva York.

Reynolds, J., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annu. Rev. Neurosci.*, *27*, 611–647.

Roberts, R., Hager, L., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, *123*(4), 374.

Roelfsema, P., Khayat, P., & Spekreijse, H. (2003). Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5467.

Roelfsema, P., Lamme, V., & Spekreijse, H. (2000). The implementation of visual routines. *Vision Research*, *40*(10-12), 1385–1411.

Roland, P., & Gulyas, B. (1994). Visual imagery and visual representation. *Trends in Neurosciences*, *17*(7), 281–287.

Rothenstein, A., & Tsotsos, J. (2008). Attention links sensing to recognition. *Image and Vision Computing*, *26*(1), 114–126.

Salgian, G., & Ballard, D. (1998). Visual routines for autonomous driving. In *Computer vision, 1998. sixth international conference on* (pp. 876–882).

Schroeder, C., & Foxe, J. (2005). Multisensory contributions to low-level,[] unisensory'processing. *Current Opinion in Neurobiology*, *15*(4), 454–458.

Sejnowski, T., & Paulsen, O. (2006). Network oscillations: emerging computational principles. *The Journal of neuroscience*, *26*(6), 1673.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424.

Shadlen, M., & Movshon, J. (1999). Synchrony Unbound: Review A Critical Evaluation of the Temporal Binding Hypothesis. *Neuron*, *24*, 67–77.

Shulman, G., Remington, R., & Mclean, J. (1979). Moving attention through visual space. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 522.

Simons, D. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, *4*(4), 147–155.

Sprague, N., & Ballard, D. (2001). A visual control architecture for a virtual humanoid. In *IEEE-RAS Int. Conf. on Humanoid Robots.*

Steinberg, R., Reid, M., & Lacy, P. (1973). The distribution of rods and cones in the retina of the cat (felis domesticus). *The Journal of Comparative Neurology*, *148*(2), 229–248.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*(6747), 869–873.

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5).

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*(6), 599–606.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *nature*, *381*(6582), 520–522.

Tipper, S. (2001). Does negative priming reflect inhibitory mechanisms? A review and integration of conflicting views. *The Quarterly Journal of Experimental Psychology A*, *54*(2), 321–343.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Tsotsos, J. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13*(3), 423–469.

Tsotsos, J. (2010, October). *Re-visiting visual routines: A white paper* (Tech. Rep. No. CSE 2010-11). Dept. of Computer Science Engineering, York University, Toronto, Canada. (`http://www.cse.yorku.ca/csresearch/tech_reports/index.html`)

Tsotsos, J. (2011). A Computational Perspective on Visual Attention.

Tsotsos, J., Culhane, S., Kei Wai, W., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial intelligence*, *78*(1-2), 507–545.

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

Tsotsos, J., & Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia*, *6*(1), 6201.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Ullman, S. (1984). Visual routines. *Cognition*, *18*(1-3), 97–159.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2010). Attentional selection for object recognitiona gentle way. In *Biologically motivated computer vision* (pp. 251–267).

Wang, Z., & Klein, R. (2010). Searching for inhibition of return in visual search: A review. *Vision research*, *50*(2), 220–228.

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

Yantis, S. (1993). Stimulus-driven attentional capture. *Current Directions in Psychological Science*, *2*(5), 156–161.

Yarbus, A. (1967). *Eye movements and vision.* Plenum press.

Yi, W., & Ballard, D. (2009). Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics*.

Zaharescu, A., Rothenstein, A., & Tsotsos, J. (2005). Towards a biologically plausible active visual search model. In *Attention and performance in computational vision: second international workshop, wapcv 2004: Prague, czech republic, may 15, 2004: revised selected papers* (Vol. 3368, p. 133).

Zhang, W., & Luck, S. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zylberberg, A., Dehaene, S., Roelfsema, P., & Sigman, M. (2011). The human turing machine: a neural framework for mental programs. *Trends in Cognitive Sciences*.